This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.
Sex determination, sex chromosomes, and supergenes: evolution of an unusual reproductive system

Robert Baird

THE UNIVERSITY
of EDINBURGH

Submitted for the degree of Doctor of Philosophy
School of Biological Sciences
2023
Abstract

Sex determination systems are remarkably diverse and undergo frequent turnovers, yet our understanding of many of these systems is still fragmentary and the forces that generate this diversity are not yet well understood. Sexual reproduction is also associated with the evolution of sex chromosomes, but there remain many gaps in our knowledge surrounding the dynamics of sex chromosome evolution. One of the most unusual sex determination systems is found in dark-winged fungus gnats (Diptera: Sciaridae). Sciarids break virtually all the conventional rules when it comes to sexual reproduction: males only transmit their maternally-inherited chromosomes, sex is pre-determined by mothers, and females routinely produce single-sex broods. This latter phenomenon, known as monogenic reproduction, occurs in three clades of flies, where it is often associated with chromosomal inversions. In this thesis, I assemble and characterise an inversion-based X-linked supergene associated with monogenic reproduction in the sciarid *Bradysia coprophila*. I determine the age and strata of the supergene and the extent of its functional degradation. I then review current knowledge of monogenic reproduction in three dipteran clades: Sciaridae, Cecidomyiidae (gall midges), and Calliphoridae (blowflies). I also propose a model for the evolution of monogeny in sciarids and analyse early embryonic expression data to identify candidate genes that control it. Finally, I use expression, population, and outgroup whole-genome sequencing data to examine the consequences of this unusual inheritance system for the evolution of its sex chromosomes. I show how studying such systems with unusual genetics provides opportunities to learn more about the evolution of sex determination, supergenes, sex ratios, and sex chromosomes.
Lay summary

Sexual reproduction is a feature common to nearly all multicellular life, yet there exists a vast diversity in the ways in which an individual's sex is determined. However, we still understand only a small fraction of this diversity, much less the forces that generate it. Sex is also usually associated with sex chromosomes. We know that sex chromosomes evolve differently to the rest of the genome, but how and why remain unresolved questions. In this thesis, I study sex determination and sex chromosome evolution in dark-winged fungus gnats, a family of flies that break the fundamental rules when it comes to virtually all perceived wisdom around sex determination, sex chromosomes, and many aspects of genetics. Males only transmit their maternally inherited chromosomes, the sex chromosomes do not necessarily determine sex, and mothers produce bizarre offspring sex ratios. I identify and characterise the chromosome region controlling the unusual form of sex determination and identify candidate genes that may be involved. I also uncover some interesting evolutionary dynamics about their sex chromosomes which may help us to better understand evolutionary processes that affect the genome more generally. Beyond learning about this interesting system, my broader goal is to determine what we can understand about biological norms by studying the exceptions. My findings have provided a foundation for discovering more about these systems with unusual genetics.
Acknowledgements

I have had the privilege during my PhD to work alongside many people whom I have come to admire a great deal. I owe an awful lot to Laura and Andrew, who have not only imparted onto me a vast wealth of knowledge but also critical scientific skills. Laura is a wonderful supervisor who has a proclivity to foster peoples’ curiosity and engage their wacky ideas, no matter how much HiFi reads cost. I hope that my baking has made up for a fraction of the sequencing splurges.

I would like to thank the Ashworth community. It is easy to feel unintelligent amongst so many brilliant minds, but fortunately everyone is nice enough not to draw obvious attention to my unintelligence. Anyway, there are many members of the department from who I have learned a great deal, including, but not limited to, Brian and Deborah, Konrad, Darren, as well as everyone at GJC. Discussions with my secondary supervisor, Simon, have been particularly invaluable. I thank all past and present members of the Ross lab, particularly Kamil and Christina, who have taught me everything about K-mers and fungus gnats, respectively. Thanks to Elpida and Maria, who agreed to work on systems that I convinced Laura it was a good idea to take on. Finally, all my friends in the department – Tamsin, Catherine, Lucy, Julie, Shravan, Priscilla, and many more. A special shout-out to Alex MacKintosh, who has taught me more than I’m sure he will ever realise. I would also like to thank Koorosh in particular for the indispensable intellectual discussions. The topics we covered over the years include, ‘How painful would it be to be boiled in ghee?’, ‘If you had a piece of fruit for a head, which fruit would you choose?’, ‘Are there more dust particles on Earth, or in Saturn’s belt?’, and ‘Would you rather defend your thesis, or eat it?’ Most of these we never quite got to the bottom of. Regarding the last question, I would probably finely shred the paper and mix it with water to create a malleable paste before consumption. Perhaps a thesis gruel with salt and other savoury flavourings would serve as a tasty main. I would probably reserve the supplementary material paste for pudding, with some sugar and cinnamon. Actually, I would probably eat that regardless of whether or not I had to defend the thesis afterwards.

Outside of Ashworth, I would like to thank Malte Grewoldt and Tom Hitchcock for not only being fun collaborators but also great friends. I also thank Susan Gerbi for hosting me in her lab at Brown and imparting her endless Sciara knowledge onto me. Another special shout-
out to my long-term pen-pal John, whom I have never met but have probably exchanged several tomes worth of emails with. John has taught me a great deal more than I’m sure he’s aware of – his scrupulousness when it comes to analysis has definitely rubbed off on me.

Outside of academia, I want to thank all of my close friends (Harry, Josh, Max, Fred, Jen, Andreas, Gio, other Josh… I am lucky to have too many to list). But especially Yaseen for making Edinburgh an even more enjoyable place to live than it already is. Also, of course, my parents and my sisters, Kitty and Susie, for their unyielding support. I could not have been luckier to have been able to write the majority of my final chapter in Tuscany, glass of wine in hand, before retreating to play monopoly deal or president.

I am eternally indebted to the Ventoux, Cloisters, Bennets Bar, The Stockbridge Tap, Leslie’s Bar, Teuchters, Brauhaus, The Hanging Bat, The Dog House, and The Jazz Bar. I could not have done it without you guys.

I would also like to thank the flies. I’m really sorry for the massacres, and I hope that nobody one day discovers you have feelings.
Publications

Chapters II and III of this thesis were prepared for publication, and Chapter V is currently being prepared for publication. For this reason, there is a significant amount of repetition between the introductions of the various chapters, and the placement of the methods section of chapter II was determined by the requirements of the journal to which it was submitted. Furthermore, since chapters II, III, and V were collaborative projects, they are written in first-person plural.

It is unlikely that chapter IV will be prepared for publication in its current form, however the data produced and research conducted in this chapter will likely form part of a more comprehensive paper, with follow-up work, to be submitted for publication at a later stage.

Chapters II and III have been published, respectively, as:

Declaration

I declare that this thesis was written by myself, and that the work has not been submitted for any other degree or professional qualification. I confirm that the work submitted is my own, except when specified, if for example, work has been conducted as part of jointly-authored publications. Below, I have explicitly listed the contributions of other authors involved in the production of this work.

Chapter I. I wrote this chapter with feedback from Laura Ross.

Chapter II. This concept for this chapter arose from discussions with Andrew Mongue, Kamil Jaron, Christina Hodson, and Laura Ross, all of whom contributed in supervising the study, along with Simon Martin. John Urban generated some of the Illumina and Hi-C data used in this chapter, performed some of the structural variant calling analysis and the functional annotations, and carried out some of the analysis in the supplementary material. Malte Grewoldt performed the transposable element annotations and helped to generate figure 2.5. I carried out the remaining analyses, produced the figures, and wrote the chapter with input from John Urban and Laura Ross. All of the aforementioned people reviewed the draft.

Chapter III. I conceptualised and wrote the draft for this chapter. Andrew Mongue and Laura Ross reviewed and edited the draft.

Chapter IV. I wrote and performed all data collection and analysis for this chapter. Laura Ross reviewed the draft.

Chapter V. This chapter was originally conceptualised through conversations with Tom Hitchcock, Andy Gardner, Laura Ross, and Andrew Mongue. Katy Monteith helped with collection of resequencing and outgroup data, as well as DNA extractions. I collected the remaining data. I performed all analysis in this chapter with supervision from Laura Ross and
Andrew Mongue. I wrote the draft, which was reviewed by Tom Hitchcock, Laura Ross, and Andrew Mongue.

**Chapter 6.** I wrote this chapter with feedback from Laura Ross.

Robert Baird, Tuesday 24th October 2023
# Contents

Chapter I. General introduction ............................................................................................................. 13

1.1. The evolution of sex determination systems ................................................................................. 13
  1.1.1. The diversity of sex determination ............................................................................................ 13
  1.1.2. Sex chromosomes: origins and turnovers ................................................................................. 15
  1.1.3. Evolutionary dynamics of sex chromosomes .......................................................................... 17
  1.1.4. Sex determination and the primary sex ratio ........................................................................... 18

1.2. Sex determination in fungus gnats ................................................................................................. 18
  1.2.1. An unusual chromosome cycle and maternally controlled sex determination .................... 18
  1.2.2. Monogenic and digenic reproduction in fungus gnats .............................................................. 21
  1.2.3. Comparison with gall midges and other flies of interest ......................................................... 23

1.3. Fungus gnats as a re-emerging model organism ........................................................................... 25
  1.3.1. What makes fungus gnats interesting? ..................................................................................... 25
  1.3.2. Fungus gnats as a model for studying the evolution of sex determination, sex chromosomes, and sex ratio selection ........................................................................................................... 27

1.4. Thesis aims and structure .............................................................................................................. 27

1.5. References ....................................................................................................................................... 28

Chapter II. Recent evolution of a maternally acting sex-determining supergene in a fly with single-sex broods ........................................................................................................................................... 39

2.1. Abstract ........................................................................................................................................... 39

2.2. Introduction ..................................................................................................................................... 40

2.3. Results ........................................................................................................................................... 44
  2.3.1. X'-X' divergence reveals recent evolution and stratification of the X' chromosome ............... 44
  2.3.2. De novo assembly of the X' sequence ...................................................................................... 48
  2.3.3. Functional degradation of the X' chromosome ......................................................................... 51

2.4. Discussion ....................................................................................................................................... 54
  2.4.1. The X' chromosome evolved recently and shows signs of degradation ................................. 55
  2.4.2. A role for adaptive stepwise expansion of the X' ................................................................. 57
  2.4.3. Evolutionary perspectives on monogenic reproduction in fungus gnats ............................... 58

2.5. Materials and methods .................................................................................................................. 60
  2.5.1. Data collection ......................................................................................................................... 60
  2.5.2. Analysis of X'-X' divergence and evolutionary strata .............................................................. 61
  2.5.3. Genome assembly and annotation .......................................................................................... 62
  2.5.4. Analysis of functional degradation ........................................................................................ 63

2.6. References ....................................................................................................................................... 64

Chapter III. Why put all your eggs in one basket? Evolutionary perspectives on the origins of monogenic reproduction ......................................................................................................................... 74
5.2. Introduction .................................................................................................................. 135
5.3. Results and discussion ................................................................................................. 139
  5.3.1. X chromosomes diverge more slowly under PGE ................................................. 139
  5.3.2. Purifying selection drives slower-X divergence .................................................... 141
  5.3.3. Sex-biased gene expression and hemizygous selection ........................................ 143
  5.3.4. Differences in X and autosome expression may contribute to slower X divergence 146
  5.3.5. Unusual reproductive strategies shape patterns of neutral diversity .................. 147
5.4. Conclusion .................................................................................................................. 150
5.5. Materials and methods ............................................................................................... 150
  5.5.1. Genome assembly, annotation, ortholog identification and sex-linkage assignment for species pairs ........................................................................................................ 150
  5.5.2. Population genomic analyses .................................................................................. 152
  5.5.3. Differential gene expression .................................................................................... 154
  5.5.4. Statistical analyses ................................................................................................ 155
5.6. References .................................................................................................................. 156

Chapter IV. General discussion ......................................................................................... 164
  6.1. Resolving the nature of the sex-determining supergenes ......................................... 164
  6.2. The evolution of monogenic reproduction ................................................................ 165
  6.3. How is sex determined in these systems? ................................................................. 166
  6.4. Evolution of sex chromosomes under PGE ............................................................. 167
  6.5. Conclusion and outlook ............................................................................................ 169
  6.6. References ................................................................................................................ 170

Supplementary materials for Chapter II. Recent evolution of a maternally acting sex-determining supergene in a fly with single-sex broods ................................................... 173

Supplementary figures ....................................................................................................... 173
  Supplementary text 2.1. Mapping rates of XO, XX and X’X libraries.............................. 173
  Supplementary text 2.2. De novo assembly of reads from X’X individuals .................... 175
  Supplementary text 2.3. Binning reads with K-mers ..................................................... 176
  Supplementary text 2.4. Identification of degraded genes of interest ............................ 176
  Supplementary text 2.5. Fly husbandry .......................................................................... 177
  Supplementary text 2.6. Genome annotation ................................................................. 178
  Supplementary text 2.7. Transposable element annotation .......................................... 179
  Supplementary text 2.8. Binning RNAseq reads to identify silenced genes and examine dosage compensation .......................................................... 180
  Supplementary text 2.9. Data used to examine dosage compensation .......................... 181
  Supplementary text 2.10. Removing adapter sequences found in the X’ scaffold .......... 182

Supplementary tables ........................................................................................................ 183

Supplementary figures ....................................................................................................... 190
References ............................................................................................................................................. 192

Supplementary materials for Chapter III. Why put all your eggs in one basket? Evolutionary perspectives on the origins of monogenic reproduction ...................................................................................................................... 197

Supplementary Tables ................................................................................................................................. 197

References ................................................................................................................................................... 199

Supplementary materials for Chapter IV. Molecular control of the sex ratio in dark-winged fungus gnats ...................................................................................................................................... 202

Supplementary materials for Chapter V. Faster adaptation but slower divergence of X chromosomes under paternal genome elimination ........................................................................................................... 203

Supplementary figures .................................................................................................................................. 203

Supplementary tables .................................................................................................................................... 208

Supplementary texts ....................................................................................................................................... 214

Supplementary Text 5.1. Sex-linkage assignments ...................................................................................... 214

Supplementary Text 5.2. Assignments of female genotypes in Bradysia coprophila .................................. 217

References ................................................................................................................................................... 218
Chapter I. General introduction

1.1. The evolution of sex determination systems

1.1.1. The diversity of sex determination

Sexual reproduction is ubiquitous in eukaryotic organisms. As an ancestral feature of this clade, its evolution is regarded as one of the most fundamental transitions to occur in evolutionary history (Szathmary and Smith, 1995). In many multicellular eukaryotes, sexual reproduction involves mating between two different individuals. Separate sexes have evolved independently across many clades, and there exists remarkable diversity in the mechanisms that initiate sexual development, known as the ‘primary signal’ of sex determination (Verhulst et al., 2010). The downstream gene networks regulating sexual development, in contrast, tend to be more conserved. For example, alternative splicing of the Doublesex gene at the bottom of the sex determination cascade initiates sex-specific development in clades as diverged as mammals and worms (Shukla and Nagaraju, 2010). The idea that downstream elements of the sex determination cascade remain conserved while the upstream primary signals undergo frequent turnover is known as the ‘bottom-up’ hypothesis (Wilkins, 1995; Adolfi et al., 2021).

The diversity of sex determination is evident not only between distant taxa but also among closely-related species (Ashman et al., 2014). In some taxa, including many reptiles and fishes, the primary signal is environmental (Bull, 1980; Godwin and Roberts, 2018); most species, however, exhibit genotypic sex determination, whereby genetic elements determine the sex of an embryo. For example, male heterogamety (XX/XY) is common among animals, where Y-determining alleles or dosage of X-linked factors may determine an individual’s sex (Schafer and Goodfellow, 1996; Salz and Erickson, 2010). Genotypic sex determination systems themselves also vary widely: sometimes it is the female that is the heterogametic sex (ZZ/ZW systems, Ellegren, 2011), or it is only the haploid phase of the life cycle which has separate sexes (UV systems, Coelho et al., 2019). In other cases, the sex of an embryo is determined by multiple loci (Moore et al., 2022; Schartl et al., 2023) or maternal loci (Metz, 1938; Ullerich, 1958; Stuart
and Hatchett, 1991). The extent of the diversity in sex determination systems is most obvious in arthropods, among which virtually every known type of system is represented (Blackmon et al., 2017). Understanding the evolutionary forces that drive turnover between systems to generate this diversity remains an open question.

**Table 1.1.** The diversity of sex determination systems.

<table>
<thead>
<tr>
<th>Mechanism</th>
<th>Primary signal</th>
<th>Common examples</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male heterogamety</td>
<td>Genotypic</td>
<td>Mammals; many insects</td>
<td>(Schafer and Goodfellow, 1996; Salz and Erickson, 2010)</td>
</tr>
<tr>
<td>Female heterogamety</td>
<td>Genotypic</td>
<td>Birds; moths and butterflies</td>
<td>(Traut et al., 2007; Irwin, 2018)</td>
</tr>
<tr>
<td>Hermaphroditism</td>
<td>Genotypic</td>
<td>Most plants; some fish; some molluscs</td>
<td>(Heller, 1993; Avise and Mank, 2009; Anderson et al., 2010)</td>
</tr>
<tr>
<td>Haplodiploidy</td>
<td>Genotypic (can be maternal)</td>
<td>Wasps, ants, and bees; thrips; mites; some true bugs; some beetles</td>
<td>(de la Filia et al., 2015)</td>
</tr>
<tr>
<td>Paternal genome elimination</td>
<td>Genotypic</td>
<td>Scale insects; lice</td>
<td>(Herbette and Ross, 2023)</td>
</tr>
<tr>
<td>Parthenogenesis</td>
<td>Genotypic</td>
<td>Stick insects; fairy shrimp; brine shrimp; reptiles</td>
<td>(Elkrewi et al., 2022; Jaron et al.)</td>
</tr>
<tr>
<td>Haploid-phase sex determination</td>
<td>Genotypic (haploid phase)</td>
<td>Brown algae; red algae</td>
<td>(Coelho et al., 2019)</td>
</tr>
<tr>
<td>Temperature-dependent sex</td>
<td>Environmental</td>
<td>Fish; reptiles</td>
<td>(Bull, 1980; Godwin and Roberts, 2018)</td>
</tr>
<tr>
<td>determination</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Polygenic sex determination</td>
<td>Genotypic</td>
<td>Fish</td>
<td>(Moore et al., 2022)</td>
</tr>
<tr>
<td>Maternally-controlled sex</td>
<td>Genotypic (maternal)</td>
<td>Gall midges; fungus gnats; blowflies</td>
<td>(Metz, 1938; Ullerich, 1958; Stuart and Hatchett, 1991)</td>
</tr>
</tbody>
</table>
1.1.2. Sex chromosomes: origins and turnovers

In species with genotypic sex determination, the sexes are usually distinguished by the presence of a sex-limited chromosome: the Y in male-heterogametic (XY/XX) systems or the W in female-heterogametic (ZW/ZZ) systems. Sex chromosomes generally follow a particular evolutionary trajectory: they originate from a pair of homologous autosomes on which a sex-determining locus evolves. Recombination becomes suppressed around the locus, and this suppression subsequently expands along the chromosome. Following this, the non-recombining region irreversibly accumulates deleterious mutations and transposable element (TE) insertions, as offspring recurrently inherit the full mutational load of their parents (Muller, 1964). Functional degradation of the sex-limited chromosome ensues, and in some cases, it is lost entirely. Finally, dosage compensation mechanisms evolve to re-balance expression of the now hemizygous sex chromosome in the heterogametic sex (Ohno, 1967; Charlesworth et al., 2005).

**Figure 1.1.** The classic trajectory of sex chromosome evolution. Sex chromosomes evolve from a pair of previously homologous autosomes on which a sex-determining region (SDR) evolves. Recombination suppression around the region gradually extends along the length of the chromosome, which subsequently degenerates and may be lost entirely.

However, recent studies of sex chromosomes have revealed many exceptions and peculiarities that diverge from this classical theory of sex chromosome evolution. For example, some sex chromosomes seem not to degenerate, and others do not evolve full dosage compensation (Furman et al., 2020; Yazdi et al., 2020). Furthermore, multiple theories attempt to explain specifically why recombination suppression occurs and how degradation and dosage...
compensation evolve. Until recently, prevailing theory posited that selection favours recombination suppression between sex-determining loci and sexually antagonistic alleles (Charlesworth and Charlesworth, 1980; Rice, 1987). More recent alternative hypotheses implicate roles for local adaptation (Connallon et al., 2018), regulatory evolution (Lenormand and Roze, 2022), and sheltering of deleterious mutations (Jay et al., 2022; Olito and Abbott, 2023). Nonetheless, concurrent theories and empirical studies assert that regions of recombination suppression expand in a stepwise process to encompass increasingly large portions of the chromosome, producing evolutionary strata of varying ages (Lahn and Page, 1999; Handley et al., 2004; Vicoso et al., 2013).

In many clades, sex chromosomes undergo striking rates of turnover. For example, the formation of neo-Y chromosomes via Y-autosome fusions is common in Drosophila (Mahajan and Bachtrog, 2017; Nozawa et al., 2021), and, more generally, sex chromosome turnovers have occurred many times in Diptera (Vicoso and Bachtrog, 2015). Frequent turnover of sex chromosomes happens in a diverse range of taxa including fish (Kitano and Peichel, 2012), amphibians (Jeffries et al., 2018), and plants (Tennessee et al., 2018). In contrast, the sex chromosomes of some lineages are relatively conserved, such as birds (Zhou et al., 2014), mammals (Cortez et al., 2014), and Lepidopterans (moths and butterflies, Fraïsse et al., 2017). Again, what drives this turnover, and in particular the evolutionary forces that drive and maintain recombination suppression between chromosomes, remains an open question.

Non-recombining regions on sex chromosomes also share many attributes with supergenes. Over the last decade there has been an increasing recognition of the role of supergenes in driving the evolution of diverse and complex traits across animals, plants, and fungi (Branco et al., 2018; Yan et al., 2020; Huu et al., 2020), including examples that determine sex ratios (Lagunas-Robles et al., 2021). Supergenes are typically defined as clusters of loci that become tightly linked, often by inversions, and control a polymorphic phenotype (Charlesworth, 2016). In many ways, supergenes thus resemble sex chromosomes (or a sex chromosome could be defined as a sex-determining supergene): both can evolve in a stepwise manner (Vicoso et al., 2013; Huu et al., 2020) and undergo functional degradation if they do not recombine (Bergero et al., 2015; Tuttle et al., 2016). The key distinction between non-recombining regions on sex chromosomes and supergenes is that the complex phenotype that the sex chromosome confers is sex itself, and this is the result of a sex-determining locus being
present heterozygously in one sex. Much of the theory and understanding of sex chromosome evolution can thus be applied to supergenes (Charlesworth, 2016).

1.1.3. Evolutionary dynamics of sex chromosomes

Aside from their role in determining sex, sex chromosomes have specific properties that set them apart from autosomes. As such, they are well-placed to inform us about the evolutionary forces that shape the genome. For example, sex chromosomes are predicted to undergo different rates of evolution compared to autosomes. Following recombination suppression, when a locus on the sex-limited chromosome degenerates, the corresponding allele on the homolog that still recombines in the homogametic sex is now present in a single copy in the heterogametic sex. Mutations that occur on these alleles will therefore always be exposed to selection in that sex, even if they are recessive. This improves the efficacy of selection in driving the fixation of beneficial mutations and purging of deleterious ones, which can increase or slow the rate of nonsynonymous changes, respectively, depending on the strength and direction of selection. Furthermore, following recombination suppression, these X- (or Z-) linked loci have a reduced effective population size relative to autosomes. This is predicted to reduce the efficacy of selection. X-linked loci are therefore predicted to undergo a faster rate of evolution relative to autosomes, a phenomenon known as the ‘faster-X’ effect (Charlesworth et al., 1987).

Predictions for faster rates of evolution of X- or Z-linked loci have been confirmed in many species (Kousathanas et al., 2014; Ávila et al., 2014; Wright et al., 2015; Jaquiéry et al., 2018; Bechsgaard et al., 2019; Mongue et al., 2022), although support in some lineages is mixed (Thornton et al., 2006; Pinharanda et al., 2019). Furthermore, the causes can be difficult to disentangle. Some studies find that faster-X effects are predominantly due to drift (Mank et al., 2010); others, adaptive evolution (Charlesworth et al., 2018; Mongue et al., 2022), suggesting that the causes of faster-X evolution may be lineage-specific or otherwise unclear. Systems with unusual inheritance and reproduction, such as aphids (Jaquiéry et al., 2018) or those with paternal genome elimination (discussed later in this thesis) may allow for some of these causes to be disentangled.
1.1.4. Sex determination and the primary sex ratio

The sex determination system of a species and its primary sex ratio are intricately linked. For example, where sex determination is governed by X and Y chromosomes, then the segregation of those chromosomes in XY individuals will decide the primary sex ratio. The sex determination system thus affects how the sex ratio changes (West, 2010). This creates an arena for genetic conflicts, as different genomic elements (e.g. nuclear or cytoplasmic) may favour the production of a particular sex (Werren and Beukeboom, 1998). Classical sex ratio theory predicts that frequency-dependent selection should result in a 1:1 male:female sex ratio in a population (Fisher, 1930), and there is an abundance of theoretical and empirical literature on sex ratio evolution, in particular on how parents are selected to control offspring sex (Hamilton, 1967; Clutton-Brock et al., 1986; Werren, 1987). However, there are many deviations from the norm, notably in Hymenoptera (King, 1987; Meunier et al., 2008), where as a result of their haplodiploid genetics, the sex ratio can be controlled by mothers through modulating fertilisation rate (although sometimes there is a genetic component, Pannebakker et al., 2008; Lagunas-Robles et al., 2021). There are also some strange exceptions to the rules of sex ratio selection in some dipteran clades, which will be discussed later in this thesis.

Bull (1983) highlighted that selection on sex ratios is one of the main underlying forces that drives the evolution of sex determination systems and produces the diversity that we see in nature. Because the sex ratio concerns the transmission success of various genetic elements – be they parental, zygotic, cytoplasmic, or nuclear – through male gametes versus female gametes, manipulation of the sex determination system to bias the sex ratio in favour of those elements can drive the evolution of new sex determination systems.

1.2. Sex determination in fungus gnats

1.2.1. An unusual chromosome cycle and maternally controlled sex determination

Most organisms exhibit regular Mendelian inheritance: one member of each chromosome pair segregates into the sperm or egg following meiotic division, and fertilisation restores the diploid
state. Maternally- and paternally-inherited chromosomes are therefore equally likely to be transmitted to the next generation. However, many species have inheritance systems that violate these fundamental laws of transmission (Burt and Trivers, 2006). This occurs in the dipteran family Sciaridae, known as dark-winged fungus gnats, through maternally-controlled paternal genome elimination (PGE). In this introduction, I will refer to dark-winged fungus gnats simply as ‘fungus gnats’, although note that in the literature, this common name can also refer to the family Mycetophilidae, which exhibits regular chromosome inheritance (White, 1973).

Fungus gnats constitute a large family of lower dipterans (Nematocera) that comprise over 2000 described species (Shin et al., 2013). Several species are synanthropic and are receptive to being cultured in laboratory conditions; some in the genera Lycoriella and Bradysia (previously both Sciara) have been studied since the 1920s (Metz, 1925). Their complicated system of chromosome inheritance has thus long been appreciated. The life cycle of fungus gnats involves several rounds of chromosome elimination. One occurs during spermatogenesis when all paternally-derived chromosomes are excluded. The result is a single spermatocyte which is haploid, except for two copies of the maternally-derived X (which remain as a consequence of meiotic nondisjunction). Another elimination step happens only in somatic cells, where the germline-restricted chromosomes (GRCs) are lost. These are the only paternally-transmitted chromosomes that escape elimination; not all members of the family possess them (Hodson and Ross, 2021). A further elimination occurs when sex is determined during the 7th-9th embryonic cleavage divisions. Here, either one or two X chromosomes, which are always derived from the sperm, are eliminated from the embryo, initiating female (XX) or male (X0) development, respectively (Gerbi, 2022). X elimination is presumably governed by maternally-deposited transcripts or proteins that act in the early embryo (Sánchez, 2010). The chromosomes bound for elimination fail to divide at anaphase and are left behind on the metaphase plate (DuBois, 1933).
**Figure 1.2.** Anaphase at the 7th embryonic cleavage division in the fungus gnat *Bradysia coprophila*, where paternal X chromosomes fail to segregate and remain on the metaphase plate. (A) One X chromosome is eliminated in embryos that develop into females (four divisions shown). (B) Two are eliminated in embryos that develop into males (one division shown). Drawings from (DuBois, 1933).

In fungus gnats, sex determination occurs post-zygotically. Rather than being initiated at the moment of fertilisation, as in most other systems (e.g. if an embryo inherits a Y chromosome from the sperm), zygotes are genetically sexless until X elimination. Following X elimination and zygotic genome activation, since males are X0 and females are XX, it is likely a difference in dosage of an X-linked factor that acts as the primary signal of sex determination which triggers sexual differentiation, like in many other dipterans (Sánchez, 2014). The sex determination master switch – the gene at the top of the cascade – thus probably acts in the zygote. However, in fungus gnats it is the maternal genotype that predetermines sex. There are likely what could be referred to as ‘maternally acting sex ratio modifiers’ that control the proportion of embryos in a brood that develop as male or female. The precise mechanisms of sex determination in this clade will be discussed later in this thesis.
1.2.2. Monogenic and digenic reproduction in fungus gnats

In the model species *B. coprophila*, two genotypically distinct female morphs exist; one produces only sons and the other only daughters. Production of single-sex broods is known as monogenic reproduction. Monogeny is known in several other *Bradysia* species, in addition to other fungus gnat genera including *Lycoriella*, *Scatopsiciara*, *Corynoptera*, and *Rhynchosciara* (Metz, 1938; McCarthy, 1945a, 1945b; Lara *et al.*, 1965). Outside of fungus gnats, monogeny occurs in two other clades of flies: the gall midges (Cecidomyiidae) and the blowflies (Calliphoridae). Similar phenomena are seen in some hymenopterans (wasps, bees, and ants): various parasitoid wasps are known to produce single-sex broods, and fire ant queens can specialise in the production of either future queens or male drones (‘split-sex ratios’). In the hymenopteran cases there is also sometimes a genetic component (King, 1987; Lagunas-Robles *et al.*, 2021).

In fungus gnats, male-producing females are referred to as androgenic and female-producing females as gynogenic. While male and female fungus gnats are easily distinguishable by their size and genital morphology, wild-type androgenic and gynogenic females are phenotypically indistinguishable. Fortunately, irradiation-induced X-linked genetic markers allow for the two female types to be distinguished in laboratory cultures (Metz and Ullian, 1929; Metz and Smith, 1931, *Figure 1.3*).

*Figure 1.3.* Gynogenic female, androgenic female, and male *Bradysia coprophila* individuals. Males are identifiable by the presence of their genital claspers, but they also have smaller bodies and wings compared to females. Androgenic and gynogenic females are phenotypically identical except in lines that contain mutant markers, such as the dominant marker *Wavy*, which
is present only in gynogenic females (Metz and Smith, 1931). Figure modified from (Urban et al., 2021).

In two fungus gnat species, *B. coprophila* and *B. impatiens*, monogeny is known to be associated with large inversions on the X chromosome (Carson, 1946; Crouse, 1960, 1979). The affected chromosome is termed the X' (prime). Gynogenic females are always heterozygous for the X' (X'X) while androgenic females lack it (XX). The X' is transmitted in a Mendelian fashion so that half of female offspring inherit it. The X' is therefore never found in males. The inverted region suppresses recombination between the X and X' chromosome, preserving the factor(s) for female production on the X', as well as preventing transfer of any X' or X-linked markers between the two chromosomes (Metz and Smith, 1931).

**Figure 1.4.** The XX (A) and X'X (B) salivary gland polytene chromosomes of androgenic and gynogenic *B. coprophila* females. Polytenes chromosomes result from many rounds of DNA replication that produce sister chromatids that remain synapsed. The X chromosome of *B.
*coprophila* forms a circular-like structure because two repetitive regions of the X chromosome pair with one another, causing it to fold into a loop (Crouse, 1977; Urban *et al*., 2022). In X'X individuals, the X an X' are unpaired throughout much of their length (highlighted above by yellow lines). These preparations were generated by Anne Kerrebrock in the Yamashita lab at the Whitehead Institute, MIT, and are used with permission.

Fungus gnats that produce mixed-sex broods rather than single-sex broods are referred to as digenic. However, digenic species do not routinely produce male and female offspring in equal proportions, rather, significant departures from a 1:1 primary sex ratio occur more often than not. In some species, digenic females sometimes even produce single-sex progenies (Davidheiser, 1947; Rocha and Perondini, 2000). Monogenic species are also known to produce occasional ‘exceptional’ offspring of the wrong sex (Metz and Schmuck, 1929), suggesting that the distinction between monogeny and digeny may be one of degree rather than kind. To complicate matters further, sex determination has a temperature-dependent maternal effect, with changes in temperature during oogenesis resulting in alterations of the primary sex ratio (Nigro *et al*., 2007; Farsani *et al*., 2013). The mechanism for the temperature effect, however, remains unknown.

1.2.3. *Comparison with gall midges and other flies of interest*

The common term ‘fungus gnat’ is often (correctly) used to refer to Mycetophilidae, the family from which dark-winged fungus gnats (Sciaridae) are thought to have originated, along with the gall midges (Cecidomyiidae, White, 1973). Like dark-winged fungus gnats, gall midges are very speciose, comprising over 5000 described species (Skuhrava, 2006), although a metabarcoding study estimated the total number of species to be as high as 1.8 million, which would make it the most species-rich metazoan family (Hebert *et al*., 2016). Both Sciaridae and Cecidomyiidae comprise many pests of agricultural significance and are commonly found in farms and glasshouses (Hall *et al*., 2012; Shin *et al*., 2013). The genetic similarities between the two families are analogous in such a way that makes their independent evolutionary origins seem improbable. They both have non-Mendelian inheritance systems with maternally-controlled PGE during spermatogenesis, in which the mechanism for paternal chromosome elimination is
the same (White, 1973). In both clades, sex determination happens post-zygotically via elimination of paternal X chromosomes. Some cecidomyiids are digenic and others monogenic, and in the Hessian fly *Mayetiola destructor*, monogeny is associated with an autosomal, female-limited heterozygous inversion. Another unusual genetic feature they share is the GRCs, the supernumerary chromosomes that are eliminated from somatic cells during early embryogenesis. One of the peculiarities specific to fungus gnats is the strange second male meiotic division. Spindle formation is aberrant, and results in non-disjunction (failure to segregate) of the X chromosome. The result is the production of a single sperm (rather than four), all of which are diploid for the maternally-inherited X. In gall midges, the second meiotic division is conventional, and two sperm are produced that both have one copy of the maternally-inherited X chromosome. Why this difference exists is unknown. However, gall midges only eliminate one X from male embryos post-zygotically, while no chromosomes are eliminated in female embryos. The fact that sperm are XX in fungus gnats means that zygotes begin as XXX, and so one and two X chromosomes must be eliminated from female and male embryos, respectively (Stuart and Hatchett, 1988; Gerbi, 2022).

Despite all these similarities, fungus gnats and gall midges are separated by >147 million years of evolution (Hodson and Ross, 2021), and between them lie several intermediate families, including Diadocidiidae, Mycetophilidae, Lygistorrhinidae, Keroplatidae, and Bolitophilidae (Ševčík et al., 2016). Compared to Sciariidae and Ceicodmyiidae, these families are poorly studied in terms of their karyotypes, genetics, and reproduction. Mycetophilidae have apparently orthodox chromosome cycles (White, 1973), and GRCs are absent in the 11 species (from 5 genera) that have been karyotyped (Fahmy, 1949; Le Calvez, 1950). However, it was recently found that the GRCs of the fungus gnat *B. coprophila* have closer homology to a gall midge genome than the core genome of their host, suggesting that some of the genetic features that the two families share may have been transferred in an ancient hybridisation event (Hodson et al., 2022). Despite this, their evolutionary relationship to one another remains puzzling and requires a better understanding of the biology of both clades.

Members of the other dipteran clade with monogenic reproduction, blowflies, do not seem to share the other unusual genetic features of fungus gnats and gall midges. Moreover, while fungus gnats and gall midges are lower dipterans (Nematocera), blowflies are higher dipterans (Brachycera), separated by ~200 million years of evolution (Wiegmann et al., 2011).
Consequently, a common origin for monogeny in the higher and lower dipteran families is unlikely.

1.3. Fungus gnats as a re-emerging model organism

1.3.1. What makes fungus gnats interesting?

Charles W. Metz was a graduate student in Thomas Hunt Morgan’s lab at Columbia University when he began studying Drosophila. Before long, he established colonies of B. coprophila and B. ocellaris which he had collected from a pigeon house of Cold Spring Harbour Laboratories. Through studying these lab colonies, he laid down the foundations for much of our understanding of the complex genetic system of these flies (Metz, 1938), and he was convinced that many of the unusual genetic features of sciarids must be interdependently linked. Later, Helen V. Crouse maintained Bradysia stocks first as a research associate in Metz’ lab and then as a graduate student of Barbara McClintock at the University of Missouri. In 1976, the stock centre was moved to Brown University, where it has since been maintained by Susan Gerbi. The model genus Bradysia is referred to in much of the literature as Sciara but was formally renamed by Steffan (1966); Sciara now refers to a more distantly-related (but understudied) fungus gnat genus (Shin et al., 2013). Confusingly, the B. coprophila has also been referred to as B. tilicola (Loew, 1850) and B. amoena (Winnertz, 1867). Although the species is usually referred to by taxonomists as B. tilicola (Shin et al., 2013; Ševčík et al., 2016), in all of the molecular, cell, developmental, and evolutionary biological literature, it is referred to as B. coprophila. As such, I will use the latter nomenclature throughout this thesis.

Besides their unusual chromosome inheritance and sex determination, fungus gnats have other unusual genetic features. For example, they are well-studied for their DNA puffs – chromosomal regions of developmentally regulated gene amplification – that allow for rapid production of salivary gland proteins needed for cocoon construction when larvae pupate (Gerbi et al., 1993; Andrioli et al., 2008). They also have unique DNA repair mechanisms that render them highly resistant to irradiation (Urban et al., 2021), as well as genetic pathways that provide resistance to low temperatures for long periods (Cheng et al., 2020). Their study has
contributed to our understanding of biological processes that are integral features of metazoans, including programmed cell death (dos Santos Brandão et al., 2014), insect immunity (Candido-Silva et al., 2008), ribosomal DNA and telomere organisation (Kerrebrock et al., 1989; Fernandes et al., 2012) and formation of nuclear organising regions (Madalena et al., 2010; Simon et al., 2016). *B. coprophila* was also subject to the first description of the term ‘imprinting’ in any system, which was used to describe the maternally controlled elimination of paternally-inherited chromosomes and the protection of those that are maternally-inherited (Crouse, 1960). This term is now widely used to describe patterns of parent-of-origin specific gene expression in mammals and flowering plants (Barlow and Bartolomei, 2014; Rodrigues and Zilberman, 2015).

Fungus gnats are also significant agricultural pests (Katumanyane et al., 2018; Broadley et al., 2018; Sueyoshi and Yoshimatsu, 2019). The Chinese chive maggot, *B. odoriphaga*, is studied extensively in East Asia for its insecticide resistance (Chen et al., 2017; Li et al., 2022). Many other members of the *Bradysia* genus are minor pests of glasshouses and houseplants (Katumanyane et al., 2018). Research into fungus gnats as pest species and as a model for evolutionary and developmental biology will thus be complementary in increasing the wealth of data, techniques, and other resources, to add to the repertoire that already exists to make these flies an excellent study system. Fungus gnats also have large polytene chromosomes that can be viewed at relatively low magnification (Crouse, 1977, 1979). Multiple species have been cultured in laboratory conditions (Rocha and Perondini, 2000; Nigro et al., 2007; Farsani et al., 2013; Li et al., 2022; Trinca et al., 2023), including lines that have been inbred for around a century with a catalogue of sex-linked genetic markers (Metz, 1938). There are now chromosome-level reference genomes available for *B. coprophila* (Urban et al., 2021, 2022) and *Pseudolycorella hygida* (Trinca et al., 2023), as well as emerging techniques for transgenic manipulation (Yamamoto et al., 2015; Yamamoto and Gerbi, 2022). Together, the growing pool of resources, the agricultural significance, and the interesting biology of these flies make them a promising model system.
1.3.2. Fungus gnats as a model for studying the evolution of sex determination, sex chromosomes, and sex ratio selection

More specifically, the unique features of fungus gnats allow for excellent opportunities to study the evolution of sex determination, sex chromosomes, and sex ratios. As discussed above, turnover between sex determination systems is common and the mechanisms that exist are diverse. Studying these unusual systems will help to understand the evolutionary forces that produce this diversity and turnover (discussed in Chapters II, III and V). Moreover, the way that sex chromosomes evolve compared to autosomes is affected by inheritance patterns and recombination suppression. The unusual transmission genetics and sex-linked inversions found in fungus gnats therefore allow for predictions of sex chromosome evolution to be tested in a unique context (discussed in Chapter V). Fungus gnats also deviate from the stabilised 1:1 primary sex ratios of most other organisms. Why this is the case remains a mystery and requires understanding the mechanisms that underpin sex determination in this system. In doing so, we may learn more about why deviations from the norm in this case are so rare, why they occur, and how they are resolved (discussed in Chapters III and IV). Finally, a consequence of this primary sex ratio deviation is that females produce single-sex broods. As such, large numbers of embryos of known sex can be easily collected, which allows for detailed investigation of every stage of sex-specific development, including maternal control of early embryogenesis (discussed in Chapter IV).

1.4. Thesis aims and structure

Although the peculiarities of the fungus gnat genetic system have been appreciated for around a century, they are still far from well-understood. In this thesis I am to further elucidate their sex determination system and sex chromosome evolution, and in doing so, contribute to our understanding of the wider processes that generate the diversity of sex determination systems we see in nature. I also hope to show that studying systems with unusual genetics can be important for understanding broader evolutionary processes and phenomena.
Specifically, we do not know how and why monogenic reproduction has evolved in any of the dipteran clades in which it is seen. The relationship between monogenic and digenic reproduction is also a mystery, as is the relationship between monogeny and the enigmatic X’ chromosomes of *Bradysia*, the molecular mechanisms governing all the unusual features, how sex chromosomes evolve under the system, and more generally, why these unusual systems evolve. In this thesis, I attempt to tackle these questions. To this end, I have collected and analysed genomic and transcriptomic datasets, primarily using *B. coprophila* as a model.

In **Chapter II**, I focus on assembling and characterising the X’ chromosome of *B. coprophila* as a first step towards understanding the system. Knowing the age and structure of the X’, and what exactly differentiates androgenic and gynogenic females at the molecular level, will help to address the evolutionary history of this sex determination system and lay a foundation for understanding how heterozygous inversions in fungus gnats evolve. In **Chapter III**, my findings from **Chapter II** prompt me to speculate on the evolutionary origins of monogeny in fungus gnats, and to propose a model for its evolution. I also address the evolution of monogeny in the two other dipteran clades, gall midges and blowflies. In **Chapter IV**, I analyse early embryonic gene expression data to understand the molecular underpinnings of sex determination in fungus gnats, identifying candidate loci and proposing a model for control of the sex ratio. In **Chapter V**, I conduct population genomic analyses, utilising outgroup, population data, and gene expression data, to investigate the consequences of this unusual genetic system for the evolution of its sex chromosomes.

1.5. References


Trinca V, Carli S, Uliana JVC, Garbelotti CV, da Silva MM, Kunes V, et al. (2023). Biocatalytic potential of *Pseudolycoriella* CAZymes (Sciaroidea, Diptera) in degrading plant and fungal cell wall polysaccharides. *Iscience* **26**.


Chapter II. Recent evolution of a maternally acting sex-determining supergene in a fly with single-sex broods

2.1. Abstract

Sex determination is a key developmental process, yet it is remarkably variable across the tree of life. The dipteran family Sciaridae exhibits one of the most unusual sex determination systems in which mothers control offspring sex through selective elimination of paternal X chromosomes. Whereas in some members of the family females produce mixed-sex broods, others such as the dark-winged fungus gnat Bradysia coprophila are monogenic, with females producing single-sex broods. Female-producing females were previously found to be heterozygous for a large X-linked paracentric inversion (X'), which is maternally inherited and absent from male-producing females. Here we assembled and characterised the X' sequence. As close sequence homology between the X and X' made identification of the inversion challenging, we developed a k-mer-based approach to bin genomic reads before assembly. We confirmed that the inversion spans most of the X' chromosome (approximately 55Mb) and encodes around 3500 genes. Analysis of the divergence between the inversion and the homologous region of the X revealed that it originated very recently (<0.5 mya). Surprisingly, we found that the X' is more complex than previously thought and is likely to have undergone multiple rearrangements that have produced regions of varying ages, resembling a supergene composed of evolutionary strata. We found functional degradation of around 7.3% of genes within the region of recombination suppression, but no evidence of accumulation of repetitive elements. Our findings provide an indication that sex-linked inversions are driving turnover of the strange sex determination system in this family of flies.
2.2. Introduction

Sex is an ancient feature shared by most eukaryotes, yet the sex determination systems regulating the development of males and females vary widely among animals (Beukeboom and Perrin, 2014) and can evolve rapidly (Saunders and Veyrunes, 2021). Why such a fundamental developmental process as sex determination is variable remains an outstanding question (Bachtrog et al., 2014). Insects include many examples of this diversity and are therefore an excellent model for understanding changes in sex determination systems. While most insects have genetic sex determination mechanisms with distinct sex chromosomes, different chromosomes act as the sex chromosomes in different species, and species differ in whether males (XY and X0 systems) or females (ZW and Z0 systems) are the heterogametic sex, and in divergence between the sex chromosome pair (Bull, 1983; Beukeboom and Perrin, 2014). There are also examples of complete loss of sex chromosomes, where sex is linked to ploidy differences (e.g. haplodiploidy) or elimination or silencing of paternally-derived chromosomes in males. Another remarkable case, where sex is determined chromosomally but in a way that fundamentally differs from the standard XY or ZW systems is that of monogenic sex determination. Here, sex is determined by the genotype of the mother instead of that of the zygote: mothers are genetically predetermined to produce either only male offspring or only female offspring. Monogenic sex determination has evolved in three clades of flies (Diptera): blowflies (Chrysomyinae, Ullerich, 1958), gall midges (Ceccidomyiidae, Benatti et al., 2010) and fungus gnats (Sciaridae, Metz, 1938). Little is known about control of sex determination in blowflies (Scott et al., 2014). However, in the fungus gnat and gall midge species in which karyotypes have been characterised, monogeny appears to be associated with chromosomal inversions (Carson, 1946; Crouse, 1979; Benatti et al., 2010). None of these inversions has yet been characterised and little is known about the nature and the molecular evolution of these regions. Neither the evolutionary history of monogeny, nor how selection acts on sex determining regions that occur outside the context of conventional sex chromosomes, is thus currently understood.

Suppression of recombination through chromosomal inversions occurs in some sex chromosomes (Vicoso, 2019), and several scenarios can favour a lack of recombination (Wright
et al., 2017; Connallon et al., 2018; Lenormand and Roze, 2022; Jay et al., 2022). Prevailing theory posits that this process involves selection for suppressed recombination between the sex-determining locus on the Y or W chromosome, and sexually antagonistic alleles maintained polymorphically at partially sex-linked loci, potentially encompassing increasingly large portions of the sex chromosome in a stepwise process (Charlesworth and Charlesworth, 1980; Rice, 1987; Charlesworth et al., 2005). However, several alternative hypotheses have recently been proposed, including a role for local adaptation (Connallon et al., 2018), regulatory evolution (Lenormand and Roze, 2022) and the build-up of deleterious mutations (Jay et al., 2022). Y- or W-linked inversions may create regions that never or rarely recombine with their homologous X- or Z-linked regions. This creates sex-specific transmission and ensures that the affected regions are always heterozygous, unlike autosomal inversions. Such regions are likely to accumulate adaptive mutations specific to one sex or the other (Connallon et al., 2018). If the region completely fails to recombine, it is liable to accumulate deleterious mutations and transposable elements (Felsenstein, 1974). As a result, the non-recombining Y or W chromosomes undergo functional degradation (Bachtrog et al., 2008) and become a reservoir for repetitive sequences (Chalopin et al., 2015).

In the present study we investigated a female-limited, non-recombining X-linked inversion associated with monogeny in the fungus gnat Bradysia (Sciara) coprophila. This species has been studied extensively since the 1920s (Metz, 1938) and has a complex chromosome inheritance system (Figure 1). Like all members of Sciaridae, it reproduces through paternal genome elimination, where males fail to transmit paternally-derived chromosomes to their offspring as they undergo several rounds of maternally-controlled chromosome elimination targeting the paternal genome (Metz, 1938; Gerbi, 2022). In all studied members of the Sciaridae, the somatic cells of males have an X0 karyotype, while those of females are XX. However, sex is determined by maternally-controlled X-elimination during early embryogenesis, rather than X inheritance. All zygotes begin with three X chromosomes, one inherited from the mother and two from the father - the result of aberrant spermatogenesis involving the nondisjunction of the sister chromatids in the second meiotic division. During the 7th-9th embryonic cleavage divisions, either one or both paternal X chromosomes are eliminated from somatic cells, resulting in the zygotes developing into females (XX) or males (X0), respectively. The eliminated X chromosomes fail to divide at anaphase and are left behind
on the metaphase plate (DuBois, 1933). Germ cells in both sexes eliminate a single paternal X during a resting stage later in development.

In \textit{B. coprophila} and many other Sciaridae, females are monogenic and produce single-sex progenies. Non-monogenic sciarids are ‘digenic’ and produce mixed-sex broods, although both monogenic and digenic species determine sex through X chromosome elimination. Both reproductive strategies occur in multiple Sciaridae genera (Metz, 1938), though their evolutionary relationship to one another remains unclear. Early cytological observations suggested that two monogenic species, \textit{B. coprophila} and \textit{B. impatiens}, possess single long inversions spanning most of the X chromosome (henceforth the inverted chromosome is denoted by X'), for which female-producing females are heterozygous (Carson, 1946; Crouse, 1979). Polytene chromosome staining indicates that such inversions are absent in digenic species (McCarthy, 1945a, 1945b) as well as in at least one species exhibiting mixed reproductive strategies (Rocha and Perondini, 2000). Through a series of cytogenetic studies, Crouse (Crouse, 1979) deduced the structures of the chromosomes in \textit{B. coprophila} and demonstrated that the X' inversion is paracentric and spans most of the length of the chromosome, leaving the two ends of the chromosome, which still synapse with the X, non-inverted. The genome sequence of \textit{B. coprophila}, with all three autosomes and the X chromosome, has recently been published (Urban \textit{et al.}, 2021, 2022), though the sequence and precise nature of the X' inversion remains unknown as the reference genome was generated from X0 males, which lack the X'.

Here, we have shown through comparative analysis of X and X' chromosomes in \textit{B. coprophila} that the structure of the X' is likely more complex than previously thought. Rather than a single paracentric inversion, we found that it resembles a supergene composed of multiple linked inversions that all emerged less than 0.5 mya. Our finding that the X' is young is intriguing given that monogeny is shared by multiple Sciaridae genera (Metz, 1938), and suggests that inversions may drive the turnover of reproductive strategies in this family. We used a novel process of k-mer binning to assign short reads to chromosomes prior to assembly, allowing assembly of ~55Mb corresponding to X' supergene sequence despite its high sequence similarity to the ancestral X chromosome. With assembly and annotation of the X' we compared patterns of evolution between the two homologous sequences and found that the supergene shows some early signs of degradation characteristic of other non-recombining sex chromosomes and supernes. We discuss the implications of our findings for disentangling the
evolutionary relationship between the strange genetic properties of sciarid flies, and in light of the evolution of sex chromosomes and sex-linked adaptive inversions.

**Figure 1.** Sex determination and X chromosome inheritance in *B. coprophila*. While oogenesis is regular, sperm receive two X copies due to X nondisjunction. The mother’s genotype determines offspring sex: all zygotes begin with three X chromosomes and lose either one or two paternal X chromosomes via targeted paternal genome elimination, resulting in female and male development respectively. XX females produce only sons whereas females heterozygous for the X’ (X’X) produce only daughters.
2.3. Results

2.3.1. X-X’ divergence reveals recent evolution and stratification of the X’ chromosome

We set out to identify the breakpoints of the long paracentric inversion previously described in the literature. The size of the X chromosome in *B. coprophila* is estimated as 50 to 67 Mb (Gabrusewycz-Garcia, 1964; Rasch, 2006; Urban et al., 2021; Hodson et al., 2022), and the X’ inversion spans almost the entire chromosome length (Crouse, 1979). We therefore expected the inversion to be slightly shorter than the X. We produced whole-genome sequencing Illumina libraries from X0, XX, and X’X individuals, which when aligned against the recently updated chromosome-scale reference genome that contains sequences for chromosomes X, II, III and IV, but not X’ (Bcop_v2, Urban et al., 2022), resulted in mapping rates of 93.58%, 96.68%, and 96.34%, respectively. That the X’X libraries have approximately the same mapping rate as the XX libraries indicates there is high enough sequence identity between the X and X’ to reliably call structural (SVs) and single-nucleotide variants (SNVs). We found that the lower mapping rate of the X0 reads was explained mostly by a higher microbial content in those libraries (Supplementary Text 2.1).

In an attempt to identify the breakpoints of the long paracentric inversion on the X’, we searched for SVs that could be attributed to the X’ using both Illumina short-read and PacBio long-read alignments from X’X samples, using XX and X0 samples as a control. However, this analysis demonstrated that, in the X’X samples, the region of the X chromosome corresponding to the inverted region on the X’ is highly enriched for discordant paired-read and split long-read alignments that yield long, overlapping SV signals. We interpreted the entangled and contradictory nature of many individual SV calls as suggesting the presence of multiple complex rearrangements and transpositions throughout the region rather than one single paracentric inversion (Figure 2.2A, Table 2.1, Supplementary Figure 2.1, Supplementary Table 2.6). In contrast, HiC reads from X’X and X0 genotypes mapped against the X chromosome clearly revealed the two ‘main’ breakpoints observed cytologically, as well as three repeat regions that likely correspond to folds in the X chromosome (Crouse, 1979), but did not clearly show additional breakpoints along the chromosome (Figure 2.2B,C). Nonetheless, SNV calls from
alignments of X’X Illumina reads to the X chromosome revealed multiple distinct segments of the inversion with different SNV densities, again suggesting that multiple adjacent and/or nested inversions may have occurred at different times, perhaps in a stepwise fashion (Figure 3). We used these SNV calls to delineate putative evolutionary strata using a change-point analysis, and we estimated divergence for each stratum (Figure 2.3, Table 2.2, Supplementary Table 2.1). We found that the region of recombination suppression spans between approximately 4.1 and 62.9Mb on the 67.2Mb X chromosome. All strata were predicted to have emerged less than 0.5 mya. Dxy values calculated from all sites across the chromosome region were 0.0006 for the youngest stratum and 0.0159 for the oldest stratum, corresponding to divergence in years of 0.008 and 0.107 mya, respectively (assuming a similar mutation rate to Drosophila; see methods). Notably, some of the youngest strata had exceptionally low divergence. Estimates for neutrally evolving (synonymous) genic sites ranged from 0.099 to 0.335 mya for the youngest and oldest strata, respectively. Taken together, our findings suggest that a stepwise set of genomic rearrangements formed the X’ chromosome; we therefore set out to target the X’ sequence for de novo assembly.
**Figure 2.2.** (A) SV calls from the X'X genotype are enriched across the X chromosome compared to calls from the X0 genotype, indicating more complex rearrangements for the X' than may be explained by a single paracentric inversion. Start and end positions of SVs are shown with arcs. Only SVs supported by at least 4 reads and with spans greater than 10kb are shown. (B) HiC contact heatmap across the X chromosome for reads from the X'X genotype, as well as (C) for the X0 genotype. Contact showing the two main breakpoints is highlighted by blue dashed lines in (B). Repeats present in both heatmaps are highlighted by green dashed lines in (C).
**Table 2.1.** Number of each type of SV call from X0 and X’X alignments to the X chromosome.

<table>
<thead>
<tr>
<th>Support</th>
<th>SV</th>
<th>X0 genotype</th>
<th>X’X genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short reads (Illumina)</td>
<td>Deletion</td>
<td>20</td>
<td>2697</td>
</tr>
<tr>
<td></td>
<td>Duplication</td>
<td>1</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>Inversion</td>
<td>0</td>
<td>29</td>
</tr>
<tr>
<td>Long reads (PacBio)</td>
<td>Deletion</td>
<td>9</td>
<td>4037</td>
</tr>
<tr>
<td></td>
<td>Duplication</td>
<td>3</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>Inversion</td>
<td>11</td>
<td>267</td>
</tr>
</tbody>
</table>

**Figure 2.3.** Upper panel: the distribution of variant sites between the X and X’, obtained from alignments of X’X reads to the X chromosome, along which posterior means were calculated. Lower panel: the probability of point changes between posterior means were used to delineate putative evolutionary strata. Putative breakpoints between strata are shown as dotted lines in the upper panel.
Table 2.2. Divergence estimates for putative X' strata in millions of years. $D_{xy}$ estimates are calculated from the density of heterozygous sites within each stratum; neutral estimates are calculated from the proportion of synonymous variants in single-copy X-X' homologs within each stratum. Midpoints are means between ages calculated using different estimates of mutation rates and generation times.

<table>
<thead>
<tr>
<th>Stratum</th>
<th>Length (Mb)</th>
<th>Number of homologs</th>
<th>$D_{xy}$ estimate midpoint</th>
<th>$D_{xy}$ estimate midpoint</th>
<th>Neutral estimate midpoint</th>
<th>This study's lowest estimate</th>
<th>This study's highest estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>9.00</td>
<td>326</td>
<td>0.0075</td>
<td>0.099</td>
<td>0.147</td>
<td>0.050</td>
<td>0.219</td>
</tr>
<tr>
<td>S2</td>
<td>0.50</td>
<td>15</td>
<td>0.0040</td>
<td>0.053</td>
<td>0.098</td>
<td>0.027</td>
<td>0.145</td>
</tr>
<tr>
<td>S3</td>
<td>23.80</td>
<td>904</td>
<td>0.0123</td>
<td>0.162</td>
<td>0.221</td>
<td>0.083</td>
<td>0.329</td>
</tr>
<tr>
<td>S4</td>
<td>1.00</td>
<td>35</td>
<td>0.0039</td>
<td>0.052</td>
<td>0.067</td>
<td>0.026</td>
<td>0.099</td>
</tr>
<tr>
<td>S5</td>
<td>1.65</td>
<td>55</td>
<td>0.0151</td>
<td>0.198</td>
<td>0.292</td>
<td>0.101</td>
<td>0.434</td>
</tr>
<tr>
<td>S6</td>
<td>0.10</td>
<td>2</td>
<td>0.0074</td>
<td>0.097</td>
<td>0.099</td>
<td>0.049</td>
<td>0.147</td>
</tr>
<tr>
<td>S7</td>
<td>9.40</td>
<td>350</td>
<td>0.0145</td>
<td>0.191</td>
<td>0.290</td>
<td>0.097</td>
<td>0.432</td>
</tr>
<tr>
<td>S8</td>
<td>0.10</td>
<td>0</td>
<td>0.0006</td>
<td>0.008</td>
<td>NA</td>
<td>0.004</td>
<td>0.011</td>
</tr>
<tr>
<td>S9</td>
<td>2.55</td>
<td>62</td>
<td>0.0153</td>
<td>0.200</td>
<td>0.318</td>
<td>0.102</td>
<td>0.474</td>
</tr>
<tr>
<td>S10</td>
<td>0.40</td>
<td>22</td>
<td>0.0122</td>
<td>0.161</td>
<td>0.233</td>
<td>0.082</td>
<td>0.346</td>
</tr>
<tr>
<td>S11</td>
<td>1.50</td>
<td>50</td>
<td>0.0145</td>
<td>0.191</td>
<td>0.277</td>
<td>0.097</td>
<td>0.412</td>
</tr>
<tr>
<td>S12</td>
<td>1.90</td>
<td>96</td>
<td>0.0159</td>
<td>0.210</td>
<td>0.335</td>
<td>0.107</td>
<td>0.499</td>
</tr>
<tr>
<td>S13</td>
<td>1.30</td>
<td>90</td>
<td>0.0136</td>
<td>0.179</td>
<td>0.260</td>
<td>0.091</td>
<td>0.386</td>
</tr>
<tr>
<td>S14</td>
<td>1.20</td>
<td>36</td>
<td>0.0104</td>
<td>0.137</td>
<td>0.286</td>
<td>0.070</td>
<td>0.425</td>
</tr>
<tr>
<td>S15</td>
<td>3.50</td>
<td>97</td>
<td>0.0096</td>
<td>0.127</td>
<td>0.147</td>
<td>0.065</td>
<td>0.218</td>
</tr>
<tr>
<td>S16</td>
<td>0.95</td>
<td>54</td>
<td>0.0093</td>
<td>0.122</td>
<td>0.200</td>
<td>0.062</td>
<td>0.297</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>58.85</strong></td>
<td><strong>2194</strong></td>
<td><strong>0.0054</strong></td>
<td><strong>0.136</strong></td>
<td><strong>0.218</strong></td>
<td><strong>0.070</strong></td>
<td><strong>0.304</strong></td>
</tr>
</tbody>
</table>

2.3.2. De novo assembly of the X' sequence

We attempted assembly of PacBio reads from X’X individuals, followed by chromosome assignment of scaffolds using sex differences in read depth across the genome (Supplementary Text 2.2, Supplementary Figure 2.2). This yielded a genome size of only 291Mb which was
comparable to the size of the male (X0) genome (Urban et al., 2021). Moreover, we were able to assign only around 3.6Mb as putative X’ sequence (Supplementary Table 2.2). High sequence identity between reads originating from the X and X’ chromosomes was likely leading to their collapsing together upon assembly. To overcome this, we used a process akin to haplotype resolution of diploid sequences by trio binning (Koren et al., 2018). Our approach utilises differences in k-mer frequencies in Illumina reads between sexes to assign them to chromosomes prior to assembly. Taking advantage of high homozygosity due to over a century of inbreeding (Metz, 1938) and the fact that X’ is limited to X’X individuals, we assigned k-mers specific to X’X female reads as likely to belong to the X’ (Figure 2.4A). We used these k-mers to extract the short reads from the X’X dataset as putative X’-specific reads. In contrast to long reads, which have a high likelihood of false k-mer matches due to high sequencing error rates, we found that short reads (75-150bp) can effectively be binned with k-mers due to their low error rate and short length (Supplementary Text 2.3).

**Figure 2.4.** (A) 27-mer frequency heatmap between Illumina reads from X0 versus X’X flies. 27-mers form dense clouds based on k-mer frequency, which reflects ploidy: (i) 27-mers specific to X’X are assigned as putative X’-specific 27-mers. (ii) 27-mers haploid in X0 but diploid in X’X are likely those belonging to the X and the portion of X’ shared by both chromosomes. (iii) 27-mers diploid in both sexes are likely those belonging to autosomes. (iv) 27-mers containing read errors cluster around the origin. (B) Mean normalised per-based genomic coverage across 100Kb
windows of all autosomes, the X and the \( X' \)-supergene sequence, for reads from pooled individuals of each genotype: \( X'X \) (top), \( XX \) (middle) and \( X0 \) (bottom). The main breakpoints of the \( X' \) (dotted lines) are clearly visible in coverage from \( X'X \) reads.

The resulting putatively \( X' \)-specific Illumina reads assembled into 61.7Mb across 42564 contigs, with an N50 of 10Kb and a largest contig length of 87Kb (Supplementary Figure 2.3). We performed reference-based scaffolding of these contigs, using the regular X chromosome scaffold as a reference (Urban et al. 2022), to produce a single scaffold corresponding to the \( X' \) supergene. To gap-fill and polish the \( X' \) scaffold, we combined it with the remaining chromosomes II, III, IV and X (Urban et al., 2022), then used PacBio reads from \( X'X \) individuals competitively mapped against all chromosomes to fill some remaining gaps (Supplementary Figure 2.4), and Illumina reads from \( X'X \) individuals to polish the final assembly. The resulting ~55Mb scaffold is the first model of the \( X' \) sequence contained within the long paracentric inversion breakpoints defined by Crouse (1979, Table 2.3). Due to using the uninverted X as a reference for scaffolding the \( X' \) contigs, this scaffold may be an inaccurate structural representation of the \( X' \) chromosome, thus we did not attempt to use this assembly to infer information about the structure of the \( X' \). However, alignment of Illumina reads from all three genotypes (X0, XX and \( X'X \)) to all chromosomes strongly supports its correspondence to the \( X' \) chromosome (Figure 2.4B). As expected, we observed that (i) \( X'X \) individuals had haploid coverage (1n) across the \( X' \) sequence and the corresponding inverted region of the X chromosome compared to the autosomes, (ii) XX and X0 individuals had very low coverage of the \( X' \), and (iii) XX females had relatively equal coverage across the X and autosomes in XX females (Figure 2.4B).
Table 2.3. Assembly statistics for the *B. coprophila* chromosome-level genome, also showing the X' supergene sequence as well as the portion of the X homologous to the supergene. Predicted genes shown are those predicted and annotated in this study. Gene and TE density were calculated from total gene and TE base pairs as a proportion of the length of each chromosome, excluding N bases.

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Size (Mb)</th>
<th>Size (excluding N bases, Mb)</th>
<th>Gaps</th>
<th>Predicted genes</th>
<th>Gene density</th>
<th>TE density</th>
</tr>
</thead>
<tbody>
<tr>
<td>II</td>
<td>58.34</td>
<td>57.99</td>
<td>26</td>
<td>3968</td>
<td>10.40%</td>
<td>26.06%</td>
</tr>
<tr>
<td>III</td>
<td>71.05</td>
<td>69.01</td>
<td>101</td>
<td>5919</td>
<td>12.68%</td>
<td>26.92%</td>
</tr>
<tr>
<td>IV</td>
<td>97.08</td>
<td>95.49</td>
<td>90</td>
<td>6730</td>
<td>10.57%</td>
<td>28.30%</td>
</tr>
<tr>
<td>X</td>
<td>70.51</td>
<td>67.23</td>
<td>187</td>
<td>3881</td>
<td>8.69%</td>
<td>23.06%</td>
</tr>
<tr>
<td>X (homologous portion)</td>
<td>58.80</td>
<td>56.87</td>
<td>136</td>
<td>3429</td>
<td>8.54%</td>
<td>17.32%</td>
</tr>
<tr>
<td>X'-supergene</td>
<td>54.74</td>
<td>54.50</td>
<td>3486</td>
<td>3470</td>
<td>7.66%</td>
<td>10.51%</td>
</tr>
<tr>
<td>Total (excluding X')</td>
<td>296.98</td>
<td>289.71</td>
<td>404</td>
<td>20498</td>
<td>10.60%</td>
<td>26.31%</td>
</tr>
<tr>
<td>Total</td>
<td>351.73</td>
<td>344.11</td>
<td>3194</td>
<td>23968</td>
<td>10.14%</td>
<td>23.81%</td>
</tr>
</tbody>
</table>

2.3.3. Functional degradation of the X' chromosome

We identified 3470 protein-coding genes totalling 4.2Mb, i.e., 7.7% of the 54.7Mb X' supergene scaffold. The portion of the X chromosome homologous to the supergene spans 58.8Mb and contains 3429 genes totalling 4.9Mb, i.e., 8.3% of the region. Thus, the proportion of the chromosome corresponding to coding sequence is slightly, but not significantly lower on the X' supergene relative to the homologous X region (X-squared=0.00027, df=1, p=0.9869). We also found both the X' and X to have lower gene densities than the autosomes. The difference between gene densities of the X' and X may reflect an increase in non-coding DNA within the supergene, or alternatively, may be a consequence of the assembled X' sequence having more gaps than the X chromosome sequence (Table 2.3). We identified 2321 single-copy homologs between the X' and X. A further 527 genes from the X' and 679 genes from the X across 296
orthologous groups (OGs) were categorised as duplicates. In 64 duplicate OGs, the X' and X chromosomes had the same number of gene copies. Of the remaining duplicate OGs, 162 had more copies on the X compared to the X', while 70 had more copies on the X'. These may be due to X- and X'-specific duplication events, or ancestral duplication and subsequent loss in one or the other chromosome. Unlike X mutations, X' mutations should not be purged by recombination, thus deletions and duplications on the X' are the more likely explanation. We also found that 622 genes across 603 OGs were specific to the X', and 429 (across 359 OGs) were specific to the X. Gain of novel genes and whole-gene deletions from the X' are possible explanations for these finding, although sequence divergence, misassemblies or gaps may have also led to homologs not being found.

We investigated the possibility that the X' has undergone patterns of degeneration similar to other non-recombining sex chromosomes by analysing functional degradation of genes and accumulation of repetitive elements. Of the 2321 single-copy gene OGs, we found that 123 (5.3%) contain X'-specific mutations that are likely to compromise gene function (including frameshift and/or gain or loss of stop or start codon mutations, Supplementary Table 2.3). We further analysed expression of single-copy homologs and found that fewer X' genes are transcribed compared to their X-linked homologs: across four life stages in females, 2191 X' homologs are expressed compared to 2237 X homologs, although only the larval (X-squared=6.07, df=1, p=0.014) and adult (X-squared=6.80, df=1, p=0.009) stages had significantly fewer X' copies expressed (Figure 2.5A). Overall, 7.3% of genes were classified as either silenced, disrupted or both (Figure 2.5B).

The fly stock used in this study was derived from a laboratory stock maintained by Metz (1938), in which X'X females carry an X'-linked, irradiation-induced mutation, Wavy, which alters wing phenotype. As such, it is worth noting that this mutation may cause estimates of degradation to differ from wild-type flies, though its molecular nature is unknown. Among the genes classified as degraded, we identified several with functions in wing development, including the wing polarity protein STAN and the “held out wings” protein HOW, which are both required for regular wing development in Drosophila (Zafran et al., 1997; Adler, 2012), and may serve as candidates for the Wavy mutation. Since X'X females eliminate one, rather than two, X chromosomes from their embryos to produce only daughters, we may expect that this results from the silence or disruption of a maternal-effect X-linked gene on the X' chromosome which
is somehow involved in the control of chromosome elimination. Among the genes classified as degraded, we found several candidates involved in chromatin regulation, chromosome segregation and cohesion (Supplementary Text 2.4, Supplementary Table 2.4).

Figure 2.5. (A) Number of single-copy orthologous genes expressed at four developmental stages in females. Significantly more genes expressed from the X compared to the X' chromosome in larvae and adults; other stages were non-significant (ns). (B) Proportion of functional versus nonfunctional X'-linked genes that have functional X homologs. (C) TE counts in 1Mb bins distributed on the X chromosome and within the X' supergene sequence.

We also checked X'X females for evidence of dosage compensation (DC) of X-linked genes that have corresponding degraded X' homologs. We expected that if these genes were dosage compensated, they would be upregulated in X'X females to match the expression of those genes in XX females where both copies are functional. We compared expression of X-linked genes in X'X and XX female samples (Supplementary Figure 2.5). In total, only 4 X-linked genes were significantly upregulated in the X'X females (Supplementary Table 2.5), none of
which were genes that we classified as pseudogenised. Thus, we found no evidence of DC of degraded genes.

We also analysed transposable element (TE) content across the genome. We found that TE density was lower on the sex chromosomes compared to the autosomes (Table 2.3), but the difference was non-significant (X-squared=0.0065, df=1, p =0.936). We did not find an enrichment of repetitive sequences on the X'; TE density within the X'-specific sequence was non-significantly lower (10.51%) compared to the homologous portion of the X (17.32%, X-squared=0.060, df=1, p =0.807). The fact that the X' sequence was assembled with short reads may have resulted in limited power to detect repetitive sequences compared to the X from the reference genome. Our analyses comparing structural differences between the X and X' suggested that repetitive sequences, such as TEs, may have different or additional locations within the X' sequence not present on the X (Figure 2.2A, Table 2.1, Supplementary Figure 2.1, Supplementary 2.6). However, the distribution of TEs and TE superfamilies across our assembled X' sequence appeared to be similar to that of the same chromosome region on the X (Figure 2.5C).

2.4. Discussion

Recently-evolved sex chromosomes can provide crucial insights into understanding the evolution and turnover of sex chromosomes. B. coprophila is particularly unusual in that it exhibits a major transition from an X0-like system to one not unlike a ZW system, though only half of females are heterogametic and the sex chromosome is maternally-acting rather than acting in the zygote. Thus Bradysia offers a chance to study the early differentiation between non-recombining regions of sex chromosomes in a unique evolutionary context. Our analysis and assembly of the ~55Mb X' sequence revealed that it evolved very recently (<0.5 mya), that it is more complex in structure than previously thought, and that it is beginning to undergo functional degradation characteristic of classical sex chromosome evolution.
2.4.1. The X’ chromosome evolved recently and shows signs of degradation

In insects, emergences and turnovers of sex chromosomes are common (Vicoso and Bachtrog, 2015). *Drosophila* neo-Y chromosomes are among the most well-studied cases. In *Drosophila* males are achiasmatic, so Y-fused autosomes become instantly sex-linked and non-recombining. In the absence of recombination, deleterious mutations and TEs accumulate irreversibly because offspring inherit the full mutational load of their parents (Muller, 1964). Lack of recombination further leads to hitchhiking effects of deleterious mutations, where beneficial mutations arising on the sex-limited chromosome spread to fixation and carry with them linked mildly deleterious loci (Maynard-Smith and Haigh, 1974; Rice, 1987). The inevitable result of such processes is functional degradation of the non-recombining chromosomes. Studies of neo-sex chromosome evolution in *Drosophila* have indeed found that significant degeneration occurs rapidly: In *D. pseudoobscura*, the neo-Y evolved ~15 mya and has very few genes remaining (Carvalho and Clark, 2005; Mahajan and Bachtrog, 2017). The *D. miranda* neo-Y has undergone ~4-fold expansion due to TE accumulation and 40% of the ancestral genes have become pseudogenised in only ~1.5 million years of evolution (Bachtrog, 2006; Bachtrog et al., 2008; Mahajan et al., 2018). The younger *D. busckii* and *D. americana* neo-Ys evolved ~0.8 and <0.47 mya and have 60% and 22% pseudogenised genes, respectively (Vieira et al., 2003; Zhou and Bachtrog, 2015; Nozawa et al., 2021). Our finding that around 7.3% of genes on the <0.5 mya *B. coprophila* X’ chromosome are pseudogenised is consistent with this expected trajectory of sex chromosome evolution. Moreover, because the chromosome is present in only half of females, its effective population size is half that of other sex-limited chromosomes. As such, it should exhibit an accelerated rate of decay as the effects of drift should increase the rate of evolution further at sites under purifying selection (Charlesworth et al., 1987), though the effective population size of the sex chromosomes will depend on the relative reproductive success of males and females (Vicoso and Charlesworth 2009). To explore the potential effects of drift on sex-limited chromosome degeneration further, it will be essential to compare the X’ chromosomes of other Sciaridae species which may have evolved independently and at different times, such as that of *B. impatiens* (see below, Carson 1946).

Sex chromosome degeneration is sometimes accompanied by the evolution of dosage compensation (DC) mechanisms to re-establish diploid expression of the X chromosome in
males (Ohno, 1967). In the ancestral Drosophila X chromosome this is achieved by hypertranscription of X-linked genes in males (Schulze and Wallrath, 2007). The neo-X chromosomes of various Drosophila species have achieved DC via transposon-mediated co-option of DC machinery, though the younger neo-Xs are yet to achieve global DC (Marín et al., 1996; Zhou et al., 2013; Ellison and Bachtrog, 2013). In B. coprophila there is evidence for DC in X0 males through upregulation of X expression (Urban et al., 2021). As for the X'X females, we found no evidence that X-linked genes with degraded X' homologs show DC in X'X females. Given the young age of the X'', there may not have been sufficient time for the establishment of DC mechanisms to compensate for degraded X' genes. Furthermore, because the X' chromosome is present in only half of females, conflict between XX and X'X females over gene expression may hinder the evolution of DC.

Accumulation of repetitive sequences occurs rapidly in recently-evolved sex chromosomes (Chalopin et al., 2015). The neo-Y chromosome of D. miranda has undergone massive TE accumulation and has expanded significantly as a result (Bachtrog et al., 2008; Mahajan et al., 2018), though the repetitive landscape of younger (<0.5 mya) Drosophila neo-Ys (Nozawa et al., 2021) has not been analysed and thus TE accumulation in nonrecombining regions over these shorter timescales is unclear. Despite the recent divergence between X and X' in B. coprophila, one would expect higher TE content within the X'. Differences in TE accumulation may be affected by TE content: the proportions of DNA transposons relative to retrotransposons varies widely between lineages. We find that B. coprophila, similar to the Musca, Aedes and Culex genera, has a higher proportion of DNA transposons than Drosophila (Petersen et al., 2019). Compared to the cut-and-paste mechanisms of DNA transposons, the copy-and-paste mechanisms of retrotransposons may lend themselves to more rapid accumulation (Kim et al., 2012). However, our approach using X'-specific k-mers to assemble the X' may result in failure to assemble repetitive sequences that are shared by other chromosomes, which may explain why we found a lower TE content on the X' compared to the homologous X region. Understanding the dynamics of TE accumulation in this peculiar system will require a more contiguous assembly of the X' as well as examination of other X' chromosomes in Sciaridae species (see below).
2.4.2. A role for adaptive stepwise expansion of the X’

Over the last decade, there has been a growing recognition of the importance of clusters of linked loci within inversion-based supergenes in driving the evolution of diverse and complex phenotypes. These include Batesian mimetic morphs of butterflies (Kim et al., 2022), divergent social behaviours in ants (Yan et al., 2020), mating compatibility in fungi (Branco et al., 2018), as well as several polymorphisms in birds including plumage colour (Funk et al., 2021), reproductive strategies (Küpper et al., 2016) and sperm morphology (Kim et al., 2017). This study of a recent transition in the sex-determining system in flies presents another case. It has been argued that supergenes may be more widespread than previously recognised, that they are important for co-segregation of adaptive variation within a species, and that they may even occasionally result in the spread of complex phenotypes across species boundaries (Schwander et al., 2014; Thompson and Jiggins, 2014).

The evolutionary trajectories of supergenes and sex chromosomes show similarities: some supergenes have evolved in a stepwise manner or have undergone functional degradation, and sex chromosomes also play an important role in adaptation and speciation (Presgraves, 2008; Tuttle et al., 2016; Branco et al., 2018; Kim et al., 2022). Furthermore, the evolutionary fates of inversions differ depending on whether they arise on sex chromosomes or on autosomes, with the probability of spread of an inversion through a population being higher on sex chromosomes. This is further affected by sex-biased migration patterns, dominance of locally adapted alleles and chromosome-specific deleterious mutation load (Connallon et al., 2018). Indeed, X-linked genes are predicted to disproportionately contribute to local adaptation due to exposure of recessive alleles to selection, and sex-linked inversions are therefore more likely to sweep to fixation compared to autosomal inversions (Lasne et al., 2017). For these reasons, X-linkage of this supergene in Sciaridae may have favoured its initial emergence as well as its enlargement along the chromosome. Rather than one long paracentric inversion, our analysis suggests the X’ chromosome has undergone multiple rearrangements, which may be explained by multiple adjacent and/or overlapping inversions, smaller inversions nested within larger ones, or some combination thereof, which have accumulated in a stepwise process to suppress recombination along the chromosome. Some of the smaller strata we identified appeared to be far less diverged than others, which may represent uninverted gaps.
between inversion breakpoints. Alternatively, the inversion(s) of the X’ may lead to complex pairing with the X, which may result in varying recombination rates along the chromosome and produce regions of differing divergence. Further resolution of the X’ structure will be required to determine the precise formation of the chromosome. Nonetheless, it appears that the rearrangements on the X’ accumulated rapidly between 0.02 and 0.5 mya. Expansion of the non-recombining region through additional inversions may have adaptively captured female-beneficial alleles at nearby loci, as sex chromosome evolution theory posits (Charlesworth and Charlesworth, 1980; Rice, 1987; Charlesworth et al., 2005), although this may be hindered by genetic conflict between XX and X'X females.

An alternative explanation for the stratification of the X’ is that sex determination relies on more than one locus, i.e. it is polygenic, and that successive inversions have emerged to control the sex ratio. Among digenic Sciaridae, sex ratios vary significantly: in Bradysia ocellaris and Bradysia matrogrossensis, broods frequently depart from the expected 50:50 sex ratio and are often heavily skewed in either direction (Rocha and Perondini, 2000). The sex ratio in B. ocellaris is heritable, and majority male-production can evolve from majority female-production and vice-versa in as few as six generations (Davidheiser, 1947). Taken together, these observations suggest that multiple loci are involved in sex determination. Monogenic Sciaridae presumably evolved from digenic ancestors, which may have occurred through the adaptive linkage of sex-determining alleles through inversions. The young age of the X’ indicates that repeated evolution of monogeny in Sciaridae may have been favoured under certain circumstances over the ancestral digenic sex determination system.

2.4.3. Evolutionary perspectives on monogenic reproduction in fungus gnats

Within the fungus gnat clade Sciaridae, origins of monogeny and the relationship between the monogenic and digenic reproductive strategies remains poorly understood. At least one other monogenic species, B. impatiens, is known to harbour an X-linked inversion polymorphism (Carson, 1946). Monogeny also occurs in many other Sciaridae, including other Bradysia species, but also in more distantly related members of other genera such as Lycoriella and Corynoptera (Metz, 1938). In this respect, our finding that the B. coprophila X’ chromosome evolved <0.5 mya has intriguing consequences for understanding the evolution of this reproductive strategy.
Unlike the *B. coprophila* X', the *B. impatiens* X' non-recombining region is terminal (Carson, 1946), suggesting that the X' chromosomes in the two species may not be homologous by descent (alternatively, the region has expanded in *B. impatiens* or the terminal portion has re-inverted in *B. coprophila*). Another possibility is that the X' chromosome in *B. coprophila* may be older than our findings suggest but appears younger due to occasional recombination through gene conversion or double crossovers, which occur within large inversions (Navarro et al., 1997). While crossing-over requires synapsis between chromosomes, gene conversion, i.e. the non-reciprocal copying of stretches of sequence between sister chromatids to repair mismatch errors during replication, does not (Szostak et al., 1983; McMahill et al., 2007). Furthermore, in *B. impatiens*, dicentric chromatids were observed to form through pairing between the X and X' along the length of the inversion (Carson, 1946). If such pairing occasionally occurs in *B. coprophila*, it may prevent sequence divergence between the two chromosomes.

Nonetheless, the distribution of monogenic reproduction among sciarids indicates multiple evolutionary origins. For example, within *Bradysia* alone, both monogenic and digenic species exist, and the same pattern is found within other genera (Metz, 1938; Steffan, 1974). If sex determination involves multiple loci, inversions may have emerged in some lineages to fix the production of sex-biased broods in a particular direction. However, this raises the question: what drives the turnover between reproductive strategies in this clade? Haig (1993) suggested that female production evolved as a response to a male-biased sex ratio. Fungus gnats carry a unique type of chromosome only found in the germline (germline-restricted chromosomes, GRCs), in addition to their sex chromosomes and autosomes. The GRCs are disproportionately transmitted by males, and so may have distorted the sex ratio in their favour. Presence of GRCs in Sciaridae does appear to correlate with monogenic reproduction, and many (but not all) digenic species lack them (Hodson and Ross, 2021). In support of this is the observation that a monogenic lab-reared line of *B. impatiens* reverted to digenic reproduction following loss of its GRCs (Crouse et al., 1971). However, the function of the GRCs remains unknown. Interestingly, the GRCs of *B. coprophila* were recently found to have introgressed into Sciaridae following an ancient hybridization event with Cecidomyiidae (Hodson et al., 2022), a clade that shares many features with Sciaridae including paternal genome elimination, GRCs, sex determination by chromosome elimination, and monogenic reproduction (Benatti et al., 2010). It is thus tempting
to speculate that GRCs may have spread throughout Sciaridae via similar introgression events, and that this may have also facilitated the spread of monogenic reproduction.

Most of our knowledge about sex determination in Sciaridae comes from the study of several closely-related *Bradysia* species (Metz 1938), though the more early-diverging *Trichosia splendens* is also known to share the strange genetic features of *Bradysia* (Fuge 1994). These diverged sciarid genera have the same X-linked Muller Elements (A and E), although *Phytosciara flavipes* has X-linked portions of other elements which indicates there may be some different derived states (Anderson et al. 2022). It will thus be important to survey a wider sample of sciarid species to obtain a more comprehensive understanding of sex determination and sex chromosome evolution in this clade. Nonetheless, if female-determining inversions were to repeatedly evolve, individuals lacking inversions would be selected to increase their male production as an evolutionary response, with the expected result being that the X’X genotype is maintained at 50% in the population by frequency-dependent selection. However, the fact that we observe X’ degeneration could mean that X’X females will, over time, have reduced fitness, which should favour the invasion of individuals capable of digenic reproduction, unless occasional recombination keeps the X’ from degrading significantly or DC mechanisms are able to evolve. Future work on the role of GRCs in sciarid sex determination, and on the relationship between the unusual genetic aspects across different sciarid species, will be required to elucidate the origins and turnover of sex determinations strategies in this clade.

2.5. Materials and methods

2.5.1. Data collection

The *Bradysia* (formerly *Sciara*) *coprophila* strain used in this study was obtained from the *Sciara* stock centre at Brown University (https://sites.brown.edu/sciara/). Data was produced at Edinburgh University and the Carnegie Institution for Science in Baltimore. At Edinburgh, DNA was extracted using the Qiagen DNeasy Blood and Tissue Kit, modified for high-molecular weight (HMW) extractions. DNA from 50-60 X’X heads (i.e. soma) was used for sequencing on the Illumina NovaSeq S1 platform for paired-end 150bp reads with 350bp inserts; DNA from 30-40
X'X heads was used for PacBio Single-Molecular, Real-Time long-read sequencing. DNA samples were quantified using the Qubit (ThermoFisher). HiC data was sequenced from 50 whole X'X females, which were ground using a DiagoCine Powermasher II with a Biomasher II attachment; libraries were prepared and sequenced by Science for Life Laboratory in Stockholm, Sweden. Illumina data from males (X0) previously generated for (Hodson et al., 2022) was used for the k-mer analysis (see below). At Carnegie, DNA was extracted using DNAzol (ThermoFisher) from 20-36 pooled whole-body individuals, two replicates per genotype (X'X, XX, X0), quantified with the Qubit, analysed for purity with Nanodrop (ThermoFisher), analysed for HMW integrity with 0.5% agarose gel electrophoresis, prepared for sequencing using Illumina Nextera reagents, and was sequenced on the Illumina NextSeq platform to generate 75bp paired-end reads of ~150-400bp fragments. X0 PacBio data was from male embryos, and was generated previously as part of assembling the original somatic reference genome (Urban et al., 2021). X0 HiC data was from male pupae, and was generated previously for chromosome-scale scaffolding of the somatic reference genome (Urban et al., 2022). Illumina and HiC reads were adapter- and quality-trimmed using fastp v0.2.1 (Chen et al., 2018) and quality was assessed before and after trimming using FastQC (Andrews, 2010).

2.5.2. Analysis of X-X’ divergence and evolutionary strata

To identify SVs, 75bp paired-end Illumina reads and PacBio reads from X’X and X0 samples were aligned to the X0 reference genome (Urban et al., 2022) with BWA-MEM (Li, 2013) to force-map X’ reads to the X chromosome. SAMtools (Li et al., 2009) was used to sort, merge and index BAM files prior to calling SVs from Illumina alignments with Smoove v0.2.8 (Layer et al., 2014; Pedersen et al., 2020) and PacBio alignments with Sniﬄes v2.0 (Smolka et al., 2022). svtools (Larson et al., 2019) was used to convert variant ﬁles to bedpe ﬁles. To target fixed variants between X and X’, at least 4 reads were required to support a variant call. The R/Bioconductor (R Core Team, 2023) package Sushi (Phanstiel et al., 2014) was used to plot SVs. HiC reads were aligned to the reference genome using Juicer (Durand et al., 2016), and HiC contact heatmaps were produced using the script HiC_view.py (Mackintosh et al., 2022). To call SNVs, the Illumina alignments were processed with Picardtools (‘Picard toolkit’, 2019) before calling variants with the GATK-4 best practices pipeline (McKenna et al., 2010; DePristo et al., 2011). The scripts
parseVCF.py and popgenWindows.py (https://github.com/simonhmartin/genomics_general) were then used to parse the variant files and calculate the density of SNVs (i.e. heterozygous sites) across 100 Kb windows, respectively. The R (R Core Team, 2023) change-point package bcp (Erdman and Emerson, 2008) was used to identify breakpoints between putative evolutionary strata.

Two methods were employed to estimate the ages of strata. We assumed a neutral mutation rate and a similar mutation rate to *Drosophila melanogaster*, i.e. between 2.8 x 10⁻⁹ (Keightley et al., 2014) and 4.9 x 10⁻⁹ (Assaf et al., 2017), and a 24-40 day generation time for *B. coprophila* (Supplementary Text 2.5). First, the density of heterozygous sites (i.e. number of heterozygous sites divided by the number of homozygous and heterozygous sites) across 100 Kb windows in each stratum (see above) was taken as a proxy for \( D_{xy} \). Generations since divergence were calculated as \( D_{xy}/2*r \), where \( r \) is mutation rate. Second, we targeted only synonymous (neutral) mutations. SnpEff (Cingolani, Platts, et al., 2012) and SnpSift (Cingolani, Patel, et al., 2012) were used to annotate variants and count synonymous variants per gene. The script partitionCDS.py (Mackintosh et al., 2022) was used to annotate degeneracy for all genic sites. Divergence in generations for single-copy homologs within each stratum was then calculated as \( V/(2*r*S) \), where \( V \) is the number of synonymous variants, \( r \) is mutation rate, and \( S \) is the number of synonymous sites.

### 2.5.3 Genome assembly and annotation

A genome assembly was initially generated *de novo* from PacBio reads from X’X females, but only around 3.6Mb of sequence from this assembly could be assigned to the X’ inversion because high sequence similarity between X and X’ chromosomes and high read error rates (~11-16% on average) caused them to collapse upon assembly (Supplementary Text 2.2, Supplementary Figure 2.1). To assemble the X’ inversion, putative X’-specific 27-mers were identified using KAT (Mapleson et al., 2016), counted using KMC3 (Kokot et al., 2017) and FASTA files of 27-mers were obtained using custom python scripts (Hodson et al., 2022). Cookiecutter (Starostina et al., 2015) was used to bin X’ and non- X’ (A + X) reads from X’X Illumina reads, whereby read are binned if they contain k-mers from a given k-mer library. The X’-specific reads were assembled using SPAdes (Bankevich et al., 2012) to yield 61.77Mb spanning 42564 contigs
with an N50 of 10.32Kb and a largest contig length of 86.95Kb. These contigs were then stitched together with the reference-based scaffold RagTag (Alonge et al., 2021) using the regular X chromosome as a reference (Urban et al. 2022) into a single 52.43Mb scaffold containing 9409 gaps. The X' scaffold was then combined with the male X0 genome to produce an X'X genome. PacBio long reads from an X'X female were mapped to the X'X genome with minimap2 and gaps were plugged using Racon (Vaser et al., 2017). Illumina reads from an X'X female were subsequently mapped to the X'X genome with minimap2 (Li, 2018, p. 2) and Racon (Vaser et al., 2017) was used to polish the assembly. After gap-plugging and polishing the X' inversion scaffold totalled 54.75Mb and contained 3486 gaps.

The genome was soft-masked prior to annotation. To this end, de novo repeat families were created using RepeatModeler v2.0 (Flynn et al., 2020), which were combined with known dipteran repeat families from RepBase (Bao et al., 2015). RepeatMasker v4.0 (Smit et al., 2015) was used to soft-mask the genome, which was then annotated using BRAKER2 (Stanke et al., 2008; Lomsadze et al., 2014; Hoff et al., 2016, 2019; Brûna et al., 2021) with RNAseq alignments and homology-based datasets (Supplementary Text 2.6). Functional information for the 26887 protein sequences in the resulting BRAKER2 gene annotation set was then obtained by finding the best BLASTP (Altschul et al., 1990) hits in several protein databases and with InterProScan (Jones et al., 2014, Supplementary Text 2.5). The reasonaTE module of the TranposonUltimate v1.03 pipeline (Riehl et al., 2022) was used to annotate TEs in the genome (Supplementary Text 2.7).

2.5.4. Analysis of functional degradation

Homologous X-X' gene copies were identified with OrthoFinder (Emms and Kelly, 2019). To identify disrupted X'-linked genes, genotyped variant files for X'X and X0 75bp paired-end Illumina reads mapped against the Bcop_v2 (X, II, III, IV) genome (as described above) were filtered for SNVs and indels using SnpEff (Cingolani, Plattls, et al., 2012) and SnpSift (Cingolani, Patel, et al., 2012). SNV and indel counts for 2321 single-copy homologs were then analysed in RStudio (R Core Team, 2023). Two replicates of each genotype (X'X and X0) were aligned, and only consensus variant calls between replicates were considered to identify fixed differences between X and X' and exclude low-frequency variants. Variants specific to the X' were identified
by excluding common variants also found with X0 male data, which may represent polymorphisms on the X or errors in the X chromosome reference sequence (Urban et al., 2022). Genes with loss or gain of start or stop codons or indels causing frameshifts were counted as disrupted.

To identify silenced genes, RNAseq reads were binned as X or X’ in origin using elements of a pipeline developed for (Marshall et al., 2020) to avoid mismapping between the two chromosomes (Supplementary Text 2.8). Expression of the 2321 single-copy homologs was quantified using Kallisto (Bray et al., 2016) and counts were normalised with EdgeR (Robinson et al., 2010). Genes with zero counts of TPM (transcripts per million), and those with TPMs in the bottom 0.1% of non-zero TPM counts within a sample (to account for stochastic mismapping of RNAseq reads) were assumed to be non-expressed. X’ genes were counted as ‘silenced’ if the X copy was expressed but the X’ copy was not, and if genes were both silenced and contained pseudogenising mutations we counted them as ‘silenced and disrupted’. Analysis and plotting of counts was carried out with R Studio (R Core Team, 2023).

To examine DC of X genes that have corresponding degraded X’ homologs, the same pipeline (Supplementary Text 2.8) was used to compare expression of X-linked genes between X’X and XX females (Supplementary Text 2.9). In X’X females, X-linked genes with degraded X’ homologs should have upregulated expression if they are dosage compensated to match expression in XX females. Upregulated genes were identified as those with a log2FC (fold-change) of greater than 0.5.

2.6. References


Chapter III. Why put all your eggs in one basket?

Evolutionary perspectives on the origins of monogenic reproduction

3.1. Abstract

Sexual reproduction is ubiquitous in eukaryotes, but the mechanisms by which sex is determined are diverse and undergo rapid turnovers in short evolutionary timescales. Usually, an embryo's sex is fated at the moment of fertilisation, but in rare instances it is the maternal genotype that determines the offspring's sex. These systems are often characterised by mothers producing single sex broods, a phenomenon known as monogeny. Monogenic reproduction is well documented in Hymenoptera (ants, bees and wasps), where it is associated with a eusocial lifestyle. However, it is also known to occur in three families in Diptera (true flies): Sciaridae, Cecidomyiidae, and Calliphoridae. Here we review current knowledge of monogenic reproduction in these dipteran clades. We discuss how this strange reproductive strategy might evolve, and we consider the potential contributions of inbreeding, sex ratio distorters, and polygenic control of the sex ratio. Finally, we provide suggestions on future work to elucidate the origins of this unusual reproductive strategy. We propose that studying these systems will contribute to our understanding of the evolution and turnover of sex determination systems.

3.2. Glossary

**Genetic sex determination (GSD)** – where sex is controlled by the genotype of an individual, typically the zygote, and is determined at the time of fertilisation.

**Monogeny** – systems in which females produce single-sex broods. While all examples of monogeny involve maternal control of sex determination, not all examples of maternal control over sex determination are monogenic (see ‘digeny’).
**Digeny** – systems with maternal control over sex determination in which females produce both sexes.

**Paternal genome elimination (PGE)** – a form of inheritance in which paternal chromosomes are eliminated or silenced, usually in the male germline, and are therefore not transmitted to a male’s offspring.

**Supergene** – a cluster of tightly linked loci that confer a complex phenotype.

**Inbreeding depression** – decreased fitness associated with an increase in homozygosity following mating between closely-related individuals.

**Local mate competition (LMC)** – a phenomenon observed in Hymenoptera whereby male siblings compete for mating opportunities.

**Meiotic drive** – where asymmetric segregation of chromosomes or alleles during meiosis leads their unequal transmission to offspring.

**Supernumerary chromosomes** – additional chromosomes carried by some organisms that are usually dispensable (i.e. not necessary for organism survival).

**Germline-restricted chromosomes (GRCs)** – chromosomes that are only present and passed through the germline, usually because they are eliminated from somatic cells during early development.

### 3.3. Introduction

Sexual reproduction is an ancient feature among eukaryotes and in many cases involves the evolution of two separate sexes: male and female. However, while the downstream gene networks controlling the differential development of the sexes tend to be relatively conserved (Salz, 2011), the upstream mechanisms of sex determination are strikingly diverse and undergo significant transitions over relatively short timescales (Bachtrog et al., 2014). Among animals, most species exhibit genetic sex determination (GSD). GSD mechanisms themselves are diverse and include the male (XY/X0) and female (ZW/Z0) heterogamety, haplodiploidy as well as the hermaphroditic systems that are common throughout the tree of life. In no clade is the diversity of sex determination mechanisms more obvious than in insects, where virtually every known type of sex determination exists (Sanchez, 2008; Bachtrog et al., 2014; Blackmon et al., 2017).
In most systems with GSD, sex is determined by the genotype of the offspring. For example, in X0 systems and some XY systems, including *Drosophila*, the primary signal for sex determination is the X chromosome dose (Erickson and Quintero, 2007). In other XY systems, such as the housefly *Musca domestica*, it is a Y-linked male-determining factor (Hediger *et al*., 1998). Likewise, in Z0 or ZW systems, such as moths and butterflies (Lepidoptera), sex can be determined by Z dosage (Sahara *et al*., 2012) or W-linked female-determining factors (Kiuchi *et al*., 2014). In haplodiploid sex determination systems, haploid males and diploid females develop from unfertilised and fertilised eggs, respectively (Evans, 2004). In rare instances however, an individual’s sex can be fated by the genotype of the mother instead of that of the offspring. This phenomenon, sometimes referred to as ‘sex predetermination’ (Ullerich, 1980; Nigro *et al*., 2007) is often characterised by females producing single-sex broods. In other words, mothers are genetically predetermined to produce a particular sex ratio, and those that produce predominantly or exclusively male offspring are genotypically distinct from those that produce predominantly or exclusively female offspring. When mothers specialise in producing only one sex, it is referred to as monogenic reproduction (Metz, 1938).

Monogeny, or split-sex ratios, occurs in over 20 different eusocial genera of Hymenoptera (Meunier *et al*., 2008), where colonies specialise in producing either male (drones) or female reproductives (queens), although both types of colonies also produce female workers. In the ant species *Formica glacialis*, monogeny is associated with a 5.5Mb supergene that occurs exclusively in females in a heterozygous state, causing them to produce queens (Lagunas-Robles *et al*., 2021). Some parasitoid wasps are also known to produce single-sex broods, and this is controlled by multiple factors including host size, temperature, local mate competition, diet and maternal genotype (for a review see King, 1987). Outside of Hymenoptera, monogeny is reported in three dipteran families, all of which have a solitary lifestyle and which include pests of agricultural significance (Hall *et al*., 2012; Shin *et al*., 2013; Scott *et al*., 2014, **Figure 3.1A**): the dark-winged fungus gnats (Sciaridae, henceforth ‘fungus gnats’), the gall midges (Cecidomyiidae) and the blowflies (Calliphoridae).
Figure 3.1. (A) Monogenic reproduction is reported in three families of Diptera: dark-winged fungus gnats (Sciaridae), gall midges (Cecidomyiidae), and blowflies (Calliphoridae). (B) Sex determination via postzygotic chromosome elimination in fungus gnats and gall midges. The mechanism of sex determination is similar in the two families, with the subtle distinction that zygotes in fungus gnats are triploid because sperm provide two X chromosomes. One extra chromosome is eliminated from embryos as a result. Photo credit: J Niland (C. rufifacies); S Bauer (M. destructor); RB Baird (B. coprophila).

Fungus gnats and gall midges are both members of the superfamily Sciaroidea, and are themselves large families comprising over 5000 and 2000 described species respectively (Skuhrava, 2006; Shin et al., 2013). Several fungus gnat species are synanthropic and are receptive to being cultured in laboratory conditions, and as such they are the most well-studied of the three clades in terms of their genetics and sex determination (for reviews see Sánchez, 2010; Gerbi, 2022). Fungus gnats and gall midges have a non-Mendelian inheritance system called paternal genome elimination (PGE), a phenomenon in which the paternal autosome and
sex chromosome copies are lost during the meiotic divisions of the spermatocytes and are therefore not transmitted to a male's offspring. They also determine sex via elimination of paternal X chromosomes during early embryogenesis, though the exact molecular mechanism by which this is governed remains unknown (Stuart and Hatchett, 1991; Gerbi, 2022). The monogenic blowflies are, in comparison, relatively understudied and very little is presently known about their sex determination (Scott et al., 2014). Moreover, the evolutionary origins of monogenic reproduction, and the precise mechanisms by which it occurs, remain unknown in all three clades. Elucidating how this strange reproductive strategy evolves may help our understanding of how and why some systems depart from classical Mendelian inheritance.

Here we review current knowledge of monogenic reproduction in the three dipteran clades in which it is known to exist. We choose to focus explicitly on these dipteran examples, as there is already an extensive literature on the phenomena in the Hymenoptera (Herre 1985; Greeff 1996; Meunier et al., 2008). In the literature, female fungus gnats are referred to as gynogenic if they are female-producers and androgenic if they are male-producers (e.g. Sánchez, 2010), whereas in gall midges and blowflies they are referred to as thelygenic and arrhenogenic females, respectively (e.g. Stuart and Hatchett, 1991; Scott et al., 2014). For the purpose of simplicity, for all species we will refer to female-producers as gynogenic and male-producers as androgenic. After reviewing the three clades, we discuss evolutionary forces that may drive transitions to monogeny.

3.4. Fungus gnats (SCIARIDAE)

Fungus gnats have been studied since the 1920s (Metz, 1925) and their complicated system of chromosome inheritance has long been appreciated. The majority of our knowledge comes from the study of the closely related species Bradysia coprophila, B. impatiens and B. ocellaris, though the more distant Trichosia splendens has also been studied and shares many of the unusual features found in Bradysia (Metz, 1938; Carson, 1946; Amabis et al., 1979; Fuge, 1994). Their chromosome cycle involves three rounds of PGE, one of which occurs during embryonic cleavage divisions 7-9 and is the moment when sex is determined. Unusually, fungus gnat zygotes begin with three X chromosomes. This is a result of asymmetric segregation of the X
chromosome in male meiosis II, which gives rise to XX sperm. As a result, either one or two paternally derived X chromosomes are eliminated from the embryo, which initiates female (XX) or male (X0) development, respectively (Figure 3.1B). The chromosomes bound for elimination fail to divide at anaphase and are left behind on the metaphase plate, though the precise mechanism by which this elimination is controlled is unknown (DuBois, 1933).

Not all fungus gnat species are strictly monogenic. Some are described as digenic, meaning they produce mixed-sex broods, though progeny sex ratios are highly variable (Davidheiser 1947). These variable sex ratios are also temperature-dependent, with a higher proportion of females being produced at higher temperatures. This is caused by an increase in female-production at the expense of male-production rather than higher mortality in male embryos (Nigro et al., 2007; Farsani et al., 2013). The temperature-sensitive period of development appears to be the mid-pupal to early adult stages (Nigro et al., 2007), when oogenesis takes place (Berry, 1941).

Digenic and monogenic fungus gnats determine sex via the same mechanism of paternal X elimination during embryogenesis (DuBois 1933; Perondini et al., 1986). Monogenicity is also known to be associated with chromosomal inversions in B. coprophila and B. impatiens, and in both cases these inversions are X-linked. The affected chromosome is termed the X’ (prime) chromosome. Gynogenic females are heterozygous for this chromosome and transmit it to half of their offspring (Carson, 1946; Crouse, 1960, 1979). The inverted portion is paracentric, and is terminal in B. impatiens (Carson, 1946) but in B. coprophila it lies in the middle of the left arm of the X; it is not known whether the inversions occurred prior to divergence between the two species or evolved independently. It was recently found that the X’ in B. coprophila appears to carry a supergene of multiple, linked inversions that span ~55Mb of the ~67Mb chromosome and emerged less than 0.5 mya (Chapter II, Baird et al., 2023). Presumably, the X’ contains the locus or loci that results in one paternal X being retained in the embryos of X’X females, while those of XX females eliminate both paternal X chromosomes. Maternally produced factors are postulated to mediate X elimination by recognising an X-linked element. This ‘controlling element’ (CE) has been localised to the short right arm of the X. Rather than a control site, the CE likely acts as a recognition site for X elimination: if translocated to an autosome, the receiving autosome is instead eliminated (Crouse, 1960, 1979; de Saint Phalle and Sullivan, 1996). The X’ inversions in these monogenic species prevent homologous pairing
and recombination with the X chromosome, preserving the maternal factors responsible for X elimination (Metz, 1938).

Monogenic and digenic reproductive strategies are reported to exist within several distinct fungus gnat genera, including Bradysia, Lycoriella, Scatopsciara, and Corynoptera. Some species, such as B. ocellaris, are reported to have both monogenic and digenic strains (Metz 1938, *Supplementary Table 3.1*). Furthermore, we recently found that the X’ chromosome of B. coprophila evolved as recently as <0.5 mya (Chapter II, Baird et al., 2023). Taken together, these observations indicate that monogeny may have evolved repeatedly within the fungus gnat family, which suggests that this reproductive strategy may confer some selective advantage. The factors that drive turnover between digenic and monogenic reproduction will be discussed below.

### 3.5. Gall midges (CECIDOMYIIDAE)

Gall midges represent one of the most species-diverse families of flies, comprising over 5000 known species (Skuhrava, 2006; Dorchin et al., 2019). They are relatively closely related to fungus gnats; both are thought to have originated from the more primitive family Mycetophilidae. Gall midges exhibit a range of unusual reproductive strategies. Some genera of the more early-diverging subfamilies Heteropezinae and Lestemiinae reproduce via larval or pupal pedogenesis (a type of cyclic parthenogenesis involving asexual reproduction by immature insects), though the majority of species reproduce sexually (White, 1973). While the chromosome cycles in Mycetophilidae are orthodox, those of gall midges, like in fungus gnats, involve several rounds of maternally-controlled elimination of paternal chromosomes, including loss of the paternal homologs during spermatogenesis (White, 1973). *Mayetiola destructor* is the most well-characterised cecidomyiid in terms of sex determination and chromosome inheritance (Stuart and Hatchett, 1991). This species has two pairs of nonhomologous sex chromosomes, X1 and X2. All zygotes begin with the same chromosome constitution, X1Mx2Mx1Px2P (X_M = maternally derived; X_P = paternally-derived), following the fusion of X1X2-bearing eggs and X1X2-bearing sperm. Like in fungus gnats, sex is determined when a round of X chromosome elimination occurs during the early cleavage divisions (*Figure 3.1B*).
Embryos that lose the paternal set develop into males (X1M X2M 00); those that retain their X chromosomes develop into females (X1M X2M X1P X2P). X elimination is presumably governed by maternally deposited factors in the early embryo, although this has not been confirmed.

While gall midge species from various genera have been documented as strictly monogenic, some species exhibit both monogenic and digenic reproductive strategies. (Supplementary table 3.2). The model species M. destructor is one example of a species with both monogenic and digenic females. The mechanism of sex determination via X elimination is the same in M. destructor embryos regardless of whether the broods are single- or mixed-sex. In this species, gynogenic and androgenic females are distinguished by an autosomal inversion, for which gynogenic females are heterozygous. Because the inversion is present only in female-producers, it is only ever found in females and is inherited by half of the offspring in a regular Mendelian fashion, such that an equal ratio of gynogenic and androgenic females are produced (Stuart and Hatchett, 1991). Presumably, the inversion contains one or more loci that repress X elimination in the embryo, while acting to suppress recombination and prevent the transfer of the locus or loci to the homologous autosome. The inversion spans approximately 2Mb, corresponding to around 1.3% of the haploid genome (Benatti et al., 2010; Vellacott-Ford, 2020). Some populations of M. destructor also have a second, ~3Mb nonoverlapping inversion present only when the first inversion is also present in cis. No recombination has been observed between the two inversions, suggesting that the second may have been selected for because it further suppresses recombination along the chromosome (Benatti et al., 2010).

Despite being separated by over 147 million years of evolution and several intermediate families (Ševčík et al., 2016; Hodson et al., 2022), Sciaridae and Cecidomyiidae share many features including monogeny, a near-identical chromosome cycle with PGE, as well the presence of germline-restricted chromosomes (GRCs) which are eliminated from somatic cells in early development. It is therefore tempting to speculate on a common origin for some of the features of these two clades. It was recently discovered that the GRCs in B. coprophila share little homology with the core chromosomes of their host species, but rather are closer in sequence similarity to the core genome of M. destructor, likely being acquired by fungus gnats from gall midges via introgression between 114 and 50 mya (Hodson et al., 2022). If PGE did not evolve independently in the two lineages then it must have either been lost in the intermediate families, or otherwise perhaps also transferred through introgression. Moreover, if monogenic
reproduction has evolved repeatedly in the fungus gnats then it may have also done so in the gall midges, though phylogenetic information on gall midges with different reproductive strategies is lacking.

3.6. Blowflies (CALLIPHORIDAE)

Among the blowflies, only two species, Chrysomya rufifacies and C. albiceps, have been described as monogenic (Wilton, 1954; Ullerich, 1958); other members of this genus have male heterogamety with differentiated X and Y sex chromosomes. In the genera Lucilia and Cochliomyia, which lack monogeny, sex is controlled via a male-determining Y factor that initiates autoregulatory splicing of the sex determination cascade gene transformer (Concha and Scott, 2009; Li et al., 2013). In the non-monogenic Chrysomya species C. chlorophyga, aberrant X0 and XXY embryos develop into females and males, respectively, which suggests that the ancestral mechanism for monogenic species involves a Y-linked male determining locus (Ullerich, 1976). In contrast, monogenic Chrysomya reportedly have undifferentiated sex chromosomes (Ullerich, 1975). Andere et al. (2020) performed coverage-based assignment of approximately 3.3 and 1.5 Mb worth of sequence to putative X and Y chromosomes, respectively, suggesting that there may be some sex-linked regions, but this requires further work to fully resolve. The mechanism by which sex is determined in monogenic blowflies remains unknown, but it is likely to be fundamentally different from fungus gnats and gall midges since chromosome transmission behaviour appears to be regular, with no reports of PGE or X elimination. Furthermore, fungus gnats and gall midges belong to the lower Diptera (Nematocera) superfamily Sciariodea and are therefore relatively closely related (Ševčík et al., 2016). Blowflies, on the other hand, are higher dipterans (Brachycera), which diverged from lower dipterans ~200 mya (Wiegmann et al., 2011). A common origin for monogenic reproduction between the three families can thus be confidently ruled out.

Gynogenic female blowflies produce androgenic and gynogenic female offspring at a 1:1 ratio, and therefore the gynogenic females are thought to be heterozygous for a dominant allele that is inherited in a regular Mendelian fashion and pre-determines female sex in their offspring (Ullerich 1996, Figure 3.2). Transplantation of ovaries and pole cells between androgenic and
gynogenic females revealed that this sex-determining factor is synthesised by the germline during early oogenesis and maternally deposited in the embryo (Ullerich, 1980, 1984). Studies of the inheritance of various genetic markers showed incomplete linkage between the markers and the locus that pre-determines offspring sex (Ullerich, 1996), and translocation experiments revealed that it is situated on the proximal half of the long arm of chromosome 5 (Ullerich, 1975). However, it has not yet been demonstrated whether there is recombination suppression between chromosome regions of gynogenic females, like there is in B. coprophila and M. destructor, and cytogenetic analysis of the polytene chromosomes have not revealed any obvious chromosomal rearrangements (Puchalla, 1994). The genomes of male and androgenic and gynogenic female C. rufifacies were recently published (Andere et al., 2020), though their poor contiguity (>100,000 contigs per genome) makes identifying the control locus in gynogenic females challenging, particularly if sex is under the control of a small genomic region.

Previously, transformer (tra) has been proposed as a candidate for the sex determining locus in monogenic Chrysomya (Scott et al., 2014). Tra is one of a set of genes in a conserved cascade that regulates sexual development in many insects (Hopkins and Kopp, 2021). Interestingly, mutant housefly, M. domestica, females that lack tra default to male-production (Hediger et al., 2010). If such a transition were favoured in populations of Chrysomya, the
resulting male-biased population sex ratio might subsequently drive the evolution of maternally-acting factors that cause female production. Identification and characterisation of tra in monogenic Chrysomya and its potential role in sex determination remains to be investigated.

3.7. How does monogeny evolve?

Why would mothers evolve to produce single sex broods? Attempts to explain the split sex ratios found in Hymenoptera focus mainly on kin selection, (Meunier et al., 2008; Kobayashi et al., 2013), inbreeding, and local mate competition in the context of their eusocial lifestyle (Herre, 1985; Greeff, 1996; Schrempf et al., 2006). In social Hymenoptera, diploid females and haploid males are produced from fertilised and unfertilised eggs, respectively. This results in workers being more closely related to sisters than to brothers, though this depends on the number of queens in the colony: workers are more closely related to one another when there are fewer queens. Kin selection predicts that workers should favour the production of females where relatedness between workers is higher, and that when relatedness is lower, more males should be produced (Boomsma and Grafen, 1990). Empirical results indeed show this to be the case (Meunier et al., 2008). Previous hypotheses for the evolution of monogeny in fungus gnats and gall midges have focused on inbreeding depression and conflict over the sex ratio, respectively (Haig, 1993; Tabadkani et al., 2011). We elaborate on both below, and suggest an additional hypothesis based on multi-locus control of the sex ratio that has some support from previous studies of fungus gnats and gall midges.

3.7.1. Inbreeding depression

Inbreeding, e.g. mating between siblings, is widespread in natural populations of animals (Lacy, 1993). Inbreeding increases homozygosity which leads to phenotypic expression of deleterious recessive mutations and resulting fitness costs (Pusey and Wolf, 1996; Crnokrak and Roff, 1999; Keller, 2002; Mongue et al., 2016). These effects have been found to result in the evolution of
diverse inbreeding avoidance mechanisms such as dispersal (Szulkin and Sheldon, 2008), intentional avoidance of kin (Facon et al., 2006) and polyandry (Firman and Simmons, 2008).

Monogenic reproduction has been suggested as an alternative mechanism for inbreeding avoidance (Tabadkani et al., 2011; Andere et al., 2020). A consequence of monogeny is that mating between siblings is impossible because the progeny in any one brood are of the same sex. Offspring must therefore disperse in order to mate, which will at worst result in mating with half-siblings. Simulations suggest that monogeny provides a potentially effective route to inbreeding avoidance, particularly when populations are small (Tabadkani et al., 2011), though empirical evidence to support this is lacking.

3.7.2. Sex ratio selection

Fisherian sex ratio theory posits that in a large, randomly mating population, frequency-dependent selection should result in a 1:1 male:female sex ratio (Fisher, 1930). However, there are circumstances where biased sex ratios can be advantageous. Probably the most frequent scenario is local mate competition (LMC), which occurs when matings frequently occur between close relatives. Under the most extreme scenario where matings occur between full-sibs, extremely female-biased sex ratios are selected for. Generally, in species with frequent LMC, mothers are able to facultatively adjust their brood sex ratio relative to the expected degree of sibmating. However, experimental evolution studies in mites show that this sex ratio strategy can be genetically determined (Macke et al., 2014). While LMC can drive the evolution of a female-producing strategy, it is difficult to envisage how a male-producing strategy can evolve, because in the absence of LMC mothers should produce equal, not male-biased sex ratios (West, 2009).

Sex ratio distortions can also arise where a particular sex, chromosome, or endosymbiont favours the production of one sex over another (Sandler et al., 1959; Jones, 1991; Hurst, 1993). One scenario that can lead to biased sex ratios is sex chromosome meiotic drive, where the transmission of one sex chromosome is favoured over the other (Jaenike, 2001; Lindholm et al., 2016). For example, Gershenson (1928) showed that an X-linked factor in male D. obscura kills Y-bearing sperm, resulting in a female-biased sex ratio. In some cases, autosomal segregation distorters are also known to cause sex ratio distortions (Larracuent
Presgraves, 2012). Meiotic drive and segregation distortion is well-studied in *Drosophila* (Courret *et al.*, 2019) and also occurs in other Diptera (Wood and Newton, 1991; Fry and Wilkinson, 2004). Unlike with female-bias caused by LMC, significant departures from an even population sex ratio that occur due to drive may provide a selective advantage to parents who are able to specialise in producing the rarer sex, such that the population sex ratio returns to 1:1. Haig (1993) suggested that a driving X chromosome arising in a fungus gnat ancestor initiated the evolution of its strange chromosome cycle. Following the female-biased sex ratio that results from X-drive, mothers began converting XX daughters into X0 sons by eliminating a paternal X in the embryo, and ensuing conflict over the sex ratio ended with mothers specialising in the production of a particular sex.

The presence of supernumerary chromosomes can also cause departures from an even sex ratio. For example, B chromosomes found in many species favour the production of individuals in which they are carried, and are able to bias the sex ratio through association with nuclear transmitted segregation distorters (Jones and Rees, 1982). B chromosomes drive male-biased sex ratios in a variety of systems including fairy shrimp *Branchipus schaeferi* (Beladjal *et al.*, 2002), the teleost fish *Astyanax scabripinnis* (Vicente *et al.*, 1996) and the wasp *Nasonia vitripennis* (Nur *et al.*, 1988). Supernumerary chromosomes that somewhat resemble B chromosomes (GRCs, or germline restricted chromosomes) are found in gall midges and fungus gnats. The GRCs are eliminated from somatic cells early in development of both sexes, but are retained in the ovaries or testes (Hodson and Ross, 2021). Haig (1993) noted that, because GRCs in *B. coprophila* are disproportionately transmitted by males, they should favour male-biased sex ratios. The GRCs would have thus favoured the conversion of XX daughters into sons by mothers, which may have spurred the evolution of the X’ chromosome that suppressed the actions of the GRCs. In contrast, GRCs in gall midges are exclusively transmitted through females and should therefore favour female-production. The function of GRCs in fungus gnats and gall midges, and whether they have any effect on sex determination, remains to be explored.

Maternally inherited microorganisms present another route by which conflict over the sex ratio can arise. *Wolbachia* are common reproductive parasites, and the feminisation, parthenogenesis, male-killing and cytoplasmic incompatibility that they induce is well documented in insects (Werren *et al.*, 2008). *Wolbachia* are found in blowflies (Mingchay *et al.*, 2014; Xu *et al.*, 2022), including the monogenic *C. albiceps* (Şaki and Şimşek, 2014), and have
been suggested as a mechanism for biological pest control for members of this family (Caleffe et al., 2019). A *Rickettsia* genome was sequenced along with the recently sequenced *B. coprophila* genome (Urban et al., 2021). *Rickettsia* are a group of proteobacterial endosymbionts related to *Wolbachia* that are also known to exhibit meiotic drive behaviour (Werren et al., 1994; Lawson et al., 2001; Giorgini et al., 2010), providing another potential mechanism that may have favoured the evolution of these monogenic systems.

### 3.7.3. Polygenic control of the sex ratio

Within the digenic (mixed-sex brood producers) fungus gnats and gall midges, significant departures from a 1:1 progeny sex ratio are the norm (Davidheiser, 1947; Mcclay, 1996; Nigro et al., 2007), and some species are described as exhibiting mixed (both monogenic and digenic) strategies (McCarthy, 1945; Steffan, 1974; Stuart and Hatchett, 1991; Rocha and Perondini, 2000, Supplementary tables 3.1 and 3.2). Sex ratios in these families appear to exist along a continuum, with extreme sex ratios (i.e. monogeny) fixed in some species (in blowflies variable sex ratios have not been reported). Even strictly monogenic females of *B. coprophila* do occasionally produce ‘exceptional’ offspring of the wrong sex, showing that the capacity for producing both sexes is retained in females of monogenic species (Metz and Schmuck, 1929). Metz (1938) originally suggested that in *B. coprophila* the X and X' are distinguished not by a single allele but rather a series of alleles of varying potency, and that the difference between monogenic and digenic species is the ‘strength’ of the X' chromosome.

It is now known that the X' in *B. coprophila* is distinguished from its X homolog by a large region of recombination suppression composed of inversions (Crouse, 1979; Chapter II, Baird et al., 2023). While all females of digenic species like *B. ocellaris* are XX, there must be something that genotypically distinguishes females that produce male-biased broods from those that produce female-biased broods. Davidheiser (1947) reported that the sex ratio in this species is heritable: the female offspring of female-biased broods and male-biased broods also produce female-biased and male-biased broods, respectively. In the same study it was shown that it was possible to artificially select for predominantly male production from predominantly female production, and vice versa, in only a handful of generations. These observations of (i) continuous variation of this phenotype, (ii) inheritance of the sex ratio and (iii) rapid artificial
selection provide a strong indication that the sex ratio has an additive genetic component. It follows that in this system, particular combinations of alleles at multiple loci may determine the amount of maternally deposited factors in the oocytes, which then affects the proportion of embryos that develop as male or female. Recombination between these loci in digenic lineages produces the different sex ratios observed (Figure 3.3A); their fixation in some lineages leads to monogenic females with single-sex broods (Figure 3.3B). Under this scenario, the term ‘monogenic’ refers to lineages in which the production of single-sex broods has become the dominant strategy in a population. This may occur via recombination suppression via inversions around the dominant female-determining alleles. Alternatively, inversions may occur first and then alleles that modify the sex ratio may migrate to inverted regions. Females without inversions should then evolve more male-biased production as an evolutionary response, with the expectation being that the genotype heterozygous for the inversions is maintained at 50% in the population by frequency-dependent selection.

The distribution of monogenic, digenic and mixed reproductive strategies across fungus gnats suggests multiple evolutionary origins for monogeny within the family, or perhaps frequent reversions to digeny (Supplementary table 3.1). This is further supported by the finding that the X' chromosome in B. coprophila emerged only <0.5mya. Turnover between monogenic and digenic reproduction may therefore be common. Since the non-recombining region degenerates over time (Chapter II, Baird et al., 2023), decreased fitness in X'X females may present opportunities for females that produce mixed-sex progenies to invade, resulting in reversions to digeny (Figure 3.3C).

Why would monogeny evolve under this scenario? Skewed sex ratios in digenic species may be evidence of divergent selection acting on the sex ratio. If control of the sex ratio is indeed polygenic, an initial distorting driver may not be required. Sex determination systems that are under polygenic control are thought to be inherently unstable, because if one sex-determining locus provides a fitness benefit over others then that locus should eventually fix as the sole sex-determiner (Rice, 1986). Instability of polygenic sex determination is also thought to be exacerbated in small populations where it is more likely to produce skewed sex ratios and where rarer alleles may be lost more frequently by drift or selection (Bateman and Anholt, 2017). In monogenic populations with a non-recombining X', the trapped sex ratio alleles act as a single locus which may resolve instability in digenic populations.
Figure 3.3. A model for the evolution of monogenic from digenic reproduction in fungus gnats (also applies to gall midges, though inversions in the gall midge *M. destructor* are autosomal).

(A) Different combinations of X-linked alleles are responsible for variable sex ratios among digenic females. ‘Male-determining’ alleles should result in X elimination and ‘female-determining’ alleles should result in X retention. (B) The ‘trapping’ of female-determining alleles through recombination suppression (e.g. inversions) leads to the fixation of monogenic reproduction in a population. Alternatively, inversions may occur first, onto which female-determining alleles migrate. (C) Non-recombining X’ chromosomes degenerate and their carriers suffer reduced fitness. Individuals with X chromosomes bearing female-determining alleles reinvade the population, spurring turnover in reproductive strategy.
3.7. Conclusions and future perspectives

Monogenic reproduction is one of the most unusual forms of genetic sex determination and its origins remain elusive. In the blowflies, too little is currently known about sex determination in the *Chrysomya* genus to speculate on the origins of monogeny in this clade. It will be essential to first determine what distinguishes androgenic and gynogenic females genotypically in monogenic *Chrysomya*, as well as to characterise the sex determination systems and sex ratios in non-monogenic members of the genus.

As for fungus gnats and gall midges, the relationship between digenic and monogenic reproduction is unclear and the question of multiple independent origins for monogeny remains open. To answer these questions, it will be essential to sequence and compare the genomes of different monogenic species within the families, especially species that are more distantly related to the *Bradysia* models studied thus far. In particular, the development of chromosome-level assemblies will aid in comparative genomics and identification of inversions associated with monogeny. The closely related *B. coprophila* and *B. impatiens* both harbour X' chromosomes that are slightly different in structure. Their relationship is unknown, but it could provide an indication as to whether and how monogeny evolves repeatedly. It will also be necessary to uncover the molecular control of the sex ratio in monogenic and digenic species, and to determine the role of the GRCs, if any, in sex determination. Lab colonies of *B. coprophila* have been maintained since the 1920s (Moses and Metz, 1928), and more recently, colonies of other species including digenic *Lycoriella ingenua* have been established (RB Baird and L Ross, unpublished studies). The genome of the model *B. coprophila* is now available (Urban et al., 2021), and site-specific insertions of DNA as well as piggyBac-mediated transformation techniques have recently been developed for this species (Yamamoto et al., 2015; Yamamoto and Gerbi, 2022); as such there are increasingly available opportunities to understand more about this system.

Moreover, further work is required to determine the selective forces that drive transitions to monogeny. Inbreeding may appear an unlikely explanation, since inbreeding is widespread (Lacy, 1993) while monogeny is not. As for resolution of sex ratio distortions, evolving monogeny from the ancestral sex determination system may be more difficult than
simply evolving suppressors of drive (Atlan et al., 2003). However, intrinsic properties of fungus gnats and gall midges may mean that they are more amenable to evolving monogeny. Since non-monogenic members of these families already have variable progeny sex ratios, the transition to extreme sex ratios may be a relatively straightforward solution to suppress drive or resolve inbreeding depression. Furthermore, sex ratios in digenic fungus gnats may be face instability due to the temperature effect on their progeny sex ratios: environmental sex determination is thought to be unstable in the face of environmental perturbations which destabilise sex ratios, giving rise to GSD (Van Dooren & Leimar, 2007). In contrast, blowflies may represent a more major transition from an XY system to monogeny, which might require a stronger selective pressure to evolve.

Also striking are some of the features that are shared by these systems. For example, monogeny seems to be associated with chromosomal inversions. The study of these systems may therefore broaden our understanding of how inversion-based supergenes are associated with the evolution of complex traits (Schwander et al., 2014). Furthermore, so far it always appears to be the female-producers that are the heterogametic morphs that possess these inversions. If it were instead male-producers that carried heterozygous inversions, then they would need to pass through males which might present opportunities for genetic conflicts between the sexes.

More generally, these systems offer unique opportunities to study the evolution of sex determination systems and sex ratios. The consensus in the literature is that the optimal mean sex ratios that individuals produce should be broods of equal numbers of males and females (Frank, 1990). Producing anything other than 1:1 progeny sex ratios is rare, and understanding why this is the case requires studying systems that deviate from the norm. Together, exploration of the sex determination systems of these peculiar flies may help inform us about how some of the most fundamental mechanisms in evolution - of sex determination systems and sex ratios - evolve.
3.8. References


Sánchez L (2010). Sciara as an experimental model for studies on the evolutionary relationships between the zygotic, maternal and environmental primary signals for sexual development. *J Genet*** **89**: 325–331.


Chapter IV. Molecular control of the sex ratio in dark-winged fungus gnats

4.1. Abstract

Sex is widespread among eukaryotes, yet the mechanisms that govern sex determination are diverse and undergo significant turnover in short evolutionary timescales. In most systems, the sex of offspring is established at the moment of fertilisation. Dark-winged fungus gnats (Diptera: Sciaridae) have an unusual mechanism of sex determination where maternally deposited factors predetermine offspring sex. Sex becomes established when paternal X chromosomes are eliminated during early embryogenesis, but the precise molecular mechanisms governing this are unknown. In some species, mothers produce single-sex broods, allowing collection and sequencing of entire clutches of eggs of known sex. In the present study, I generated sequence data from clutches of male and female early embryos and identified alleles differentially expressed between the maternally deposited transcriptomes of male- and female-producing mothers. I identified a list of candidate sex ratio modifier genes based on their expression patterns and predicted functions. I also find evidence for onset of zygotic gene activation between 4h and 8h after egg deposition. These findings provide a foundation for further exploration of the mechanism of sex determination in dark-winged fungus gnats as well as for their early embryonic development more generally.

4.2. Introduction

Sexual reproduction is an ancient feature among eukaryotes and its origin is regarded as one of the most important transitions to occur in evolutionary history (Szathmary and Smith, 1995). It is perhaps paradoxical, therefore, that the primary signal that initiates sex determination is not evolutionarily conserved, but rather shows extraordinary diversity between species (Bachtrog et al., 2014). In contrast, the gene cascades that regulate sexual development are generally
highly conserved. Among insects, for example, which show immense diversity in their sex determination mechanisms (Blackmon et al., 2017), it is often the same set of genes, namely *Transformer, Doublesex,* and *Fruitless,* that are involved in a splicing-regulation cascade that culminates in sex-specific somatic and germline tissue development (Sanchez, 2008; Hopkins and Kopp, 2021).

In most systems, the process of sex determination is not complete until after fertilisation and the onset of zygotic transcription. For example, in generic XY systems, the sex of an embryo depends on whether the egg fuses with a Y- or X-bearing sperm, and the sex-determining signal is activated following zygotic genome activation (Casper and Van Doren, 2006). In some rare cases, maternal effects determine the fate of embryos prior to any male contribution, and this depends on the maternal genotype. This is known to occur in three families of Diptera (true flies): blowflies (Ullerich, 1958), gall midges (Stuart and Hatchett, 1991), and dark-winged fungus gnats (Metz, 1938). Within these families, there are species within which females produce single-sex broods, a phenomenon termed monogenic reproduction. Gynogenic females produce only daughters and androgenic females produce only sons. In two Sciaridae species, *Bradysia coprophila* and *B. impatiens,* and one Cecidomyiidae species, *Mayetiola destructor,* chromosomal inversions distinguish gynogenic and androgenic females (Carson, 1946; Crouse, 1979; Benatti et al., 2010). These inversions are always present heterozygously in gynogenic females, where they serve to suppress recombination and preserve the factors for production of female progeny. In *Bradysia,* the inversions are X-linked, and the affected chromosome is denoted the X' (Metz, 1938).

While maternal effects predetermine the sex of each embryo in these flies, the process of sex determination begins post-fertilisation when paternal X chromosomes are eliminated from the developing embryo. All zygotes begin with three X chromosomes, one maternally inherited and two paternally-inherited (spermatogenesis is atypical; sperm always contribute two X chromosomes). During the 7th-9th cleavage divisions, either one or two X chromosomes, which are always paternal in origin, are eliminated from the somatic cells of embryos that develop into females or males, respectively. The long arms of the eliminated X chromosomes fail to divide at anaphase and are left behind on the metaphase plate (DuBois, 1933; de Saint Phalle and Sullivan, 1996). The embryos produced by gynogenic females thus become XX and develop into females, while those produced by androgenic females become X0 and develop as
males (Gerbi, 2022). Following X elimination, sexual development is presumably initiated by dosage of an X-linked factor, as in other X0 and many XY systems (Sanchez, 2008). Crouse (1979) identified an element on the short arm of the X which acts as a recognition site for elimination: if the element is translocated to an autosome, the receiving autosome is eliminated (de Saint Phalle and Sullivan, 1996). A maternal imprinting mechanism is thought to protect maternally inherited X chromosomes from elimination while paternally-inherited chromosomes are left exposed, although the mechanism by which this occurs is unknown (Crouse, 1979). Previous hypotheses posit that maternally deposited factors in the egg interact with a recognition element on the X to direct its elimination (de Saint Phalle and Sullivan, 1996; Sanchez and Perondini, 1999).

In monogenic Sciaridae, one or more differences between the X and X’ chromosomes are presumed to determine the fates of embryos laid by gynogenic (X’X) and androgenic (XX) females. In Chapter III and Baird, Mongue, et al., (2023), I discussed why this is likely to be a difference in alleles at more than once locus, rather than a single gene. Briefly, some sciairds are digenic rather than monogenic, meaning they produce mixed-sex broods. Among digenic species, primary sex ratios are highly variable (Rocha and Perondini, 2000; Nigro et al., 2007; Farsani et al., 2013) and appear to have a heritable component (Davidheiser, 1947). Furthermore, early studies suggest that primary sex ratios can be artificially selected towards male-bias or female-bias in only a handful of generations (Davidheiser, 1947). Taken together, these findings implicate multiple sex ratio modifier (SRM) loci in control of primary sex ratios in Sciaridae. The finding that the X’ chromosome is evolutionarily young in B. coprophila suggests that monogeny is probably the derived state (Chapter II, Baird, Urban, et al., 2023). Recombination suppression between SRM loci via chromosomal inversions may have given rise to monogeny in some lineages (Figure 4.1).
**Figure 4.1.** A multi-locus model for control of the sex ratio in sciarid fungus gnats. (A) Mothers have two X chromosome copies on which there are multiple loci that have male-determining and female-determining alleles (sex ratio modifiers). Varying combinations of these alleles give rise to the different sex ratios observed among females. (B) X'X females in monogenic species have recombination suppression between female- and male-determining alleles.

In developing sciarid embryos, males lose two of their three X chromosomes while females lose just one. It follows that the difference in number of chromosomes to be eliminated is likely caused by a difference in dosage or presence of the factor causing elimination. A model has been proposed which posits that a SRM is deposited at a greater dose in female embryos and acts to suppress elimination of one paternal X in female embryos (the SRM-suppressor model, de Saint Phalle and Sullivan, 1996; Sanchez and Perondini, 1999). Alternatively, the SRM could be deposited at a greater dose in male embryos and facilitate X elimination (the SRM-activator model, Figure 4.2B). An important distinction between the two models is that if the SRM suppresses elimination then it should be exclusively produced by gynogenic females, while
if it activates elimination, then it should be produced by both females types but deposited at a double-dose in male embryos.

**Figure 4.2.** Two models for the elimination of paternal X chromosomes in sciarids. (A) In the SRM-suppressor model, the SRM prevents elimination of one paternal X chromosome and is only maternally deposited by gynogenic females. (B) In the SRM-activator model, the SRM facilitates the elimination of paternal X chromosomes, and is present at a higher dose in male embryos.

The set of transcripts supplied by mothers to embryos is referred to as the maternal transcriptome, and they are essential for the early stages of embryonic development until zygotic genome activation (ZGA) takes over (Hamm and Harrison, 2018; Tora and Vincent, 2021). In conventional systems, mothers deposit essentially identical maternal transcriptomes into each embryo, and so there are no sex differences in gene expression during the early embryonic divisions (Lott et al., 2014). However, in systems with maternal control of sex determination, sex is controlled by differences in the maternal transcriptome supplied to each embryo. Fortunately,
the nature of monogenic systems means that embryos of known sex are easily obtainable, and entire clutches of eggs of a single sex remove the need for single-embryo RNAseq and sexing of individual embryos, as is required for *Drosophila melanogaster* (Lott *et al.*, 2011) or *Anopheles gambiae* (Kalita *et al.*, 2023). Furthermore, *B. coprophila* embryos can be induced into egg-laying such that precise time stages can easily be sampled (Reynolds, 1938) and detailed examination of early embryogenesis can be investigated. In this chapter, I generate and analyse early embryo gene expression data in *B. coprophila* to investigate the molecular basis for sex determination to identify candidate genes controlling primary sex ratios in Sciaridae. I also present the most parsimonious model for control of the sex ratio, building on those developed previously (de Saint Phalle and Sullivan, 1996; Sanchez and Perondini, 1999).

4.3. Materials and methods

4.2.1. Data collection

The *B. coprophila* colonies used in this experiment were from the Holo2 strain obtained from the Sciara stock centre at Brown University (https://sites.brown.edu/sciara/), which are derived from highly inbred lines maintained in laboratory conditions since the 1920s (Metz, 1938). Gynogenic (X’X) and androgenic (XX) females are distinguishable by an X’-linked dominant wing mutant marker, *Wavy* (Metz and Smith, 1931), such that embryos of known sex can be obtained. I collected embryos using a method similar to that described by (Reynolds, 1938), where flies are induced into egg-laying by crushing the head and/or thorax. Since eggs are fertilised just before they are laid (DuBois, 1933), they can be sampled at desired stages of development.

In *Drosophila*, ZGA begins at the 8th nuclear cycle and accelerates around the 14th cycle following cellularisation; soon after this, maternal factors begin to degrade (Lott *et al.*, 2011). In *B. coprophila*, which has a longer generation time than *Drosophila*, cellularisation occurs during interphase of nuclear cycle 11 at 9.3 +/- 1.1 hours after egg deposition (AED) at 22°C. Cell cycle 12 begins at 11.4 +/- 1.4 hours AED (de Saint Phalle and Sullivan, 1996). Assuming ZGA also occurs during, or just before, cellularisation in *B. coprophila*, ZGA may initiate as early as ~8h AED and should have begun by ~12h AED. The X chromosome eliminations occur during the 7-
9th cleavage divisions at approximately 4-6h AED (DuBois, 1933), at which stage the embryos should contain exclusively maternal transcripts. I aimed to collect embryos before (0-4h AED) and after (4-8h AED) X elimination with the aim of sequencing the maternal transcriptome and capturing any maternally deposited transcripts involved in X elimination. To this end, mass matings (typically 10 females and 5 males) were set up between 1–2-day-old adults in 25mm x 95mm glass vials containing 2.2% Bacteriological agar and left overnight (approximately 18h) at 18°C. I then pinned females by the wing in a petri dish containing 2.2% Bacteriological agar and crushed the thorax to trigger egg deposition. To sample the 0-4h time stage I collected embryos after 4h; to sample the 4-8h time stage I removed females after 4h and then collected embryos after 8h. I collected 3 replicates per time stage for each sex for a total of 12 samples, totalling approximately 600-800 eggs per sample (each female lays around 50-100 eggs).

Figure 4.3. *B. coprophila* embryos at 3 and 5 hours after egg deposition. At 3h, the embryo has 32 nuclei and has therefore undergone 5 cleavage divisions. At 5h, the number of nuclei exceeds 128 and the embryo has therefore undergone at least 7 divisions. Image credit: Christina N. Hodson.

I extracted RNA using a modified version of the PureLink RNA Mini Kit (ThermoFisher Scientific) with a non-column-based purification step using TRIzol (Invitrogen). The samples were sequenced for 50M 150bp paired-end Illumina reads on the NovaSeq 6000 platform for 15Gb of data per sample. Prior to analysis, I adapter- and quality-trimmed reads using fastp v0.2.1 (Chen et al., 2018) and assessed their quality using FastQC (Andrews, 2010).
I used two complementary methods to identify candidate genes. First, I conducted a differential gene expression (DGE) analysis between male embryos (from XX mothers) and female embryos (from X’X mothers), utilising both sexes to identify genes upregulated in one versus the other. Second, I analysed allele-specific expression (ASE) of X-linked genes (i.e. expression from the X and X’ alleles) in just female embryos to identify genes with expression from only one allele. I then assessed functional annotations of candidates from the two analyses to generate a shortlist. This workflow is summarised in Figure 4.4.

**Figure 4.4.** Two methods were used to identify candidate sex determination genes. (A) The differential gene expression (DGE) workflow involved comparing expression between male embryos (from XX mothers) and female embryos (from X’X mothers). (B) The ASE workflow involved comparing expression between X and X’ alleles within female embryos (from X’X mothers).
4.2.2. Differential gene expression (DGE) analysis

To examine DGE between male and female embryos, I quantified RNAseq reads using Kallisto (Bray et al., 2016). I generated a Kallisto index from predicted transcript sequences from the autosomes and the X chromosomes (Chapter II, Baird, Urban, et al., 2023). I selected the longest transcript sequence for each gene using a custom shell script, and I intentionally excluded the X'-linked genes so that reads originating from the X' were force-mapped to X-linked genes. Using this method allows for identification of X'-linked alleles that are up- or down-regulated relative to their X-linked counterparts. I conducted a principal component (PCA) analysis on a Log transformation of the raw read counts using the function plotPCA from the DESeq2 package (Love et al., 2014). In addition to this, I assessed Log10 transcript per million (TPM) distributions for each sample.

Prior to downstream analysis, I performed quantile normalisation (i.e. made the distribution of each sample equal) on counts between samples within each time-stage to remove variability due to e.g. the presence of more tissue (i.e. more embryos) in some samples relative to others (Hansen et al., 2012; Zhao, Wong, et al., 2020). I removed genes with normalised counts lower than 4 to exclude those that may appear to be expressed due to stochastic mismapping of reads. One limitation of this approach is that potential candidates with low expression may be filtered out; however, such candidates are likely to still be captured through the allele-specific expression analysis (see below). To visually represent the extent of sex-biased gene expression at each time stage, I calculated the specificity metric (SPM) for male versus female expression for each gene (Kryuchkova-Mostacci and Robinson-Rechavi, 2017) as the mean of normalised counts for females divided by the sum of mean normalised counts for each sex. SPM values thus range from 0 to 1, representing the proportion of expression of any one gene that is unique to females, i.e. 1 = female-limited, 0 = male-limited expression, and 0.5 = equal expression between the sexes. I assigned genes as female-biased if they had SPM > 0.7, and male-biased if they had SPM < 0.3, which corresponds to a 1.5x difference in expression between the sexes and is typical in analysis of sex-biased gene expression (e.g. Pinharanda et al., 2019; Mongue et al., 2022).

To identify genes as significantly differentially expressed with statistical support, I conducted DGE (differential gene expression) analyses using three DGE packages: EdgeR
(Robinson et al., 2010), DESeq2 (Love et al., 2014), and Limma-voom (Law et al., 2014) in RStudio v4.2.2 (R Core Team, 2023). The DGE methods I used assume that (i) most genes are not differentially expressed, (ii) every gene’s read counts follow a similar distribution, and (iii) the number and magnitude of up- and down-regulated genes are comparable (Robinson et al., 2010; Law et al., 2014; Love et al., 2014; Zhao, Ye, et al., 2020). Since I was comparing maternal transcriptomes between early embryos, which are expected to be identical in conventional systems, I presumed that assumptions (i) and (iii) would hold for the majority of genes. I performed the DGE analyses on quantile-normalised read counts, and assigned genes as differentially expressed if they had a Log2 fold change (Log2FC) of greater than or equal to 0.6, corresponding to approximately 1.5x difference in expression between the sexes. I chose only to assign genes differentially expressed between the sexes if they had adjusted P-values of <0.05 and Log2FC values of >0.6 (~1.5x expression) in at least two of the three methods, as in Fruchard et al., (2020). EdgeR and DESeq2 use the same false discovery methods but EdgeR generally produces fewer false positives (Robinson et al., 2010; Love et al., 2014). Limma-voom has higher specificity and produces fewer false positives than either method (Law et al., 2014). Since the three methods have different characteristics, using them in combination should maximise the number of true positives while minimising the number of false positives, thereby reducing the overall number of potential candidate genes while retaining those that have the strongest evidence. I carried out DGE analyses on the two time stages (0-4h and 4-8h) separately, between male and female embryo samples.

4.2.3. X and X’ allele-specific expression (ASE) analysis

As a complementary analysis to the DGE analysis above, I also employed an allele-specific expression (ASE) pipeline similar to the parent-of-origin expression analysis used by Marshall et al., (2020). This method allowed for identification of genes where only either the X or X’ allele is expressed in early female embryos, and which may be missed through analysis of differential expression between male and female embryos. To this end, I mapped Illumina reads generated in Chapter II and Baird, Urban, et al., (2023) from X’X individuals to the male (X0) reference genome (Urban et al., 2021, 2022) using Bowtie2 (Langmead and Salzberg, 2012), processed BAM alignment files with Picardtools (http://broadinstitute.github.io/picard/) and SAMtools (Li
et al., 2009), and identified and filtered variant sites using freebayes (Garrison and Marth, 2012) and VCFtools (Danecek et al., 2011), respectively. I assumed that the alternate alleles of all heterozygous variants called, with at least 5 supporting reads (where median diploid coverage was approximately 30x), were specific to the X'. This is a relatively safe assumption because the line used likely has high homozygosity due to roughly 100 years of inbreeding. I used BEDtools (Quinlan and Hall, 2010) to create an N-masked version of the X0 genome to which I mapped the early embryo RNAseq data with STAR (Dobin et al., 2013). N-masking heterozygous single-nucleotide variants on the X chromosome in the core genome served to ensure that RNAseq reads originating from the X and X' chromosomes mapped equally well to the X-linked genes. I then binned RNAseq reads according to whether they originated from the X or X' chromosomes using SNPsplit (Krueger and Andrews, 2016), based on whether a read contained the reference allele (X) or the alternate allele (X'). As a control, I also binned reads from the male embryo samples, with the expectation that around twice as many reads should be assigned to X alleles in the male samples (XX mothers) compared to the female samples (X'X mothers), while very few reads should be assigned to X' alleles.

I then quantified the binned female embryo X and X' RNAseq reads against a Kallisto index as described above; this time the index was built from just X-linked genes inside the non-recombining region (i.e. between 4.1 and 62.9Mb on the X chromosome), again such that both X' and X transcripts would be force-mapped to the X-linked genes. Since X and X' reads were derived from the same samples, it was not necessary to perform quantile-normalisation on the counts. Instead, I used mean TPM across samples for comparisons between expression of X and X' alleles. I assigned genes as having expression limited to either the X or X' allele if one allele had a TPM of zero or was in the bottom 1% of non-zero TPMs (to account for potential stochastic mismapping of reads), and where the other allele did not. This method of filtering is more conservative than used in the DGE analysis above, but it still carries the limitation that some genes with extremely low expression, which may be indistinguishable from sampling noise, may be filtered out.

4.2.4. Functional annotations
I shortlisted a series of potential candidate genes for the SRM-suppressor and SRM-activator models using expression evidence from the analyses above. For the SRM-suppressor model, I identified genes that had strongly female-biased (>10-fold in females compared to males) expression in the DGE analysis, or expression limited to X' alleles in female embryos, based on the assumption that the candidate should be expressed only from the X' chromosome (i.e. only deposited in female embryos). For the SRM-activator model, I identified genes with male-biased expression, or with expression limited to the X allele in female embryos, assuming that the candidate should be expressed only from the X chromosome (i.e. deposited at double the dose in male embryos).

To narrow-down candidates further, I analysed putative functional annotations of predicted transcripts (generated in Chapter II and Baird, Urban, et al., 2023). I excluded genes that had assigned functions that rendered them unlikely to be involved in chromosome elimination, such as those with physiological or structural functions (e.g. ion transporters, membrane proteins), while including genes with assigned functions in roles such as chromosome behaviour, RNA- and DNA-binding, transcription factors, zinc finger proteins, or ubiquitination proteins. Where genes had no annotated function, I performed discontiguous megablast (Altschul et al., 1990) searches against the NCBI RefSeq database (O’Leary et al., 2015), which it more sensitive than BLASTn and thus may allow for assignment of putative functions based on distant homology to annotated proteins in other organisms (McGinnis and Madden, 2004). For many genes (191, 35% of all candidates) I was unable to assign putative functions, and while I cannot exclude them as potential candidates, I did not consider them further.

4.2.5. Duplicated gene copies

To identify candidates that had multiple paralogs, I ran Orthofinder (Emms and Kelly, 2019) on all predicted transcripts. Since Orthofinder requires two input fasta files (e.g. one per species, if the comparison is between species), I ran the set of predicted transcripts against themselves. I then identified orthogroups and gene counts of shortlisted candidate genes using a custom shell script.
4.2.6. Sex determination cascade and zygotic genome activation (ZGA) genes

I also analysed expression of genes known to be involved in the sex determination cascade of *Drosophila*, which, with the exception of *Sex-lethal* (*Sxl*, Ruiz *et al.*, 2003), are generally conserved across insects: *Transformer* (*Tra*), *Transformer-2* (*Tra-2*), *Doublesex* (*Dsx*) and *Fruitless* (*Fru*, Hopkins and Kopp, 2021). To this end, I identified the *Sxl*, *Dsx*, *Fru* and *Tra-2* transcripts sequences using functional annotations of the *B. coprophila* genome generated in Chapter II and Baird, Urban, *et al.*, (2023). Since *Tra* was not annotated, I performed a BLASTn (Altschul *et al.*, 1990) search against the set of predicted transcripts using the *Tra* sequence of the closest available outgroup, *Contarinia nasturtii* (the swede midge), downloaded from NCBI (Ševčík *et al.*, 2016). In addition, I explored the expression of genes that are known to be expressed at the onset of ZGA: *Zelda*, *Knirps* and *Dichaete* (Lee *et al.*, 2014; Li *et al.*, 2014). I performed BLAST searches against the set of predicted transcripts from *B. coprophila* using the longest transcript variants of these genes, downloaded from NCBI, from the closest available species: *D. melanogaster* *Zelda*, *Sitodiplosis mosellana* (Cecidomyiidae) *Knirps*, and *B. coprophila* *Dichaete*. I then analysed their expression, along with the sex determination cascade genes above, in the 0-4h and 4-8h embryos.

4.4. Results

4.3.1. Differential gene expression in early embryos

I found TPM distributions between male and female embryos at 0-4h to be similar (**Figure 4.5A**). TPM distributions for 4-8h samples revealed one outlying male embryo replicate (**Figure 4.5B**). Variance among all 0-4h samples was also comparable, suggesting that all mRNA at this stage is likely maternal in origin. At 4-8h, the male and female samples clustered independently in their variance, suggesting divergence in gene expression between the sexes at this stage and possible onset of ZGA (**Figure 4.5C**). Binning reads to the X and X’ chromosomes (see below) also revealed one 4-8h female replicate as an outlier. I removed the two outliers (female 4-8h replicate 2, male 4-8h replicate 3) prior to downstream analysis.
Figure 4.5. TPM distributions and count variance in the *B. coprophila* early embryo samples. (A) TPM distributions at 0-4h. (B) TPM distributions at 4-8h. (C) PCA of a Log transformation of the raw read counts for all samples.

Of the genes that were expressed at 0-4h or 4-8h, most (9,939 of 10,123 at 0-4h; 12,569 of 12,790 at 4-8h) were not significantly sex-biased in their expression. The majority (67% at 0-4, 68% at 4-8h) of genes with sex-biased expression were those in the region of recombination suppression on the X chromosome, i.e. genes with non-recombining X and X’ alleles. This is despite these genes constituting only 17% of total predicted genes (Chapter II, Baird, Mongue, et al., 2023). Inside as well as outside the non-recombining region, more genes were female-biased in their expression than male-biased (Table 4.1). The SPM distributions also revealed that there were disproportionately more genes that were sex-biased at 4-8h compared to 0-4h ($\chi^2 = 32.93$, $P < 0.001$), disproportionately more genes with sex-limited expression at 0-4h ($\chi^2 = 465.58$, $P < 0.001$), and that there was a significant excess of female-biased genes (506) compared to male-biased genes (277) inside the non-recombining region at 4-8h AED ($\chi^2 = 89.52$, $P < 0.001$).
\( P < 0.001, \textbf{Figure 4.6} \). Of these, 107 female-biased and 44 male-biased genes had statistical support from DGE packages (Table 4.1).

**Table 4.1.** Number and proportions (in parentheses) of sex-biased genes, inside and outside the non-recombining X region (including all autosomes) of *B. coprophila*, with statistical support from at least two DGE packages.

<table>
<thead>
<tr>
<th>Partition</th>
<th>Stage</th>
<th>Male-biased genes</th>
<th>Female-biased genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inside non-recombining</td>
<td>0-4h AED</td>
<td>53 (2.6%)</td>
<td>71 (3.4%)</td>
</tr>
<tr>
<td>region</td>
<td>4-8h AED</td>
<td>44 (1.9%)</td>
<td>107 (4.5%)</td>
</tr>
<tr>
<td>Outside non-recombining</td>
<td>0-4h AED</td>
<td>28 (0.3%)</td>
<td>32 (0.4%)</td>
</tr>
<tr>
<td>region</td>
<td>4-8h AED</td>
<td>13 (0.1%)</td>
<td>57 (0.5%)</td>
</tr>
</tbody>
</table>

**Figure 4.6.** SPM distributions for all genes in male and female early embryos at 0-4h and 4-8h after egg deposition (AED). Note that “outside non-recombining region” includes all autosomal
genes, as well as those on the X but outside the region of recombination suppression. SPM values range from 0-1, representing the proportion of expression that is unique to females, i.e. 1 = female-limited, 0 = male-limited, 0.5 = equal expression between the sexes. Note that not all genes assigned as female-biased or male-biased above have statistical support from DGE packages (see Table 4.1).

I assigned RNAseq reads from female embryos to X and X' alleles using ASE. The mean proportion of reads assigned to the X and X', respectively, was 6.27% and 5.69% at 0-4h, and 6.03% and 5.64% at 4-8h. This is excluding one female 4-8h replicate, which had around twice as many X reads (7.97%) as X' reads (3.61%) assigned and was excluded from further analysis. The difference in percentage of reads assigned to either allele for this sample was due to a difference in expression for many X-linked genes, rather than for a few, suggesting that this sample was likely an outlier due to developmental timing at the point of collection resulting in more X alleles being expressed (Supplementary Figure 4.1, Supplementary Table 4.1). For other replicates, the differences in reads assigned is likely to be due to degeneration of X' alleles, which is likely to confound analysis of differences in X and X' expression. As a control, I ran the same pipeline on the male early embryo data and found that approximately 2x as many reads – 11.94% at 0-4h and 11.36% at 4-8h – were assigned to the X chromosome, while a negligible amount (0.02%) were assigned to the X'.

There are 3,429 genes within the non-recombining region of the X chromosome, of which 2,344 showed expression from both X and X' alleles during in early embryos. 329 genes showed expression exclusively from the X allele, while 115 were expressed only by the X' allele (Figure 4.7). However, many of the genes that had expression limited to one allele had very low expression levels: of the 329 X-limited genes, 256 had TPM < 1; of the 115 X' limited genes, 104 had TPM < 1. This may be a result of technical stochasticity, or, alternatively, it could be reflective of functionally unimportant genes being those that are first lost from degrading non-recombining sex chromosomes (Furman et al., 2020).
4.3.2 Candidate sex ratio modifiers

Based on the expression evidence from the DGE and ASE analyses above, I shortlisted a series of potential SRM candidates for the two models. The SRM-suppressor model posits that a maternal-effect SRM allele is produced only by gynogenic females and suppresses X elimination in the embryo, while under the SRM-activator model, the SRM should be deposited at double the dose in male embryos. The simplest versions of the models would assume that the SRM is therefore either produced only from the X’ (SRM-suppressor), or only from the X (SRM-activator).

I identified genes with strong female-biased expression (DGE analysis), and/or expression limited to the X’ allele in female embryos (ASE analysis), as SRM-suppressor candidates, and genes with male-biased expression (DGE analysis), and/or expression limited to the X in female embryos (ASE analysis), as SRM-activator candidates. For each model, I initially focused on genes within the non-recombining region. However, since genes outside the non-recombining region may be regulated by non-recombining alleles, I also considered sex-biased genes outside this region.
Through analysis of DGE, I identified 103 genes with female-biased expression. Of these, I selected 19 that had 10-fold higher expression in female embryos relative to male embryos. I also identified 81 genes with male-biased (Log2FC > 0.6) expression as candidates for the SRM-activator model. Through the ASE analysis, I identified 115 genes with expression limited to the X’ allele (SRM-suppressor model) and 329 genes with expression limited to the X allele (SRM-activator model). I assigned putative functions to these candidates. Many (191; 35% of all expression-based candidates) had no annotated function. I shortlisted a series of candidates for the SRM-suppressor (Table 4.2, Table 4.4) and SRM-activator (Table 4.3, Table 4.5) models based on their putative functions. Of the candidates, some encoded zinc-finger proteins (e.g. genes 6, 7, 15, 26), which often have roles in transcriptional regulation (Cassandri et al., 2017). I also found some genes associated with ubiquitin proteins (e.g. genes 16, 22), which can have diverse functions including chromatin modification and regulation (Vaughan et al., 2021). Some candidates I also putatively assigned as SLACS retrotransposons (e.g. genes 11, 24), which are often implicated in genomic rearrangements and evolutionary changes in gene regulation (Patrick et al., 2008).

Table 4.2. Candidate sex ratio modifier genes for the SRM-suppressor model (female-biased/limited expression) identified through differential gene expression between male and female embryos. Putative functions (also for Tables 4.3-4.5) were obtained from Uniprot (https://www.uniprot.org/) or NCBI (https://www.ncbi.nlm.nih.gov/). Counts shown are quantile normalised.

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Linkage</th>
<th>Number of X-linked paralogs</th>
<th>Annotation</th>
<th>Putative function</th>
<th>Count in females</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>III</td>
<td>NA</td>
<td>Adarb1 Double-stranded RNA-specific editase 1</td>
<td>Modification of gene expression and function</td>
<td>927.67</td>
</tr>
<tr>
<td>2</td>
<td>X</td>
<td>3</td>
<td>Igf9 protein turtle homolog A</td>
<td>Possible involvement in embryo development</td>
<td>1791</td>
</tr>
<tr>
<td>3</td>
<td>X</td>
<td>1</td>
<td>TCP11LT T-complex protein 11-like</td>
<td>Fertility</td>
<td>269.75</td>
</tr>
</tbody>
</table>
Table 4.3. Candidate sex ratio modifier genes for the SRM-activator model (male-biased expression) identified through differential gene expression between male and female embryos. Counts shown are quantile normalised.

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Linkage</th>
<th>Number of X-linked paralogs</th>
<th>Annotation</th>
<th>Putative function</th>
<th>Count in males</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>X</td>
<td>1</td>
<td>Mns1 meiosis-specific nuclear structural protein 1</td>
<td>Spermatogenesis, meiotic division, germ cell differentiation</td>
<td>2288.67</td>
</tr>
<tr>
<td>5</td>
<td>X</td>
<td>1</td>
<td>RREB1 Ras-responsive element-binding protein</td>
<td>Transcriptional regulation</td>
<td>7094.33</td>
</tr>
<tr>
<td>6</td>
<td>X</td>
<td>1</td>
<td>Znf12 Zinc finger protein 12</td>
<td>Developmental control of gene expression</td>
<td>197</td>
</tr>
<tr>
<td>7</td>
<td>X</td>
<td>1</td>
<td>Znf182 Zinc finger protein 182</td>
<td>Transcriptional regulation</td>
<td>5128</td>
</tr>
</tbody>
</table>

Table 4.4. Candidate sex ratio modifier genes for the SRM-suppressor model (X'-limited expression) identified through allele-specific expression between X and X’ alleles in female embryos. Note that all genes here are X-linked.

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Number of X-linked paralogs</th>
<th>Annotation</th>
<th>Putative function</th>
<th>X’ allele TPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>1</td>
<td>ASCC2 Activating signal cointegrator</td>
<td>DNA damage repair; binds to polyubiquinated proteins</td>
<td>5.62</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>Jockey/pol RNA-directed DNA polymerase from mobile element jockey</td>
<td>DNA polymerase</td>
<td>0.10</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>Protein with forkhead-associated (FHA) domain</td>
<td>Domain present in diverse proteins including TFs and RNA-binding proteins</td>
<td>0.39</td>
</tr>
<tr>
<td>Gene ID</td>
<td>Number of X-linked paralogs</td>
<td>Annotation</td>
<td>Putative function</td>
<td>X allele TPM</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------------------</td>
<td>------------------------------------------------</td>
<td>------------------------------------------------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>11</td>
<td>16</td>
<td>RFX Regulatory Factor X DAF-19 / SLACs-associated transposon</td>
<td>Diverse functions including transcription</td>
<td>0.13</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>RTase probable RNA-directed DNA polymerase from transposon BS</td>
<td>DNA polymerase</td>
<td>0.13</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>Transposable element Tcb1 transposase</td>
<td>TE transposition</td>
<td>0.37</td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>tRNA-splicing ligase RtcB homolog</td>
<td>Catalyses tRNA splicing; may have functions with other RNAs</td>
<td>0.04</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>ZBED1 Zinc finger BED domain-containing protein 1</td>
<td>Transcriptional regulation</td>
<td>0.26</td>
</tr>
</tbody>
</table>

**Table 4.5.** Candidate sex ratio modifier genes for the SRM-activator model (X-limited expression) identified through allele-specific expression between X and X' alleles in female embryos. Note that all genes here are X-linked.
4.3.3. Sex determination cascade and ZGA genes

The sex determination cascade genes *Sxl, Tra-2, Dsx,* and *Fru* were all present in the *B. coprophila* genome, but I found no evidence of the presence of *Tra*. The four genes that were present were all expressed at 0-4h and at 4-8h, but I found no significant sex-biased expression.

I did not find a *B. coprophila* homolog for the ZGA genes *Zelda* or *Knirps*, although *Dichaete* was present and expressed in both sexes at 4-8h. *Dichaete* did not show significant sex-biased expression.

4.5. Discussion

Maternally supplied transcripts are essential for the early stages of embryonic development for a fertilised egg to transition into a totipotent state (Hamm and Harrison, 2018). Upon fertilisation, zygotes are initially transcriptionally silent, with early development being under exclusive control of these maternal factors. Female gametes are vastly larger than male gametes, so almost all the cytoplasm of a fertilised embryo, which contains the maternal transcriptome, comes from the eggs rather than the sperm. The maternal transcriptome is produced via active transcription during oocyte development, towards the end of which, transcription ceases and it is left to the fertilised zygote to translate the maternally-deposited transcripts (Tora and Vincent, 2021). Maternally deposited factors include many essential epigenetic regulators for the initial developmental steps prior to ZGA (Lott *et al.*, 2011; Ray *et al.*, 2011;...
In *Drosophila*, maternal mRNA deposition should not differ between early male and female embryos (Lott *et al.*, 2014), with sex-specific or sex-biased gene expression beginning following ZGA as the sex determination cascade initiates (Salz and Erickson, 2010) and X hemizygosity in males produces X dosage differences between males and females (Prayitno *et al.*, 2019). Dark-winged fungus gnats (Sciariidae) offer an excellent opportunity to study gene expression in early embryos, because not only does monogamy allow for entire clutches of known sex to be analysed, but the existence of genotypically distinct reproductive female morphs means that females can deposit different maternal transcriptomes into oocytes.

Here, I analysed differential patterns of maternal transcript deposition in early embryos of the monogenic species *B. coprophila*, with the aim of identifying candidate genes involved in the X chromosome elimination process that determine offspring sex. I found that mRNA in early embryos prior to 4h AED is likely exclusively maternal in origin, but that many genes show sex-biased or allele-specific expression, largely as a result of non-recombining X and X’ alleles. Through analysis of expression and functional annotations, I have identified a list of candidate genes as maternal-effect SRMs in this family of flies.

### 4.4.1. Early embryo expression patterns: maternal versus zygotic gene expression

The two time stages (0-4h and 4-8h AED) showed different patterns of sex-biased gene expression. Some differences may be due to degradation of maternally deposited transcripts: although the maternal transcriptome is critical for early embryogenesis, proper zygotic development requires degradation of the maternal transcriptome (Svoboda *et al.*, 2017), and maternal transcripts are actively degraded from meiotic maturation until ZGA (Vastenhouw *et al.*, 2019). I found that there were disproportionately fewer genes with sex-limited expression at 4-8h compared to 0-4h. One explanation for this is that transcripts at 0-4h are exclusively maternal, while some of those at 4-8h are zygotic. Because many of the maternal alleles are non-recombining in this system, and thus diverging in their expression, sex-limited or sex-biased expression of the maternal transcriptome is expected. The lack of sex-limited transcripts at 4-8h suggests that many of those that were supplied maternally may have degraded.

Patterns of sex-biased gene expression at 4-8 in *B. coprophila* are indicative of ZGA having begun by the end of this time stage. Following the onset of ZGA, most zygotically active
genes should not exhibit sex-biased or sex-limited expression: in insects, sex-biased gene expression increases throughout development, accompanying an increase in phenotypic sexual dimorphism (Djordjevic et al., 2022). However, in insects, an excess of female-biased expression is typical between ZGA, when the haploid X is expressed at half the level of the diploid autosomes in males, and the onset of dosage compensation (DC) which restores diploid expression of X-linked genes to balance X-autosome expression (Charlesworth, 1996; Kalita et al., 2023). In Bradysia, DC is achieved by hypertranscription of the male X (da Cunha et al., 1994; Urban et al., 2021), although the precise timing and mechanism are unknown. In Drosophila, ZGA coincides with embryonic cellularisation (Lott et al., 2011). In Bradysia, cellularisation occurs around the 11th nuclear cycle, which may be as early as 8h AED, and the germline nuclei cellularise a little earlier at the 7th cycle, around 4h AED (de Saint Phalle and Sullivan, 1996).

However, in B. coprophila, while male somatic cells are X0, their germline cells are XX. Sex-biased expression in the germline should therefore not result from a difference in X dosage between the sexes. My finding that there is an excess of female-biased X-linked genes at 4-8h indicates that somatic ZGA has begun at this stage, but that DC has not. In further support of this, I also found evidence for expression of the known ZGA gene Dichaete at 4-8h, which has conserved function between Drosophila and Anopheles (Kalita et al., 2023). Further work will be required to more precisely determine the timings of ZGA and DC in B. coprophila.

4.4.2. Early embryo expression patterns: sex determination and sex ratio modifiers

In most insects, the genes Dsx and Fru undergo sex-specific splicing and initiate sexually dimorphic development. Dsx and Fru are regulated by a Tra-Tra-2 complex. In Drosophila, Tra is regulated by Sxl, which in turn undergoes sex-specific splicing regulated by maternally deposited factors (Ray et al., 2023). The function of Sxl as a master sex-determiner is specific to Drosophila, while the other genes in the cascade are relatively conserved across insects (Sanchez, 2008). There is no sex-specific splicing of Sxl in the sciarids B. coprophila, Rhynchosciara americana and Trichosia pubescens, thus Sxl probably plays no discriminatory role in sex determination in this family (Serna et al., 2004). There is not yet an account of Tra in any Sciaridae genomes, and I found no evidence of this gene being present in B. coprophila. Braydsia does possess Tra-2, which appears to show conserved sex determination functions
when transgenically expressed in *Drosophila* (Martín *et al.*, 2011), suggesting that *Tra*-2 may still have sex determining functions in sciarids. *Fru* remains unstudied in Sciaridae, although *Dsx* has been explored and does not exhibit sex-specific splicing but is reported to show sex-specific transcript abundance (Ruiz *et al.*, 2003). I found *Tra*-2, *Dsx*, and *Fru* to be transcriptionally active at 0-4h and 4-8h, suggesting that the sex determination cascade may be initiated prior to 4h AED (or alternatively, some of these factors are maternally deposited). However, I did not find evidence for significant sex-biased expression for any of these canonical sex determination genes, suggesting that they may not play discriminatory roles in sex determination at these time stages.

There was also no significant sex-biased gene expression for the vast majority of genes in *B. coprophila* early embryos. This is consistent with the general finding that in conventional systems, with no maternal effects on sex determination, mothers supply the same set of maternal transcripts to each embryo (Lott *et al.*, 2014). Unsurprisingly, a disproportionate number of genes with sex-biased expression were those with non-recombining X and X’ alleles, where lack of recombination should lead to divergence in gene expression due to degradation, drift, selection, or movement of transposable elements (Bachtrog, 2006; Pucholt *et al.*, 2017; Dechaud *et al.*, 2019). However, a minority of genes with significant sex-biased expression were located elsewhere in the genome. Potential explanations for this finding include mismapping of X or X’ RNAseq reads to autosomal paralogs, or trans-regulation of autosomal genes by elements that lie within the non-recombining region. Determining the causes of differential expression of these genes will require further investigation into their functions and regulation.

Many of the 3,429 genes in the non-recombining X region had expression limited to one allele: 12% to the X allele and 4% to the X’ allele. Since the female-limited X’ is undergoing degradation (Chapter II, Baird, Urban, *et al.*, 2023), expression exclusive to the X allele is expected for genes where the corresponding X’ allele has degenerated. In contrast, the X chromosome still recombines during meiosis in androgenic females and so should be more resilient to degeneration. However, even recombining sex chromosomes can degrade (Nozawa *et al.*, 2016, 2021), as X-linked genes with functional Y gametologs can still accumulate slightly deleterious mutations (Mrnjavac *et al.*, 2023). This evolutionary pattern may explain why a significant minority of genes showed X’-exclusive expression. However, there may be other processes that explain this finding. The non-recombining portion of the X’ chromosome is
exclusively transmitted between females. As such, it may be a hotspot for positive selection acting on female-beneficial alleles. However, positively selected divergence in expression of $X$ and $X'$ alleles is unlikely for dosage-sensitive genes (Xu and Zhou, 2020), and should be further constrained by conflict between androgenic and gynogenic females over gene expression, especially considering that the two female morphs are likely to have similar fitness optima for most genes. Divergence in gene expression between intrasex morphs is a known phenomenon, however (Pointer et al., 2013; Todd et al., 2018; Mank, 2023).

Another possibility is that some loci with expression limited to the $X'$ allele are involved in control of the sex ratio. This would require production of the SRM to be limited to gynogenic females (and thus the $X'$ chromosome), consistent with the SRM-suppressor model. Interestingly, a loss-of-function mutation in the digenic species $B. ocellaris$, which has no $X'$ chromosome, caused an increase in male production at the expense of female production (Mori et al., 1979). This finding is consistent with only gynogenic females producing a SRM which suppresses $X$ elimination, as a loss-of-function mutation targeting the SRM should reduce female production. As discussed above, due to the variability of sex ratios in digenic species, the SRM is also likely to be an allele present at more than one copy. Among the candidates I identified, some had multiple paralogs across the $X$ chromosome. The spread of multiple copies of the SRM may be facilitated by retrotransposons (Patrick et al., 2008), of which I found many that were sex-limited. However, it is worth noting that the SRM may not interact directly with the chromosomes to control $X$ elimination, but rather act via any number of intermediate factors, as suggested by Sanchez and Perondini (1999). Some of the candidates I identified had roles in transcriptional regulation or RNA-binding, suggesting that they may be involved in regulatory cascades. Some regulatory elements, such as small RNAs, may be difficult to detect through mRNAseq, which is biased towards longer and more abundant transcripts (Zhao et al., 2021). Small RNAs such as micro RNAs and PIWI-interacting RNAs have been shown to have diverse functions in chromatin, translational, and transcriptional regulation (Mattick and Makunin, 2005). As such, small RNAseq may be required to fully elucidate the mechanisms of $X$ elimination in Sciaridae. Alternatively, the maternal product may be translated in the ovary and could therefore be a maternally-supplied protein, in which case gene expression during female meiosis may be required to identify candidates.
4.4.3. A model for molecular control of the sex ratio in Sciaridae

A model for control of the sex ratio in Sciaridae via X elimination in early embryos must be consistent with the following observations: (i) sex of the offspring is determined by the maternal genotypes (Metz, 1938); (ii) sex ratios are highly variable in digenic species (Rocha and Perondini, 2000; Nigro et al., 2007); (iii) sex ratios are probably heritable and fast-evolving in digenic species (Davidheiser, 1947); (iv) loss-of-function mutations in digenic species result in more male offspring (Mori et al., 1979); (v) cases of exceptional offspring of the wrong sex in monogenic species (Metz and Schmuck, 1929); (vi) expression of X'-limited alleles (this study).

The most parsimonious model should therefore involve multiple SRM alleles to account for the variable sex ratios. The number of alleles a female has determines the amount of SRM produced, which in turn decides the proportion of embryos into which SRM transcripts are deposited. The model should also involve the SRM acting as a suppressor of X elimination, where if an embryo is supplied with SRM transcripts, the SRM product stops the elimination of one of the paternal X chromosomes, causing the embryo to develop into a female. The SRM may suppress the action of another factor which causes X elimination, or it may simply be required for proper segregation of the paternal homologs.

Such a model would be consistent with the observations listed above. Under the SRM-suppressor model, higher production of SRM results in more female offspring. That loss-of-function mutations in digenic species result in more males, rather than more females, is consistent with this model. Furthermore, in monogenic species, exceptional male offspring in female broods are far more common than exceptional female offspring in male broods (Metz and Schmuck, 1929). Exceptional males may occur through loss-of-function mutations in SRM alleles, while exceptional females may be due to X-X' recombination events or transposition of SRM alleles from the X' to the X, which are likely to be less frequent than loss-of-function mutations. Finally, as argued above, the expression of many X'-limited alleles is challenging to explain through X degradation or positive hemizygous selection acting in gynogenic females and may require an alternative explanation. Female-limited deposition of factors involved in maternally controlled sex determination may explain some of this allele-specific expression.

A significant challenge in identifying candidate SRM loci in a monogenic sciarid species arises from degeneration of X' alleles due to lack of recombining, which confounds analysis of
allele-specific expression. As demonstrated by this study, there are hundreds of genes with sex-limited expression in early embryos, and many more with some degree of sex-biased expression. One advantage of digenic sciarids is that they still determine sex via X elimination (Goday and Esteban, 2001), but they have fully recombining X chromosomes with no sex-limited inversions (Rocha and Perondini, 2000). However, they do often produce strongly sex-biased primary sex ratios, and even sometimes single-sex broods (Davidheiser, 1947; Rocha and Perondini, 2000). As such, digenic sciarids with strongly biased primary sex ratios are well-placed to identify SRM alleles, since there should be very few differences in the set of maternal transcripts provided to male and female embryos aside from those involved in determining the sex ratio.

4.4.4. Towards transgenic validation of candidate genes

Further research will be required to determine whether candidates presented here are involved in molecular control of the sex ratio in sciarids. In particular, cytogenetic and functional evidence is lacking.

Transgenic techniques have previously been applied in sciarids. Rezende-Teixeira et al., (2008) performed microinjection of double-stranded RNAi into Rhynchosciara americana larvae. More recently, site-specific insertions of DNA and piccyBac-mediated transformation techniques have been developed for B. coprophila (Yamamoto et al., 2015; Yamamoto and Gerbi, 2022). Knock-down or knock-up experiments in gynogenic and androgenic females should alter the proportion of offspring that develop as the other sex accordingly. Knockdowns or knock-ups could be achieved via microinjection of RNAi or mRNA, respectively. As X elimination occurs at the 7th-9th cleavage divisions around 4h AED, individual embryos should be injected as early as possible. Alternatively, the germ tissue of mothers during or following oogenesis could be targeted.
4.6. Conclusion

This is the first study aiming to identify candidate maternal-effect sex determination genes in Sciaridae. Previously, models have been proposed for differential elimination of X chromosomes from male and female embryos (de Saint Phalle and Sullivan, 1996; Sanchez and Perondini, 1999), but importantly, these models have not addressed the observation of the highly variable sex ratios in digenic species. Here, I aimed to synthesise past observations with new genomic and transcriptomic data to provide an up-to-date model and a list of potential candidate genes for molecular control of the sex ratio. The new model invokes multiple SRM alleles that stop elimination of one paternal X in female embryos, for which I have provided a set of candidate loci that should be investigated further through cytogenetic or transgenic work.

More generally, these findings provide a foundation for future study of early embryo expression in insects. How the maternal transcriptome regulates early development and zygotic transcript activity and diversity remains poorly understood (Ray et al., 2023). Additionally, monogenic sciards, where females have diverging gene expression profiles, allow for the study of divergence in gene expression between recently non-recombining chromosomes, or conflict over gene expression between intrasexual morphs. Moreover, this system offers excellent opportunities to study the evolution of sex determination, sex ratios, and chromosome behaviour, particularly when it comes to understanding deviations from systems that have been considered the norm.

4.7. References


Xu L, Zhou Q (2020). The female-specific W chromosomes of birds have conserved gene contents but are not feminized. *Genes* **11**: 1126.


Chapter V. Faster adaptation but slower divergence of X chromosomes under paternal genome elimination

5.1. Abstract

Comparisons between sex chromosomes and autosomes can provide fundamental insights into how the genome evolves and the forces that shape it. Specifically, differences in their transmission and ploidy drive their divergent evolutionary trajectories, with sex chromosomes being predicted to adapt and consequently evolve faster. However, in the well-studied model systems these effects are confounded: sex-linked genes are under less efficient selection due to their reduced effective population size, as well as more efficient selection because recessive mutations are exposed to selection in the heterogametic sex. The fly families Sciaridae (fungus gnats) and Cecidomyiidae (gall midges) present a unique opportunity to disentangle these effects. They have XO sex determination, but males transmit only maternally inherited chromosomes. This phenomenon, known as paternal genome elimination (PGE), results in equal transmission of the X and autosomes, thus allowing the effect of hemizygous selection to be studied in isolation. We examine rates of divergence under PGE and find that, surprisingly, X chromosomes diverge more slowly than autosomes. Using population resequencing and gene expression data, we find that stronger purifying selection on the X drives its slower divergence. Our findings demonstrate the potential of systems with unusual inheritance for understanding fundamental evolutionary processes.

5.2. Introduction

In most sexually reproducing organisms, an individual’s sex is determined by sex chromosomes (Bachtrog et al. 2014). Aside from their role in sex determination, sex chromosomes are distinct from autosomes in other ways that have important consequences for their evolutionary fates. Sex chromosomes typically evolve from a pair of homologous autosomes on which a sex-
determining locus emerges. Recombination becomes suppressed around the locus, expands along the chromosome, and degeneration of the non-recombining region, which is found only in the heterogametic sex (the Y and W in XY and ZW systems, respectively) ensues; sometimes the sex-limited chromosome is lost completely (Charlesworth et al. 2005). This process leads to sex-linked alleles being expressed and transmitted differently to those on autosomes (**Figure 5.1A**). The patterns of evolution that result from this are influenced by a complex interaction of evolutionary processes (Meisel and Connallon 2013). Understanding how sex chromosomes and autosomes evolve differently, and the processes that drive these differences, can contribute to a more comprehensive understanding of how genomes evolve.

Genomes are shaped by both selective and neutral processes, as well as the rate, effect, and dominance of new mutations (Ohta 1992). Where recessive alleles are hemizygous (i.e. X- or Z-linked in the heterogametic sex), they are always exposed to selection. Mutations should therefore be swiftly purged if deleterious or invade and become fixed more quickly if beneficial, known as ‘faster-X’ effect (Charlesworth et al. 1987). X- or Z-linked genes are indeed often found to diverge faster between species (i.e. have higher $dN/dS$) than autosomal genes (Kousathanas et al. 2014; Ávila et al. 2014; Jaquiéry et al. 2018; Bechsgaard et al. 2019; Hayes et al. 2020; Mongue et al. 2022; Chase et al. 2023; Mongue and Baird 2023). Conversely, because sex chromosomes are only transmitted to half of offspring by the heterogametic sex and thus have a smaller effective population size ($N_e$) than the autosomes ($N_eX/N_eA = ¾$), selection on new mutations should be less efficient on the sex chromosomes. This may also drive a faster-X effect, if, for example, less efficient selection on the X leads to more rapid accumulation of weakly deleterious non-synonymous polymorphisms.

Faster-X evolution may therefore be driven by a combination of positive selection on recessive beneficial mutations (**adaptive faster-X**) or less efficient purifying selection on weakly deleterious mutations (**nonadaptive faster-X**) (Meisel and Connallon 2013). Although numerous studies across a range of taxa have found evidence for accelerated rates of X/Z chromosome evolution compared to autosomes, support in some lineages is mixed (Thornton et al. 2006; Rousselle et al. 2016; Pinharanda et al. 2019; Hayes et al. 2020) and the precise causes can be difficult to disentangle. Some studies find support for relaxed purifying selection as the predominant driver (Mank et al. 2010a; Hayes et al. 2020; Mongue and Baird 2023), others, adaptive evolution (Kousathanas et al. 2014; Sackton et al. 2014; Charlesworth et al. 2018;
Mongue et al. 2022). A key limiting factor in systems with regular chromosome inheritance is that these two effects are naturally confounded: X-linked genes are simultaneously hemizygous in males and have a smaller $Ne$ than those that are autosomal. As a result, it is often unclear what drives the different patterns of genetic diversity we see on sex chromosomes and autosomes.

Dark-winged fungus gnats (Diptera: Sciaridae) and gall midges (Diptera: Cecidomyiidae) are two distantly related families of flies that have independently evolved an unusual form of haplodiploidy known as germline paternal genome elimination (Herbette and Ross 2023). They have XO sex determination, but all paternally inherited homologs are eliminated during the first meiotic division during spermatogenesis. Males therefore inherit but do not transmit their fathers’ genes, instead passing on only the maternally inherited genome, including an X chromosome to every offspring (Figure 5.1B). The result of this is that both X chromosomes and autosomes are transmitted in a haploid state through males, and thus $Ne_X/Ne_A = 1$. However, elimination of X chromosomes in somatic cells during embryogenesis, which establishes the sex of an individual, produces XO males and XX females. As such, the X is haploid in males but diploid in females, while the autosomes are diploid in both sexes.

**Figure 5.1.** Chromosome dynamics in different reproductive systems. (A) In classic heterogametic systems, the presence of a sex-limited chromosome (e.g. the Y in XX/XY systems)
results in the X being haploid in males and transmitted less frequently than autosomes. (B) In systems with paternal genome elimination (PGE), the males are also haploid for the X but eliminate paternally inherited chromosomes during spermatogenesis, so that every sperm contains the maternal haploid genome and the X and autosomes are transmitted equally. (C) In systems with PGE, sex is determined after fertilisation, establishing female (XX) and male (XO) karyotypes. Note that in fungus gnats, sperm have two duplicate copies of the maternally inherited X, and fertilised embryos are therefore XXX. Male and female embryos therefore lose 2 and 1 Xs, respectively. For a detailed overview of current understanding of the mechanisms underlying PGE, see Gerbi (2022) and Herbette & Ross (2023).

PGE therefore negates the effect of Ne on the efficacy of selection when it comes to comparisons between sex chromosomes and autosomes. Moreover, identical transmission of the X and autosomes makes the system invariant to many processes that affect patterns of evolution, such as variance in reproductive success (Vicoso and Charlesworth 2009a), demography (Charlesworth 2001), total effective population size (Pool and Nielsen 2007), and the effect of drift on the rate of fixation of neutral and nearly neutral mutations (Mank et al. 2010a). Recombination rates and mutation rates, which usually differ between the X and autosomes due to differences in transmission (Vicoso and Charlesworth 2006, 2009b), should also be equal under PGE. Many of these factors can differ between populations of closely related species (Hutter et al. 2007), which may explain why faster-X results from Drosophila are difficult to interpret and often seem contradictory (Thornton et al. 2006; Garrigan et al. 2014; Ávila et al. 2014; Charlesworth et al. 2018). Under PGE, hemizygosity of the X should be the sole driver in differences in autosome and X chromosome evolution (Hitchcock et al. 2024). This may be particularly useful when investigating, for example, patterns of neutral genetic diversity, which are influenced by multiple processes (Charlesworth 2001; Kirkpatrick et al. 2010; Sayres 2018), as well as the interaction between dominance or recessivity and selection, which has been a longstanding mystery in the field (Meisel and Connallon 2013; Lasne et al. 2017).

In the present study, we investigated patterns of evolution in fungus gnats and gall midges and find that, in stark contrast to studies on systems with regular inheritance, X chromosomes diverge more slowly than autosomes under PGE. In two fungus gnat species we use population resequencing and gene expression data to estimate polymorphism, selection,
and examine the contribution of hemizygous selection on male-biased genes. We find that, although X-linked genes likely adapt faster than autosomal genes, purifying selection on weakly deleterious mutations being stronger than positive selection on beneficial mutations drives the slower-X divergence we observe. We find that this purifying selection is likely a combination of hemizygosity and higher expression of X-linked genes compared to autosomal genes. Our results demonstrate the value of using systems with unusual genetics in understanding complex evolutionary processes.

5.3. Results and discussion

5.3.1. X chromosomes diverge more slowly under PGE

To estimate whether X chromosomes evolve faster than autosomes, studies often examine whether X-linked genes accumulate more nonsynonymous substitutions than do autosomal genes by measuring the scaled rate of divergence, i.e. $dN/dS$ (Meisel and Connallon 2013). We calculated $dN/dS$ between 8 species pairs within the Dipteran infraorder Bibionomorpha: 4 fungus gnat pairs (Sciaridae), 3 gall midge pairs (Cecidomyiidae), and one fever fly pair (Bibionidae, Table 5.1). Fever flies are a non-PGE outgroup with male heterogamety and differentiated X and Y chromosomes (White 1973), thus serving as a control (Figure 5.2A). We calculated $dN/dS$ from single copy orthologs between each species pair (Figure 5.2B), as well as $dN/dS$ computed across all sites (Supplementary Table 5.1). Surprisingly, in 2 of the PGE pairs we found no significant difference between the X and autosomes and in 5 pairs we found significantly lower $dN/dS$ for the X, indicating slower divergence of X-linked genes (Figure 5.2B). This result places PGE in stark contrast with the faster-X divergence seen in many classic reproductive systems. In contrast, for the non-PGE outgroup, median $dN/dS$ per gene was around 1.3x greater on the X chromosome compared to autosomes, which is more in-line with comparisons of sex chromosomes and autosomes in other systems (although $dN/dS$ was not higher for the X when computed across all sites; Kousathanas et al. 2014; Ávila et al. 2014; Jaquiéry et al. 2018; Bechsgaard et al. 2019; Hayes et al. 2020; Mongue et al. 2022; Chase et al. 2023; Mongue and Baird 2023).
Table 5.1. Species pairs analysed in this study, genetic distance ($D_{XY}$) between species within each pair, median $dN/dS$ on for autosomal and X-linked genes, and numbers of single-copy orthologs between species within each pair from which $dN/dS$ was calculated.

<table>
<thead>
<tr>
<th>Focal species</th>
<th>Outgroup species</th>
<th>$D_{XY}$</th>
<th>$dN/dS$ (autosomes)</th>
<th>$dN/dS$ (X)</th>
<th>Orthologs (autosomes)</th>
<th>Orthologs (X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phytosciara flavipes</td>
<td>Bradysia pectoralis</td>
<td>0.176</td>
<td>0.095</td>
<td>0.068</td>
<td>968</td>
<td>76</td>
</tr>
<tr>
<td>Bradysia coprophila</td>
<td>Bradysia odoriphaga</td>
<td>0.162</td>
<td>0.121</td>
<td>0.098</td>
<td>8089</td>
<td>2211</td>
</tr>
<tr>
<td>Bradysia confinis</td>
<td>Bradysia desolata</td>
<td>0.157</td>
<td>0.434</td>
<td>0.259</td>
<td>3078</td>
<td>1607</td>
</tr>
<tr>
<td>Lycoriella ingenua</td>
<td>Lycoriella agraria</td>
<td>0.138</td>
<td>0.240</td>
<td>0.243</td>
<td>4685</td>
<td>1672</td>
</tr>
<tr>
<td>Contarinia nasturtii</td>
<td>Contarinia rumicis</td>
<td>0.208</td>
<td>0.075</td>
<td>0.071</td>
<td>2125</td>
<td>1904</td>
</tr>
<tr>
<td>Sitodiplosis mosellana</td>
<td>Aphidoletes aphidimyza</td>
<td>0.189</td>
<td>0.075</td>
<td>0.070</td>
<td>1931</td>
<td>1964</td>
</tr>
<tr>
<td>Mayetiola destructor</td>
<td>Mayetiola hordei</td>
<td>0.179</td>
<td>0.180</td>
<td>0.168</td>
<td>3183</td>
<td>3882</td>
</tr>
<tr>
<td>Dilophus febrilis</td>
<td>Dilophus femoratus</td>
<td>0.041</td>
<td>0.203</td>
<td>0.263</td>
<td>6106</td>
<td>219</td>
</tr>
</tbody>
</table>
Figure 5.2. Rates of evolution in species pairs within the PGE families Sciairidae and Cecidomyiidae and the non-PGE outgroup family Bibionidae. (A) BUSCO-based phylogenetic tree including all species pairs analysed (see Supplementary Figure 5.1 for untransformed tree). Numbers next to nodes represent ultrafast bootstrap support (%). (B) Comparisons of $dN/dS$ for autosomal and X-linked genes for each species pair (F = focal (ingroup) species, O = outgroup). Asterisks represent significance levels for Mann-Whitney U tests between the X and autoosmes (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, NS = not significant). Images were obtained from https://www.inaturalist.org/, credit: Sciaridae (Odontosciara nigra), J Gallagher; Cecidomyiidae (Contarinia pseudotsugae), G San Martin; Bibionidae (Dilophus febrilis), P Le Mao.

5.3.2. Purifying selection drives slower-X divergence

One limitation of $dN/dS$ as a measure of faster-X evolution is that, although it is informative about rates of sequence divergence, it cannot distinguish between adaptive and non-adaptive causes of divergence (Meisel and Connallon 2013). For example, positive, hemizygous selection on the X can increase the rate of non-synonymous substitutions by fixing recessive beneficial mutations (e.g. in the Monarch butterfly, Mongue et al. 2022), but relaxed purifying selection,
due to the smaller effective population size of the X (in XX/XY systems without PGE), can also lead to a faster accumulation of non-synonymous weakly deleterious mutations and thus contribute to sequence divergence (e.g. Parker et al. 2022). As such, faster-X evolution does not always imply faster-X adaptation, as the latter concerns the rate at which recessive beneficial mutations become fixed. Since X chromosomes in PGE systems are diverging more slowly, this implies that purifying selection is stronger than positive selection (i.e. the distribution of fitness effects is biased away from beneficial recessive mutations towards weakly deleterious recessive mutations). However, this does not preclude the X from undergoing faster adaptation relative to the autosomes. To examine the contribution of adaptive evolution, we utilised population resequencing data to calculate alpha (\(\alpha\)), the proportion of substitutions driven by positive selection (Stoletzki and Eyre-Walker 2011). A high proportion of genes (44% for B. coprophila, 42% for L. ingenua) had negative \(\alpha\) values, which is likely due to an excess of weakly deleterious mutations (Charlesworth and Eyre-Walker 2008), and resulted in a lower median \(\alpha\) for X-linked genes (Supplementary Figure 5.2). Comparing only genes with positive \(\alpha\), we found that \(\alpha\) values were significantly higher on the X chromosome compared to the autosomes in B. coprophila (\(P < 10^{-13}\)), indicating that positive selection may indeed by stronger for X-linked genes. There was no significant difference for L. ingenua (Figure 5.3A).

Notably, our estimates of nonsynonymous (\(pN, P < 10^{-15}\)) and scaled rates (\(pN/pS, P < 0.05\)) of polymorphism in B. coprophila were significantly lower for the X chromosome compared to the autosomes, indicating strong purifying selection removing polymorphisms. Differences for L. ingenua were not significant, and \(pN\) was extremely low across the genome, suggesting that the individuals sequenced may be closely related (see methods, Figure 5.3B-C). Synonymous polymorphism (\(pS\)) was also lower on the X (Figure 5.3D) for both B. coprophila (\(P < 10^{-15}\)) and L. ingenua (\(P < 10^{-5}\)), which could be indicative of background selection (Charlesworth 1994). To examine the contribution of purifying selection, we employed the Direction of Selection (DoS) statistic (Stoletzki and Eyre-Walker 2011), calculated using summed variant counts across all autosomal versus all X-linked genes, to estimate the strength of positive versus purifying selection; a positive value indicates the former and a negative value the latter. We found positive DoS for the autosomes but negative for the X chromosome (Figure 3E), suggesting that slower divergence of X chromosomes is due to a stronger prevalence of purifying selection on weakly deleterious mutations.
5.3.3. **Sex-biased gene expression and hemizygous selection**

Since genes are only exposed to selection if they are expressed, and the X is only hemizygous in males, the effect of selection on genes should increase with male-biased expression. Moreover, because sex chromosomes spend disproportionately more time in one sex, they should accumulate more genes with sex-biased expression than autosomes (Rice 1984; Klein et al. 2021). Accordingly, X and Z chromosomes are often found to have an excess of female-biased (feminisation) or male-biased genes (masculinisation), respectively (Wright et al. 2012; Allen et al. 2013; Albritton et al. 2014; Mongue and Walters 2018), a pattern which tends to be driven primarily by sexually dimorphic expression in reproductive tissues (Grath and Parsch 2016).

We analysed differential gene expression in and bodies and reproductive tissues of *B. coprophila* and *L. ingenua* and found extremely sexually dimorphic gene expression that was not
limited to the gonads, with 73%-78% of genes showing sex-biased expression in the body and 77-80% in the reproductive tissues (Figure 5.4A). Such extreme patterns of sex-biased gene expression are often associated with sexual dimorphism (Harrison et al. 2015) or resolution of intragenomic conflicts (Gallach et al. 2011; Wright et al. 2018), and also appears to be associated with PGE in scale insects (Bain et al. 2021) and lice (Marshall et al. 2020), although whether there is a causal link in the case of PGE is unclear.

For genes expressed in the body, the X appeared to have an excess of female-biased genes and a dearth of male-biased genes relative to the autosomes in both B. coprophila ($\chi^2 = 19.3$, $P < 0.001$) and L. ingenua ($\chi^2 = 28.8$, $P < 10^{-5}$). Interestingly, we found the opposite pattern for reproductive tissues in B. coprophila ($\chi^2 = 25.8$, $P < 10^{-4}$).

We found that genes with sex-biased expression generally evolve more quickly than those with unbiased expression, in line with findings from previous studies (Yang et al. 2016; Jaquíery et al. 2018; Mongue et al. 2022); in the case of B. coprophila and L. ingenua, strongly female-biased genes had the highest $dN/dS$ (Figure 5.4B). Consistent with hemizygous selection in males, in B. coprophila we found that unbiased (body: $P < 0.01$, reproductive tissues: $P < 0.01$), male-biased (body: $P < 0.001$, reproductive tissues: $P < 0.05$) and strongly male-biased (body: $P < 0.05$, reproductive tissues: $P < 0.05$) genes had higher $dN/dS$ if they were X-linked compared to if they were autosomal. We found a similar pattern in L. ingenua, although differences were not significant (Figure 5.4B). That we found patterns suggestive of hemizygous selection in genes expressed predominantly in the body as well as in the reproductive tissues indicates that differences in $dN/dS$ are not driven primarily by genes expressed in sexually dimorphic tissues, which was surprising given that secondary sexually characteristics are often the most rapidly evolving traits (Kopp et al. 2000). Similarly, Pinharanda et al. (2019) found no particular enrichment for genes expressed in ovaries when it came to faster-evolving Z-linked genes in Heliconius.

We did not find that significant differences in $pN/pS$ between the X and autosomes for unbiased or male-biased genes. Interestingly, female-biased X-linked genes expressed in reproductive tissues in B. coprophila had significantly higher $pN/pS$ than autosomal genes ($P < 0.05$, Figure 5.4C). We also unexpectedly found that X-linked strongly female-biased genes had lower $dN/dS$ than autosomal strongly female-biased genes. This effect was significant for body-expressed genes in both B. coprophila ($P < 0.01$) and L. ingenua ($P < 0.05$, Figure 5.4B). This result
suggests that hemizygous selection may not be the only factor driving differences in \(\frac{dN}{dS}\) between the X and autosomes.

**Figure 5.4.** Evolution of differentially expressed genes in *Bradyia coprophila* and *Lycoriella ingenua*. (A) Proportions of genes with strongly sex-biased (>90% in one sex), sex-biased (>70%
in one sex), and unbiased (30-70% in one sex) expression on the autosomes (A) and X chromosome in the body and reproductive tissues. (B) Scaled divergence \( (dN/dS) \) in differentially expressed gene classes. Expressed in the body and reproductive tissues (designated as such if >50% of a gene’s expression was arising from either tissue type). Note that we excluded testes and ovaries from reproductive tissues as germline karyotypes differ (see methods). (C) Scaled polymorphism \( (pN/pS) \) in differentially expressed gene classes. SMB = strongly male-biased, MB = male-biased, UB = unbiased, FB = female-biased, SFB = strongly female-biased. Asterisks represent significance levels for Mann-Whitney U tests between the X and autosomes (*P < 0.05, **P < 0.01, ***P < 0.001, NS = not significant) See Supplementary Table 5.2 for all statistical test results.

5.3.4. Differences in X and autosome expression may contribute to slower X divergence

We investigated differences in expression levels between autosomal and X-linked genes, which could influence the exposure of alleles to selection: genes with higher expression are likely to have a larger effect on and should therefore be under stronger purifying selection. For example, (Jaquiéry et al. 2018) found in the pea aphid system, in which parthenogenetic sex determination also predicts that \( \text{Ne}_X/\text{Ne}_A \) should equal 1 (Jaquiéry et al. 2012), that X-linked genes have much lower expression than autosomal genes and show correspondingly higher \( dN/dS \). In B. coprophila and L. ingenua, for genes that we found to be expressed, we found that X-linked genes had higher expression than autosomal genes (Figure 5.5). Despite this, L. ingenua had disproportionately fewer autosomal genes with zero expression compared to X-linked genes, which may be an artefact of a poorer quality assembly and gene annotation for this species (Supplementary Table 5.3).
Figure 5.5. Expression (mean counts across 3 replicates within each class) of autosomal and X-linked genes in the body, somatic reproductive tissues, and germline reproductive tissues (testes/ovaries) in males and females of *Bradyisia coprophila* and *Lycoriella ingenua*. Asterisks represent significance levels for Mann-Whitney U tests between the X and autosomes (*P* < 0.05, **P** < 0.01, ***P*** < 0.001, NS = not significant).

5.3.5. *Unusual reproductive strategies shape patterns of neutral diversity*

Under PGE, we predict equal effective population sizes for autosomes and X chromosomes (*Ne_X/Ne_A* = 1, Hitchcock et al. 2024). One method of testing this prediction is through comparisons of neutral diversity, i.e. diversity at sites that should be evolving neutrally, such as fourfold degenerate sites (Mank et al. 2010b). Many factors contribute to driving different patterns of neutral diversity on autosomes and X chromosomes, such as variance in male reproductive success and X/A differences in recombination and mutation rates (Vicoso and Charlesworth 2006, 2009b,a). PGE should be invariant such processes because they rely on
differences in autosome and X transmission. However, selective processes such as background selection or genetic hitchhiking, which also affect patterns of neutral diversity, are affected by hemizygosity (Gordo and Charlesworth 2001). PGE systems may therefore be particularly useful for disentangling such effects.

We used population resequencing data to estimate neutral diversity in *B. coprophila* and *L. ingenua*. As a measure of neutral diversity, we calculated heterozygosity ($\theta$) at fourfold degenerate sites. For *L. ingenua*, $\theta_X/\theta_A$ was 0.97, closely aligning with predictions. However, for *B. coprophila*, $\theta_X/\theta_A$ was surprisingly low at 0.28. The species *B. coprophila* is unusual in that it has a reproductive strategy called monogenic reproduction (Baird et al. 2023a). Females produce single-sex broods, and this is associated with a large inversion-based X-linked supergene (the $X'$ chromosome) which is always present heterozygously in female-producing (X'X) females but lacking in male-producing (XX) females (Figure 5.6A, Baird et al. 2023b). In contrast, *L. ingenua* is not monogenic (females produce mixed sex broods), and therefore presumably does not have an associated supergene and suppressed recombination on the X (Shlyakonova et al. 2024).

We estimated $dN/dS$ across genes within the supergene region and and found that it is diverging significantly faster compared to its homologous, standard-X haplotype ($P < 10^{-15}$, Figure 5.6B). This is likely a result of lack of purging of weakly deleterious alleles due to recombination suppression (Charlesworth et al. 2005) but may also result from relaxed purifying selection on the supergene due to its drastically reduced effective population size, which should be 1/6 that of autosomes (Hitchcock et al. 2024). One consequence of the presence of the X-linked supergene is an overall reduction in the rate of recombination on the X chromosome: male meiosis is achiasmatic (Amabis et al. 1979), and most of the X is under suppressed recombination in female-producers due to the inversion-based supergene. Lack of recombination can increase the effect of background selection as linkage groups are broken up less frequently (Nordborg et al. 1996). Thus, lower $\theta_X/\theta_A$ for *B. coprophila* could be a consequence of reduced recombination rate on the X chromosome. To test this, we conducted a sliding window analysis of heterozygosity along the X in *B. coprophila*. Consistent with increased background selection on the X, we found severely reduced heterozygosity on the X regions associated with the supergene in individuals homozygous for the standard (XX) haplotype (Figure 5.6C). However, assessing heterozygosity in species with and without
Supergenes would be required to verify this effect of recombination suppression on a reduction in neutral diversity.

Figure 5.6. Evolutionary dynamics of an X-linked supergene in *B. coprophila*. (A) Schematic of the X chromosomes. Female-producing females are heterozygous for a large X-linked supergene composed of paracentric inversions on the long arm of the chromosome. (B) The supergene is diverging faster (has higher dN/dS) than the autosomes due to accumulation of nonsynonymous substitutions (dN). (C) As a result of the presence of the supergene (visible as raised heterozygosity in X’X female individuals that are heterozygous for the supergene), the X suffers reduced recombination and associated reduction in heterozygosity (visible in XX individuals that are homozygous for the standard haplotype) relative to the three autosomes (AII, AIII, AIV).

A reduced recombination rate may contribute to the slower-X divergence we found in most of the fungus gnat species pairs we analysed, as background selection should slow the accumulation of nonsynonymous substitutions. However, we also found slower-X divergence in all three gall midge species pairs. Gall midges are also known to reproduce via monogenic reproduction (Chapter III; Baird, Mongue et al. 2023), but the associated inversions are notably smaller and autosomal rather than X-linked (Benatti et al. 2010) and should therefore only slow the rate of autosomal divergence. The slower-X effect in fungus gnats is therefore unlikely to be explained by differences in recombination rate alone and is instead probably a result of stronger purifying selection discussed previously.
5.4. Conclusion

Most species in which evolutionary rates of autosomes and sex chromosomes have been compared show faster rates of divergence for X- or Z-linked genes, and this is often attributed to faster adaptation of sex-linked genes. However, results are often mixed, and the causes of faster divergence are a challenge to disentangle. Here, we found that under PGE in fungus gnats and gall midges, where \( \frac{N_{eX}}{N_{eA}} = 1 \), X chromosomes diverge more slowly than do autosomes. The slower-X divergence appears to be driven by stronger purifying selection of X-linked genes, resulting from a combination of hemizygous selection in males and purifying selection housekeeping genes on the X. Interestingly, inversions that reduce recombination rates on the X in fungus gnats also reduce genetic variation on the X chromosome. Our findings demonstrate how species with unusual inheritance are uniquely placed to provide insights into complex questions surrounding sex chromosome evolution.

5.5. Materials and methods

5.5.1. Genome assembly, annotation, ortholog identification and sex-linkage assignment for species pairs

We used publicly-available genomes for *A.aphidimyza, B. coprophila, B. odoriphaga, C. nasturtii, D. febrilis, M. destructor, P. flavipes, and S mosellana*. To compare divergence between the *B. coprophila* X supergene and homologous X haplotype, we used the publicly available scaffold corresponding to the supergene (see Supplementary Table 5.4 for all accessions). For all remaining species, we generated *de novo* genome assemblies using WGS Illumina data. For *D. femoratus, C. rumicis*, and *M. hordei* we used publicly available data. We collected and generated WGS data for *L. ingenua, L. agraria, B. pectoralis, B. confinis, and B. desolata*. *L. ingenua* was collected from Mycobee Mushroom Farm in North Berwick, UK. *B. confinis, B. desolata* and *B. pectoralis* were collected from Cerová Vrchovina forest, Cezchia. *L. agraria* was collected from...
Whytham woods, Oxford, UK. All collected specimens were identified from genital clasper morphology and amplification and Sanger-sequencing of the COI barcode sequence using the primer pair LCO1490/HCO2198 (Folmer et al. 1994).

Genomic DNA was extracted from collected flies using a protocol modified to obtain high-molecular weight DNA with low-input material, developed by C Laumer at the Wellcome Sanger Institute (dx.doi.org/10.17504/protocols.io.bypypvpw). Quantification and quality control was performed with Qubit and Nanodrop (ThemoFisher). The samples were sequenced for 150bp paired-end Illumina reads on the Novaseq 6000 platform (Supplementary Table 5.5 for per-sample coverage). Raw reads generated and those downloaded for this study were trimmed with fastp v0.12.4 (Chen et al. 2018). We generated de novo short read assemblies for L. ingenua, B confinis, B. desolata and B. pectoralis using SPAdes v3.14.1 (Bankevich et al. 2012) and used Blobtools v1 (Laetsch and Blaxter 2017) to identify and remove contaminant (non-metazoan), low-coverage (<5x) and short (<500bp) contigs. Assembly quality was assessed with BUSCO v5 (Seppey et al. 2019) (Supplementary Figure 5.3). We constructed a phylogenetic tree to determine species pairs using BUSCO Phylogenomics (github.com/jamiemcg/BUSCO_phylogenomics) and IQTree v2.2.5 (Minh et al. 2020). FigTree v1.4.4 (Rambaut 2010) was used to plot the tree (Supplementary Figure 5.1).

For B. coprophila, B. odoriphaga and C. nasturtii, genome annotations were publicly available (Supplementary Table 5.4). For all other species we generated de novo annotations. The genomes were soft-masked with Red v2 (Girgis 2015) and annotated with BRAKER2 (Stanke et al. 2008; Lomsadze et al. 2014; Hoff et al. 2016, 2019; Brůna et al. 2021) using the OrthoDB v10 Diptera protein sequences (Kriventseva et al. 2018). Since the L. ingenua genome was to be used for further analyses, we sought to improve the assembly by including RNAseq libraries generated in this study (see below) mapped to the genome with STAR v2.7.8a (Dobin et al. 2013) in the annotation. Prior to this, contiguity of coding areas was improved by scaffolding with RNAseq libraries, which were aligned using TopHat v2.1.1 (Kim et al. 2013) before scaffolding with Rascaf (Song et al. 2016). The assembly was then polished once using Minimap v2.17-r941 (Li 2018) and Racon v1.4.2 (Vaser et al. 2017) to reduce the number of gaps. We used Orthofinder (Emms and Kelly 2019) on the predicted sets of genes to identify singlecopy orthologs for which we calculated dN/dS (see below).
We identified autosomal and X-linked genes based on read depth from the homogametic sex (males), as autosomal sequences should have diploid (2n) coverage while X-linked sequences should have haploid (1n) coverage. Where male reads were available for both members of a species pair, we aligned them with BWA-MEM (Li 2013) and used per-scaffold read depth, computed with SAMtools v1.14 (Li et al. 2009), to identify autosomal and X-linked sequences. For 3 species pairs (B. coprophila/B. odoriphaga, L. ingenua/L. agararia, M. destructor/M. hordei), we inferred X-linkage from only one species since male reads were not available for both. For one species pair (C. nasturtii/C. rumicis), only female reads were available, so we inferred X-linkage based on alignments, using Minimap v2.17-r941 (Li 2018), to the S. mosellana, which was the closest chromosome-level outgroup for which sex linkage information was available (for more information on sex-linkage assignments see Supplementary Text 5.1).

5.5.2. Population genomic analyses

For each pair, the species with the better quality (more contiguous) genome was chosen as the ingroup to ensure better mapping rates and reduce mapping times (see Supplementary Table 5.6 for assembly statistics). Outgroups were aligned to their respective ingroup species with BWA-MEM (Li 2013) using default parameters. PicardTools (http://broadinstitute.github.io/picard/) was used to sort alignments, mark and remove duplicates, and add read groups. Variants were called using the GATK-4 best practises pipeline (McKenna et al. 2010; DePristo et al. 2011; Van der Auwera and O’Connor 2020), and were filtered for Quality by Depth > 2, Fisher Strand-bias < 60 and Mapping Quality > 40. Variant annotations were generated using SnpEff (Cingolani et al. 2012b) and SnpSift (Cingolani et al. 2012a). The script partitioncds.py (Mackintosh et al. 2022) was used to annotate degeneracy for all genic sites. Finally, a custom R script (R Core Team 2023) was used to calculate scaled divergence ($dN/dS$), i.e. the ratio between non-synonymous variants per non-synonymous site ($dN$) and synonymous variants per synonymous site ($dS$), for single copy orthologs within each species pair.
To estimate scaled rates of polymorphism (i.e. $p_N/p_S$), we utilised population sequencing data for *B. coprophila* and *L. ingenua*. We collected 11 *B. coprophila* females and 5 *L. ingenua* XX females for resequencing. *B. coprophila* were collected from houseplants in different apartments in Edinburgh, as well as from the Royal Botanic Gardens, Edinburgh UK. Because *B. coprophila* females can either be X’X (heterozygous for an X-linked supergene) or XX, and we identified XX and X’X females based on distributions of heterozygosity using a sliding window analysis (Supplementary Text 5.2). We identified 3 individuals as having the XX genotype and the remaining 8 as having the X’X genotype. Since we were interested in polymorphisms on the standard X chromosome, we analysed only the 3 XX individuals. We collected *L. ingenua* from a single mushroom straw log obtained from Mycobee Mushroom Farm, North Berwick, from which thousands of individuals were emerging, and may therefore have been descended from relatively fewer individuals, which may explain why we observed lower polymorphism in this species (see main text). DNA extractions and sequencing were performed as described above, for ~20x coverage per individual. We calculated $p_N/p_S$ using the same method described above for $dN/dS$ but aligning resequencing libraries to their respective genomes and combining VCFs with GATK-4 combineGVCFs prior to genotyping.

To assess the extent of positive (adaptive) selection on the X versus the autosomes, we calculated alpha ($\alpha$) for each gene using a calculation of the Neutrality Index ($1 - NI$, Stoletzki and Eyre-Walker 2011). A more positive value of $\alpha$ indicates a higher proportion of fixed non-synonymous differences that are driven by positive selection. To examine the contribution of relaxed purifying selection, we used the Direction of Selection (DoS) statistic, which is defined as the difference between the proportion of non-synonymous fixed differences and the proportion of non-synonymous polymorphisms. A DoS value of zero indicates neutral evolution, while a positive value indicates positive selection and a negative value purifying selection and segregating slightly deleterious mutations (Stoletzki and Eyre-Walker 2011). Rather than calculating DoS for each gene, we summed $dN$, $dS$, $pN$, and $pS$ across each class of genes (i.e. the X and autosomes) prior to calculations to gain more power in estimating DoS.

We calculated neutral diversity (heterozygosity, $\theta$), at fourfold degenerate (i.e. putatively neutral) sites to test our prediction that under PGE, $Ne_X/Ne_A$ should equal 1. To this end, we used the alignments of resequencing datasets for *B. coprophila* and *L. ingenua* described above, and the population genomic tool ANGSD v0.940 (Korneliussen et al. 2014) to calculate $\theta$ for each
scaffold. To analyse divergence along the X supergene, which we suspected was causing a reduction in $\theta$ on the $B.\ coprophila$ X, we aligned reads from the outgroup $B.\ odoriphaga$ to the $X'$-supergene scaffold, calculated $dN/dS$ as described above, and compared this to the regular X haplotype. We analysed $\theta$ across sliding windows of the X chromosome in $B.\ coprophila$ individuals using resequencing datasets. To this end, we called genotypes at all sites with GATK-4 using the output mode EMIT_ALL_CONFIDENT_SITES with the option -include-non-variant-sites when genotyping VCFs, in order to call all (variant and invariant) genotypes at every site. The scripts parseVCF.py and popgenWindows.py (https://github.com/simonhmartin/genomics_general) were used to calculate $\theta$ in 100kb sliding windows across the X in XX and X'X individuals.

5.5.3. Differential gene expression

We mass-reared lab cultures of $B.\ coprophila$ (Metz 1926) and $L.\ ingenua$ (Shlyakonova et al. 2024) and performed dissections to separate gonadal tissue from the remainder of the body (Supplementary Figure 5.4). Germline (i.e. testes or ovaries) and somatic reproductive tissues were separated and sequenced separately, since male germ cells are XX in Sciaridae (Gerbi 2022). Adult flies were anaesthetised on ice, dissected in 1X phosphate buffered saline on a glass slide under a Leica EZ4 microscope, and flash-frozen at -70°C. 20-60 individuals were pooled per replicate and 3 replicates were collected for each sex; for $B.\ coprophila$ we used male-producing (XX) females which lack the female-limited X’ supergene. RNA was extracted using a modified version of the PureLink RNA purification kit (ThermoFisher) with a TRizol solubilisation step. All samples were quantified and quality-checked using the Qubit and Nandrop (ThermoFisher) and were sequenced for 30M 150bp paired-end reads on the NovaSeq 6000 platform, for 9G data per sample.

To analyse sex-biased gene expression and classify genes as differentially expressed, we used Kallisto (Bray et al. 2016). Indices were generated from the $B.\ coprophila$ (Baird et al. 2023b) and $L.\ ingenua$ (this study) gene predictions, against which RNAseq libraries were quantified. Data quality control (PCA and comparing TPM distributions, Supplementary Figure 5.5) and quantile-normalisation of transcript counts were performed in R. Sex-biased expression was
assessed using the specificity metric (SPM) for male versus female expression by dividing the squared mean counts for each gene in females by the sum of squared mean male and female counts (Kryuchkova-Mostacci and Robinson-Rechavi 2017), such that male-limited genes have an SPM of 0 and female-limited genes an SPM of 1. Genes with SPM < 0.3 were assigned as male-biased and SPM > 0.7 as female-biased, corresponding to a difference of 1.5x in expression between the sexes. Genes with SPM < 0.1 and > 0.9 were assigned as strongly male-biased and strongly female-biased, respectively. The aim of assigning strongly female-biased genes was to examine the effect of hemizygous selection, to which such genes should be least exposed. Prior to calculating SPM, we excluded genes with normalised counts < 4 to ensure that our results were not driven by potentially unreliable assignment of bias in genes with very low expression. Because sexually dimorphic reproductive tissues generally contribute disproportionately to sex-biased gene expression, we categorised genes into those that were predominantly (>50%) expressed in the body (i.e. non-sexually dimorphic tissues) and those predominantly expressed in the somatic reproductive tissues (i.e. sexually dimorphic tissues). For analyses of $dN/dS$ and $pN/pS$ for differentially expressed genes, we only included somatic gonadal tissue (i.e. excluding testes and germline) since germ cells are XX in both sexes and should thus not be under hemizygous selection.

We also performed a complementary differential gene expression analysis on the counts with DESeq2 (Love et al. 2014), which is more conservative in defining sex-biased genes compared to using SPM. We used an adjusted $P$-value cut-off of < 0.05 to define expression as significant and a Log2 fold-change of > 0.6 to categorise genes as sex-biased, again corresponding to a difference in expression between the sexes of approximately 1.5x. DESeq2 assigned fewer genes as sex-biased compared to SPM, but there were no major inconsistencies between the two methods (i.e. no genes were categorised as male-biased by one method but female-biased by the other, **Supplementary Table 5.7**).

### 5.5.4. Statistical analyses

Significant differences in $dN/dS$, $pN/pS$, and $\alpha$ were tested with Mann-Whitney $U$ tests for single comparisons (e.g. autosome versus $X$) and Kruskal-Wallis tests for multiple comparisons (e.g.
male-biased versus unbiased versus female-biased). Sex-biased composition of chromosomes was assessed using Chi-squared ($\chi^2$) tests. To test for significance differences in DoS between the X and autosomes, we used a permutation test framework as in (Mongue et al. 2019, 2022; Mongue and Baird 2023). We calculated point estimates of DoS for the X and autosomes and used the absolute value of the difference between the two estimates as the permutation test statistic. From the two gene sets being tested, we randomly sampled individual genes to create two permuted sets of equal size to the true sets, without replacement. We then calculated absolute differences in DoS between the two sampled gene sets for 10,000 permutations and constructed a distribution of differences in point estimates for DoS. These distributions were then compared to the true value, and $P$-values were calculated as the proportion of times the observed value was smaller than the values in the permuted distribution.

5.6. References


158


Rambaut A (2010). FigTree v1.3.1.


Chapter IV. General discussion

Although sexual reproduction is an ancient and ubiquitous feature of eukaryotic life, mechanisms of sex determination are diverse and the various systems and their sex chromosomes undergo striking rates of diversification and turnover (Werren and Beukeboom, 1998; Bachtrog et al., 2014). Moreover, beyond their role in sex determination, sex chromosome evolution has important consequences for adaptation (Zhou and Bachtrog, 2012; Connallon et al., 2018), genomic conflicts (Meiklejohn and Tao, 2010; Mank et al., 2014), and divergence and speciation (Yoshida et al., 2014; Bracewell et al., 2017). As such, the evolution of sex determination mechanisms and sex chromosomes clearly constitutes an important part of biological diversity. In this thesis, I have focused on the particularly unusual sex determination system of dark-winged fungus gnats (Diptera: Sciaridae). I have investigated how this system has evolved and the evolutionary dynamics of its sex chromosomes. Below, I will briefly discuss the main results of each chapter, and in each case, suggest directions for future work.

6.1. Resolving the nature of the sex-determining supergenes

In Chapter II, I assembled and characterised the $X'$-linked supergene of Bradysia coprophila. I determined its size, age, breakpoints, and extent of degradation. My findings showed that rather than one single, long inversion, it appears that the inverted region of the $X'$ is composed of multiple, young inversions that accumulated in a stepwise process. Furthermore, the inverted genes are mostly functional, although there are some signs of degeneration. These findings led me to suspect that perhaps monogenic reproduction has evolved repeatedly in the family through the emergence of these supergenes, and that other closely related species that are known to harbour them, such as $B$. impatiens, may have evolved them independently.

There were some limitations to this study. For example, the assembly and characterisation of the $X'$ was performed using short-read data, which left the structure of the supergene somewhat ambiguous. To fully resolve the $X'$, highly accurate HiFi reads will be required so that inverted alleles can be accurately phased and repetitive sequences assembled. Identifying $X'$ supergenes in other members of the family, such as $B$. impatiens, will also help to
understand whether the *B. coprophila* supergene predates the evolution of this species. Population resequencing data will also be useful in determining the extent of ongoing recombination between X and X' arrangements: it may be that the supergene is older than it appears but is ‘kept young’ by occasional recombination events. Data from more individuals, and from more populations, will be required to resolve the supergene’s evolutionary history.

6.2. The evolution of monogenic reproduction

My finding that the X' supergene of *B. coprophila* may have undergone rapid stepwise expansion suggests that there may be an adaptive component driving this expansion. In Chapter III, I addressed how monogenic reproduction may have evolved in the dark-winged fungus gnat family via the emergence and expansion of inversions that link sex-determining alleles with one another. I also reviewed monogenic reproduction in the other two clades in which it occurs – gall midges and blowflies – and speculated on the reasons why this unusual form of reproduction may have evolved in the first place.

To take these ideas further, one could formalise the verbal arguments presented in this chapter into mathematical models that predict the scenarios in which monogenic reproduction could evolve. It would also be necessary to generate empirical results to assess potential contributing factors, such as the extent of inbreeding, the presence of drive systems, the function of the supernumerary germline-restricted chromosomes, and the number of alleles involved in the control of sex ratios. In blowflies, virtually nothing is known about their sex determination. High-quality genomic data will be essential to determine if the monogenic species have sex chromosomes and if they have any function, and what differentiates female- and male-producing females. Moreover, due to lack of genomic data across diverse species, we still cannot be sure whether turnover between monogenic and digenic reproduction occurs in dark-winged fungus gnats and gall midges, much less the extent of turnover.

More empirical data on reproductive strategies in the gall midges and dark-winged fungus gnats (i.e. whether a species is monogenic or digenic), and genomic data to construct a phylogeny of species with varying reproductive strategies, will be necessary to determine the extent of turnover. For example, the closely related species *B. coprophila* and *B. impatiens* are
known to be monogenic, while *B. ocellaris* and *B. odoriphaga* are digenic (Metz, 1938; Carson, 1946; Cheng et al., 2017). Preliminary work from another study suggests that *B. coprophila* and *B. odoriphaga* are more closely related to each other than the other two species, and that *B. impatiens* and *B. ocellaris* also form a sister-clade (Figure 6.1). Identification of the X’ supergene in *B. impatiens*, and comparative analysis with the X’ in *B. coprophila*, will be required to fully resolve the relationship between these species.

**Figure 6.1.** BUSCO-based phylogeny of four closely related *Bradysia* species (Grewoldt et al., in prep).

### 6.3. How is sex determined in these systems?

Understanding why monogeny evolves, how turnover between digeny and monogeny occurs, and the why it is associated with X’ supergenes, will be aided by insights into the molecular basis of sex determination in these systems. For example, if many loci are involved, this may help to explain rapid stepwise expansion of the supergene, as well as how transitions from digeny to monogeny occur. In Chapter III, I generated gene expression data from early embryos and analysed this to identify potential candidate genes that determine maternal control of offspring sex. Through synthesis with previous findings, I suggested that the most parsimonious model is one where mothers possess multiple copies of a sex-ratio modifier, or multiple sex ratio modifier loci, which are maternally deposited in the early embryo and act to suppress X elimination during the early cleavages. The number of copies of this modifier which a female has determines the proportion of her eggs that receive it.
To validate these candidate sex ratio modifiers, it will be necessary to perform downstream cytogenetic and transgenic experiments. For example, the activity of candidates could be tracked with RNA fluorescent *in situ* hybridisation, and candidates could be knocked-up or knocked-down using mRNA or RNAi microinjections into early embryos. Recent developments in *B. coprophila* transgenesis have made this possible (Yamamoto *et al.*, 2015; Yamamoto and Gerbi, 2022).

However, the task of identifying candidate genes was made difficult by the fact that there are two large and diverging haplotypes in the female-producing females. Many genes are therefore differentially expressed between the inverted and non-inverted haplotypes, and the list of potential candidate genes that I produced is, as a result, rather long. The task would be made easier in a digenic species, where females can still produce extremely sex-biased broods without the presence of a large non-recombining supergene (Rocha and Perondini, 2000). In this chapter, I suggested that future work on determining the molecular mechanism of sex determination in this clade should focus on this.

### 6.4. Evolution of sex chromosomes under PGE

In Chapter V, I investigated how sex chromosomes evolve under the unusual inheritance system of dark-winged fungus gnats and gall midges. The ‘Faster-X’ theory predicts accelerated rate of evolution of X chromosomes relative to autosomes, in species with more conventional genetic sex determination, due to the effects of hemizygosity on selection of X-linked genes (Charlesworth *et al.*, 1987). However, the results are usually challenging to interpret, and the causes difficult to disentangle (Charlesworth *et al.*, 2018). Dark-winged fungus gnats and gall midges have germline paternal genome elimination (PGE), which results in equal effective population sizes for X chromosomes and autosomes (in species in which the X chromosomes are fully recombining, i.e. digenic species). As such, the efficacy of selection should be equal on the X and autosomes, and hemizygous selection in recessive mutations should be the only factor driving observed differences. I found that under PGE, X chromosomes appear to show slower divergence, likely driven by a predominance of purifying selection acting on recessive deleterious alleles.
Moreover, a broader aim of this chapter was to provide insights into how sex chromosomes evolve in general. Because under PGE, X chromosomes and autosomes should have equal effective population sizes, this makes the system immutable to the many processes that affect or are affected by effective population size. Factors like the relative rates of recombination of sex chromosomes and autosomes, reproductive success, and relative mutation rates, should affect the X and autosomes equally under PGE. As such, these systems are uniquely placed to inform us about how processes such as purifying selection, background selection, and adaptive evolution, can drive patterns of genetic diversity. Future work should aim to examine and disentangle these processes and broaden the study of chromosome evolution in other PGE and non-PGE species in the wider clade. In particular, studies should focus the evolution of sex chromosomes in the intermediate (fungus gnats) and outgroup (wood gnats) families, for which there are currently no chromosome-level genomes (Figure 6.2).

**Figure 6.2.** Phylogeny of the superfamily Sciaroidea, based on Anderson et al., (2022). Dark-winged fungus gnats (Sciaridae) and gall midges (Cecidomyiidae) have germline paternal genome elimination, while the other families shown do not. Image credit: Sciaridae (*Sciara humeralis*), Jean Marc Ruiz; Keroplatidae (*Platyura marginata*), James Lindsey; Mycetophilidae
(Leia varia), Luann Wright; Cecidomyiidae (Sitodiplosis mosellana), Giles San Martin; Anisopodidae (Sylviola fuscatus), James Lindsey.

6.5. Conclusion and outlook

In this thesis, I aimed to further our understanding of the unusual sex determination and sex chromosomes of dark-winged fungus gnats, and in doing so, make a contribution towards understanding of sex determination, sex chromosome evolution, and the processes involved in generating their diversity. However, there remain virtually endless features of this system to study and understand, which will be aided by the study of more species, the generation of more genomes and genomic data, and the production of new genetic tools in this system.

More generally, systems with unusual sex determination or sex chromosomes may be informative about why, in some cases, we see extensive diversity, while in others, we do not. To understand why biological ‘norms’ exist, it may be necessary to study deviations from the norm: where do biological constraints lie, and why? What gives way to diversity, and how? As Metz (1938) stated,

“…the peculiarities in [Bradysia] serve to throw light on special subjects of general biological interest.”

Dark-winged fungus gnats, as well as the other systems discussed in this thesis, deviate from the norm in many ways: by the means in which sex is determined, in the evolutionary dynamics of their sex chromosomes, in the sex ratios they produce, and in the inheritance of their genome. Of the unusual features of several Bradysia species, Metz (1938) went on to say,

“Such phenomena show that we are dealing with a group of organisms in which wide and apparently fundamental modifications of the genetic mechanism have become established. In some respects these phenomena violate apparently well-established principles of chromosome behaviour, and reveal what may be called a greater flexibility or wider range of potentiality in the genetic mechanism than had formerly been considered possible.”
Furthermore, it is tempting to speculate that all these unusual features are linked somehow, but it is still puzzling to imagine how. The study of more species and sequencing of more genomes, along with a larger and more connected research community, will help to resolve this. The above statement serves to point out that what had previously been considered fundamental genetic mechanisms were largely those that had been discovered in *Drosophila*, and subsequently became established as the wisdom of the time. Only a very small fraction of organisms have been well-studied by biologists, and the already apparent diversity in sex determination, and indeed wider genetic mechanisms, shows how often what may be considered the ‘norm’ may, in fact, be the exception.

6.6. References


Supplementary materials
Supplementary materials for Chapter II. Recent evolution of a maternally acting sex-determining supergene in a fly with single-sex broods

Supplementary figures

**Supplementary text 2.1. Mapping rates of X0, XX and X’X libraries.**

In the *B. coprophila* system, males are X0 and females are either XX or X’X; the X and X’ are distinguished by a large non-recombining region. X0 and XX individuals have essentially the same genome. Therefore, when mapping DNaseq libraries of each genotype to the reference genome, which is assembled from male data and so has the X but not the X’ (Urban et al., 2021; 2022), mapping rates are expected to be the same between X0 and XX libraries. X’X libraries should have a slightly lower mapping rate due to some expected level of divergence between the X and X’ chromosomes. However, while we found lower mapping rates for the two X’X libraries (95.64% and 95.96%) compared to the XX libraries (96.76% and 96.61%), we unexpectedly found that X0 libraries had the lowest mapping rates (93.23% and 93.85%). This ~3% difference in mapping rate, consistent across replicates, warrants an explanation.

To resolve this question, we focused on two potential explanations for the observed mapping rates. Firstly, the differences may be due to human and/or microbial contamination or differences in microbiome content between the sexes. Secondly, the male libraries may contain a higher proportion of reads that originate from the germline-restricted chromosomes (GRCs, Hodson et al., 2022). Though both sexes possess GRCs, and although the germline likely constitutes a small proportion of total cells, males have many more sperm cells than females do egg cells. Furthermore, sperm cells have two copies of the GRCs while eggs only have one (Goday and Esteban 2000). Since the GRCs are not included in the *Bradysia coprophila* reference genome, we reasoned that some of the difference may be due to a higher GRC content in the male libraries.
To assess the extent of contamination, we analyzed the unmapped reads from each library. Unmapped read were extracted from BWA-mapped (Li 2013) alignment files using SAMtools (Li et al. 2009). Kraken2 (v 2.1.2, Wood et al. 2019) was used to analyze read pairs and singletons separately using the PlusPF 20210517 database (https://genome-idx.s3.amazonaws.com/kraken/k2_pluspf_20210517.tar.gz). Sensitive and lightweight parameters were used (-confidence 0 –minimum-hit-groups 1 –memory-mapping). Classified and unclassified reads were written to FASTQ files (--classified-out / --unclassified-out). The Pavian R package (Breitwiese and Salzberg 2016) was used to interpret and plot the Kraken2 results. The contamination report showed that the X0 libraries had a higher proportion of contaminating chordate (likely human) and bacterial DNA compared to the XX and X'X libraries. Moreover, approximately 2.93% of read pairs from the X0 library were identified as contaminants (of which the vast majority was bacterial), compared to 1.30% and 1.44% for the XX and X'X libraries, respectively (Supplementary Table 2.7).

In an attempt to classify some of the remaining unmapped reads, 1000 unmapped reads that were not classified by Kraken2 were randomly sampled from each mate in unmapped-unclassified pairs separately and independently. Each random sample of unmapped-unclassified reads was subject to mapping with BLAST (v 2.13.0+, Altschul et al. 1990) to the full NT dataset (downloaded/updated 2023-03-29) using sensitive parameters for short sequences. The taxonomic identifier (taxid) from the top hit for each read was used to classify them, using custom Python scripts (https://github.com/JohnUrban/sciara-project-tools/tree/master/taxon) that depend on the ‘cogent’ Python library (Knight et al. 2007), specifically returning the node in the taxonomic tree closest to ‘Bradysia coprophila’. If the closest node was at least as close to Bradysia as Arthropoda (e.g. Insecta or Diptera), then the read was not considered a contaminant. Using this method and extrapolating to the rest of the input libraries, we classified a further 2.51% of X0 reads as contaminants, compared to 1.25% and 1.29% for the XX and X'X libraries, respectively (Supplementary Table 2.7). Together, from Kraken2 and BLAST, we estimate that 89.6%, 87.3% and 85.5% of the X0, XX and X'X unmapped reads, respectively, were from contaminating sources, which translates to 5.4%, 2.5% and 2.7% of the corresponding input libraries. Thus, contamination explains ~92-94% of the difference in mapping rates between males and females.
We also mapped the remaining unmapped-unclassified reads to the GRC scaffolds (Hodson et al., 2022) using BWA-MEM (Li 2013) and obtained mapping rates using SAMtools (Li et al. 2009). We identified more GRC reads in the male libraries compared to the female libraries, though they only comprised a small proportion of total reads (0.30% for X0, 0.09% for XX and 0.12% for X'X). In total, we explained approximately 90-95% of the unmapped reads as contaminants or GRC in origin, with the vast majority being explained by bacterial contamination.

**Supplementary text 2.2. De novo assembly of reads from X’X individuals.**

Initially, PacBio sequel 3.0 reads from X’X females were assembled *de novo* using wtdbg2 (Ruan and Li 2020) with a lower read length threshold of 2500bp, into 2382 contigs spanning 307Mb with an N50 of 353Kb. The assembly was polished twice with the PacBio reads and three times with 150bp Illumina reads also from X’X individuals using Racon (Vaser et al. 2017). Blobtools v1.1.1 (Laetsch and Blaxter 2017) and BUSCO v4 with the insecta_odb10 dataset (Seppey et al. 2019) were used to assess assembly quality and a custom R script was used to remove contaminant and low-coverage contigs. The final BUSCO score was 96.9% (93.5% complete and single-copy BUSCOs, 3.4% complete and duplicated BUSCOs). Scaffolds were assigned to chromosomes based on differences in X0 male and X’X female read coverage. To this end, reads were mapped to the assembly using BWA-MEM (Li 2013) and per-based genome coverage was calculated using BEDTools (Quinlan and Hall 2010). A custom R (R Core Team 2022) script was used to assign scaffolds to chromosomes: scaffolds with a relative coverage of 2x in both males and X’X females were assigned as autosomal; scaffolds with a relative coverage of 1x in males and 1-2x in X’X females were assigned to the X chromosome; scaffolds with low (<0.5x) male coverage and 1x coverage in females were assigned to the inversion (**Supplementary Figure 2.1**). A custom R script was used to calculate and plot mean coverage across 20Kb windows of the assembly. In total, only 3.63Mb of sequence was assigned as X’ inversion sequence using this method. The amount of sequence assigned to the autosomes and X chromosome was similar to that assigned in (Urban et al. 2021a).
**Supplementary text 2.3.** Binning reads with K-mers.

To assign K-mers as X'-specific, 27-mers were counted in 150bp Illumina read libraries for female-producing female (X'X) and male (X0, from 9) datasets using KMC (Kokot et al. 2017). 27-mers with a frequency of 0 in the X0 library and >5 in the X'X library (to exclude K-mers containing read errors) were assigned as X'-specific; this decision was based on the distribution of 27-mer frequencies visible in Figure 2.4. The cookiecutter (Starostina et al. 2015) option ‘extract’ was used to pull reads from the raw read file. Cookiecutter creates two separate files: one for all reads that contain a matching K-mer and one containing reads whose mate contained a matching K-mer; both files were combined into one because a read should originate from the same chromosome as its mate. This was done for both 150bp and 75bp Illumina libraries that were available. Extracting putative X’ reads with the X’-specific 27-mer library resulted in around 10-12% of reads being pulled; this was roughly in line with the expected size of the X’ relative to the rest of the haploid genome (around 40-60Mb in the expected 330-350Mb X’X genome, or 12-17%). When attempting the same process for PacBio sequel 3.0 error-prone reads, significantly more reads were pulled than expected: around 23% of X’-specific reads. We also identified putative autosomal- and X-specific K-mers based on the frequencies visible in Figure 2.4, similar to the methods used in (Hodson et al. 2022), and used these to extract putative autosomal and X reads. For Illumina reads the results were broadly as expected: around 60% for autosomal reads and 13% for X reads. Longer K-mer lengths resulted in very few reads being extracted. It is likely that shorter K-mers occur more often by chance in long reads and result in undesired reads being pulled, while longer K-mers are unlikely to have any matches due to the high error rate of the PacBio reads. For this reason, we decided to competitively map the PacBio reads post-assembly and use them for plugging gaps (Supplementary Figure 2.4).

**Supplementary text 2.4.** Identification of degraded genes of interest.

To identify degraded genes that may be of interest, we searched among functional annotations of genes that we classified as degraded using keywords. To identify genes potentially involved in sex determination by chromosome elimination, the keywords used were: ‘chromosome’, ‘chromatid’, ‘chromatin’, ‘cohesion’, ‘condensation’, ‘segregation’, ‘centromere’, ‘centrosome’
and ‘spindle’. To identify candidates responsible for the wavy wing mutation, we searched for the keyword ‘wing’. To identify potential genes with roles in male fertility, the keywords used were ‘sex’, ‘sperm’, ‘fertility’, and ‘meiosis’.

**Supplementary text 2.5. Fly husbandry.**

*Bradysia* (formerly *Sciara* *coprophila*) flies have been reared in laboratory conditions since the 1920s (Metz 1925). Colonies can be successfully maintained in a variety of ways but are healthiest when kept at around 18–21°C and around 70% relative humidity, and when conditions are not too crowded. At Edinburgh University, colonies are maintained using matings between a single female and two males (to protect against sterility) and are reared on biological agarose in 25mm x 95mm glass vials at 18°C. At the Carnegie Institution for Science in Baltimore, colonies are maintained using mass matings controlled for the sex of offspring (5 males plus 5 female-producing or 5 male-producing females) and are reared in 60z Square Bottom Polypropylene Bottles (Genesee Scientific) capped with Flystuff Flugs (Genesee Scientific) at 21°C and >70% relative humidity. At both institutions, flies are kept on 2.2% Bactoagar and are fed a mixture of yeast, powered mushroom, powdered spinach or nettle and ground straw while larvae.

Hungerford (Hungerford 1916) first reported the generation time of *B. coprophila* (from fertilized egg to eclosion) at between 24 and 32 days. Rieffel and Crouse (Rieffel and Crouse 1966) timed developmental stages so as to reflect the minimum length of each stage in ideal conditions, i.e. when crowding, food scarcity, competition and temperature do not interfere with normal development. These minimum developmental stages summed together are 24.6 days for males and 26.6 days for females. Having raised generations of *B. coprophila* in the lab for several years, the authors of this paper have observed the minimum generation time as approximately 28 days, though adults from a single progeny often continue to eclose for up to a week. The generation time fluctuates with temperature: the authors have previously raised *B. coprophila* at 25°C, which speeds up development by at least a few days, and it has previously been noted that rearing *B. coprophila* and *B. ocellaris* at temperatures as low as 12°C is possible but significantly delays development (Smith-Stocking 1936; Nigro et al. 2007). *B. coprophila* have also been observed to survive up to several months when refrigerated (R. B. Baird, C. N. Hodson & L. Ross, unpublished observations; 8). This potential adaptation to cold temperatures
may suggest that larvae lie dormant for long periods in winter, which might increase the average generation time. Harsher conditions such as crowding and food scarcity do also appear to extend the generation time beyond 40 days (J. M. Urban & R. B. Baird, unpublished observations). On the other hand, the generation time may be shorter in regions with warmer climates. Observations of other, closely-related species report varying generation times: 26-48 days or *B. odoriphaga* (Li et al. 2015), 14 days for *B. impatiens* (Kennedy 1973), 25 days at 20-25°C or 3-4 weeks at 22-24°C for *B. paupera* (Mansilla et al. 2001) and 26-28 days at 25°C for *B. difformis* (Villanueva-Sánchez et al. 2013); see (Katumanyane et al. 2018 for a review). Overall, it is difficult to estimate the generation time that best approximates *B. coprophila* development in nature, but in light of the literature cited here and observations made by the authors we decided to use a generation time 24-40 days to calculate divergence between the X and X’ chromosomes in *B. coprophila* in years.

**Supplementary text 2.6. Genome annotation.**

The RNAseq data used in annotation of the genome (including autosomes II, III, IV, X and the X’ sequence) was obtained from publicly available data previously produced by Urban *et al.* (2015; 2021a; Urban 2021b) and unpublished datasets produced by the Ross Lab. This included 18 RNAseq datasets from four life stages (3 x male and 3 x female embryos, 2 x male and 2 x female larval, 2 x male and 2 x female pupal, 2 x male and 2 x female adult, (Urban et al. 2021b), 15 datasets from dissected larval salivary glands (pooled individuals, (Urban et al. 2015), 6 datasets from radiated larvae (pooled female individuals, (Urban et al. 2021a), 6 datasets from male and female somatic tissue (3 x male, 3 x female larval/early pupal), as well as 12 datasets from early (0-8h) embryos (6 x male, 6 x female, unpublished data). All RNAseq datasets were aligned to the genome using STAR (Dobin et al. 2013) prior to being fed to BRAKER2. Homology-based datasets included all OrthoDB v10 Diptera protein sequences (751660 sequences, (Kriventseva et al. 2019), all Uniprot Diptera protein sequences (7053 sequences, (Uniprot Consortium 2015) and all Refseq Diptera protein sequences including all non-canonical isoforms (1575334 sequences, (Pruitt et al. 2007).

Functional information for the 26887 protein sequences in the BRAKER2 gene annotation set was obtained by finding the best BLASTP (Altschul et al. 1990) hits in several
protein databases and with InterProScan. Specifically, all BLASTP hits (-word_size 3 -evalue 1e-2) for all 26887 proteins were found in (i) the *Drosophila melanogaster* r6.45 (dos Santos et al. 2014) proteome, (ii) Non-redundant UniProtKB/SwissProt (Uniprot Consortium 2020) protein sequences (ftp://ftp.ncbi.nlm.nih.gov/blast/db/swissprot.tar.gz), (iii) the NCBI Landmark database for SmartBLAST (ftp://ftp.ncbi.nlm.nih.gov/blast/db/landmark.tar.gz), (iv) the NCBI RefSeq (O’Leary et al. 2015) protein database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/refseq_protein.*.tar.gz), (v) the NCBI Non-Redundant (nr) protein database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/nr.*.tar.gz), (vi) OrthoDB v10 (Kriventseva et al. 2019), (vii) and gene annotations from the original *Bradysia coprophila* reference genome (40). The best BLASTP hit for all 26887 proteins in each database was found by taking the hit with the highest bitscore. Hits with bitscores lower than 50 or e-values higher than 0.0005 were removed from consideration. The 26887 protein sequences were also extensively analyzed using the InterPro (Blum et al. 2020) protein family and domain database and InterProScan version 5.56-89.0 (-dp -iprlookup -goterms -f tsv,xml,json,gff3, (Jones et al. 2014) allowing all analyses to be run: CDD, Coils, Gene3D, Hamap, MobiDBLite, PANTHER, Pfam, Phobius, PIRSF, PIRSR, PRINTS, ProSitePatterns, ProSiteProfiles, SFLD, SignalP, SMART, SUPERFAMILY, TIGRFAM, TMHMM. Software licenses were obtained where relevant (SignalP, Phobius, TMHMM).

**Supplementary text 2.7. Transposable element annotation.**

To annotate transposable elements (TEs) in the genome, the module reasonaTE of the TranposonUltimate v1.03 pipeline (Riehl et al. 2022) was used (Supplementary Text 2.7). Combining the different annotation approaches and sensitivities of RepeatMasker (v4.1.2, (Smit et al. 2015)), RepeatModeler (v2.0.2a, (Flynn et al. 2020)), LTRharvest (2.9.0, (Ellinghaus et al. 2008)), TIRvish (v.1.10.5, (Gremme et al. 2013)), SINE-Scan (v1.1.2, (Mao and Wang 2017)), HelitronScanner (v1.0.0, (Xiong et al. 2014)), MiteFinderII (v1.0.0, (Hu et al. 2018)), MITE-Tracker (v1.0.1, (Crescente et al. 2018)), TransposonPSI (v1.0.0, https://transposonpsi.sourceforge.net/) and NCBICDD1000 (v1.0.1, (Lu et al. 2020) it enables in-depth analysis of various classes of TEs in the genome. R Studio (R Core Team 2022) was used to plot TE distribution across chromosomes.
Supplementary text 2.8. Binning RNAseq reads to identify silenced genes and examine dosage compensation.

To identify silenced genes, we parsed X and X’ RNAseq reads using elements of a pipeline from (Marshall et al. 2020, https://github.com/MooHoll/Parent_of_Origin_Expression_Bumblebee) to avoid mismapping between the two chromosomes. To this end, we created an N-masked version of the X chromosome to mask SNVs between the X and X’ by mapping reads from X’X individuals to the X0 reference genome (Urban et al. 2021b) using Bowtie2 (Langmead and Salzberg 2012), processed BAM files using Picardtools (Anon 2019) and SAMtools (Li et al. 2009) and called and filtered SNVs using freebayes (Garrison and Marth 2012) and VCFtools (Danecek et al. 2011) respectively. BEDtools (Quinlan and Hall 2010) was used to identify X- and X’-specific variants and to N-mask those variants on the X chromosome prior to mapping RNAseq reads. We mapped RNAseq reads from pooled female data from four life stages: embryo, larva, pupa and adult (Urban et al. 2021b) to the N-masked genome using STAR (Dobin et al. 2013) and subsequently assigned RNAseq reads to the X or X’ chromosome using SNPsplit (Krueger and Andrews 2016). Gene expression was quantified using Kallisto (Bray et al. 2016) with the 2321 singlecopy homologs we identified (see main text) as a Kallisto index, and counts were normalised using EdgeR (Robinson et al. 2010). Genes with zero counts of TPM (transcripts per million) were assumed to be non-expressed. We also included genes with TPMs in the bottom 0.1% of non-zero TPM counts within a sample as non-expressed to account for stochastic mismapping of RNAseq reads. Analysis and plotting of counts was carried out with R Studio (R Core Team 2022) with use of the ggplot2 package (Wickham 2016).

To examine dosage compensation of X-linked genes with degraded X’-linked homologous in the X’X females, we applied the same pipeline to extract RNAseq reads originating from the X chromosome in those females. Kallisto (Bray et al. 2016) was used to quantify the resulting X reads against the 2321 X-linked single-copy homologs and normalization between samples was carried out with EdgeR (Robinson et al. 2010). We subsequently compared expression of X-linked genes in X’X versus XX females, with the expectation that X-linked genes with degraded X’ homologs should be upregulated in X’X females if they are dosage compensated.
Supplementary text 2.9. Data used to examine dosage compensation.

To investigate potential dosage compensation (DC) of X-linked genes with degraded X' homologs, we used RNAseq data from early (0-8h) embryos. This dataset was generated by R. B. Baird for the purpose of another, unpublished study. Adult females were mated with males (generally ~10 females with ~3 males in a tube), and after ~18-24 hours, females were pinned to a petri dish containing 2.2% Bactoagar and egg laying was induced by crushing the head slightly. An entire clutch (50-100 eggs) is usually laid within the next 2 hours. Three replicates from two time stages were collected: 0-4h and 4-8h after egg deposition (AED), for the eggs of each female genotype (X'X and XX), resulting in 12 samples in total. For each sample, around ~600-800 eggs pooled from ~15 individuals were collected and sequenced for 15Gb (50m reads) of paired-end 150bp RNAseq reads on the Illumina Novaseq S1 platform.

We used these data to examine expression of X-linked genes in X'X versus XX females because data for female adults (or other developmental stages) separated by genotype are not yet available, and these early embryos are likely to contain exclusively maternal transcripts. In Drosophila, zygotic genome activation (ZGA) occurs in two waves at mitotic cycle 8 (60 genes) and following cellularization at mitotic cycle 14 (over 1000 genes, 22). De Saint Phalle & Sullivan (de Saint Phalle and Sullivan 1996) report that cellularization in B. coprophila occurs during interphase of nuclear cycle 11, which begins at 9.3 +/- 1.1 hours AED. Cell cycle 12 begins at 11.4 +/- 1.4 hours AED. However, as in Drosophila, de Saint Phalle & Sullivan (de Saint Phalle and Sullivan 1996) report that the germline nuclei cellularise slightly earlier, at cell cycle 7, which occurs around 4.2 +/- 0.5 hours AED. Nonetheless, if we assume that, as in Drosophila, ZGA occurs approximately when cellularization occurs, ZGA should begin for the majority of somatic genes by 12 hours AED.

Upon splitting RNAseq reads from the eggs of X'X mothers into X' and X reads (see methods in Supplementary Text 2.6), similar proportions of reads were assigned to the X' (5.67%) and X (6.17%) aside from one outlier sample (eggs of X'X females, 4-8h, replicate 2), which had 7.97% of reads assigned to the X and 3.61% assigned to the X'). This sample was excluded from further analysis. If transcripts were zygotic rather than maternal, then the number of X reads assigned should be three-fold higher for the X, because in the pooled eggs, maternal transcripts from X'X mothers will originate from the X' and X in equal proportions.
However, following ZGA, three times as many X transcripts should be produced for every X' transcript, owing to half the eggs being XX and the other half X'X.

**Supplementary text 2.10. Removing adapter sequences found in the X' scaffold.**

Upon submission of the X' inversion scaffold to NCBI, five sequences were flagged as potential adapter contamination. All five sequences were shorter than 48 bp, at contig ends immediately adjacent to a gap, and were at the following locations within the scaffold: 8,259,796-8,259,842; 12,677,415-12,677,459; 30,859,720-30,859,766; 44,216,421-44,216,468; 53,963,934-53,963,981. These sequences may were present in the assembly prior to scaffolding/gap-filling/polishing, so may have been missed by the trimming software. We masked out these short sequences by extending the adjacent gaps to cover them as well.
Supplementary tables

**Supplementary table 2.1.** Assembly statistics and chromosome anchorage for the *B. coprophila* genome assembled *de novo* from PacBio long reads from X’X individuals.

<table>
<thead>
<tr>
<th>Partition</th>
<th>Length (Mb)</th>
<th>N scaffolds</th>
<th>N50 (Kb)</th>
<th>Largest scaffold (Kb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autosomes</td>
<td>215.15</td>
<td>1123</td>
<td>423.42</td>
<td>2141.89</td>
</tr>
<tr>
<td>X chromosome</td>
<td>72.84</td>
<td>864</td>
<td>205.13</td>
<td>995.08</td>
</tr>
<tr>
<td>Inversion</td>
<td>3.63</td>
<td>259</td>
<td>17.69</td>
<td>73.99</td>
</tr>
<tr>
<td>Total</td>
<td>291.63</td>
<td>2246</td>
<td>342.36</td>
<td></td>
</tr>
</tbody>
</table>
**Supplementary table 2.2.** Data used to calculate age estimates for each stratum, and upper and lower estimates for neutral estimates (in millions of years, based on different estimates of mutation rate and generation time). Note that the numbers of synonymous sites are not integers because the weight of each site is normalised in accordance with codon degeneracy (see methods).

<table>
<thead>
<tr>
<th>Stratum</th>
<th>Start (bp)</th>
<th>End (bp)</th>
<th>( D_{xy} ) (het density)</th>
<th>N single-copy homologs within boundaries</th>
<th>N syn sites</th>
<th>N syn variants</th>
<th>Lower ( D_{xy} ) estimate</th>
<th>Upper ( D_{xy} ) estimate</th>
<th>Lower neutral sites estimate</th>
<th>Upper neutral sites estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>4050001</td>
<td>13050000</td>
<td>0.0075</td>
<td>326</td>
<td>155620</td>
<td>1533</td>
<td>0.050</td>
<td>0.147</td>
<td>0.075</td>
<td>0.219</td>
</tr>
<tr>
<td>S2</td>
<td>13050001</td>
<td>13550000</td>
<td>0.0040</td>
<td>15</td>
<td>11848.98</td>
<td>60</td>
<td>0.027</td>
<td>0.078</td>
<td>0.05</td>
<td>0.145</td>
</tr>
<tr>
<td>S3</td>
<td>13550001</td>
<td>37350000</td>
<td>0.0123</td>
<td>904</td>
<td>460851.1</td>
<td>6903</td>
<td>0.083</td>
<td>0.241</td>
<td>0.113</td>
<td>0.329</td>
</tr>
<tr>
<td>S4</td>
<td>37350001</td>
<td>38350000</td>
<td>0.0039</td>
<td>35</td>
<td>17063.14</td>
<td>95</td>
<td>0.026</td>
<td>0.077</td>
<td>0.034</td>
<td>0.099</td>
</tr>
<tr>
<td>S5</td>
<td>38350001</td>
<td>40000000</td>
<td>0.0151</td>
<td>55</td>
<td>19150.58</td>
<td>420</td>
<td>0.101</td>
<td>0.295</td>
<td>0.149</td>
<td>0.434</td>
</tr>
<tr>
<td>S6</td>
<td>40000001</td>
<td>40100000</td>
<td>0.0074</td>
<td>2</td>
<td>1162.64</td>
<td>3</td>
<td>0.049</td>
<td>0.144</td>
<td>0.05</td>
<td>0.147</td>
</tr>
<tr>
<td>S7</td>
<td>40100001</td>
<td>49500000</td>
<td>0.0145</td>
<td>350</td>
<td>158696.8</td>
<td>3204</td>
<td>0.097</td>
<td>0.284</td>
<td>0.148</td>
<td>0.432</td>
</tr>
<tr>
<td>S8</td>
<td>49500001</td>
<td>49600000</td>
<td>0.0006</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.004</td>
<td>0.011</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>S9</td>
<td>49600001</td>
<td>52150000</td>
<td>0.0153</td>
<td>62</td>
<td>30364.9</td>
<td>692</td>
<td>0.102</td>
<td>0.298</td>
<td>0.162</td>
<td>0.474</td>
</tr>
<tr>
<td>S10</td>
<td>52150001</td>
<td>52550000</td>
<td>0.0122</td>
<td>22</td>
<td>7521.66</td>
<td>111</td>
<td>0.082</td>
<td>0.239</td>
<td>0.119</td>
<td>0.346</td>
</tr>
<tr>
<td>S11</td>
<td>52550001</td>
<td>54050000</td>
<td>0.0145</td>
<td>50</td>
<td>26014.18</td>
<td>427</td>
<td>0.097</td>
<td>0.284</td>
<td>0.141</td>
<td>0.412</td>
</tr>
<tr>
<td>S12</td>
<td>54050001</td>
<td>55950000</td>
<td>0.0159</td>
<td>96</td>
<td>39556.22</td>
<td>873</td>
<td>0.107</td>
<td>0.312</td>
<td>0.171</td>
<td>0.499</td>
</tr>
<tr>
<td>S13</td>
<td>55950001</td>
<td>57250000</td>
<td>0.0136</td>
<td>90</td>
<td>44562.88</td>
<td>830</td>
<td>0.091</td>
<td>0.266</td>
<td>0.132</td>
<td>0.386</td>
</tr>
<tr>
<td>S14</td>
<td>57250001</td>
<td>58450000</td>
<td>0.0104</td>
<td>36</td>
<td>14329.12</td>
<td>236</td>
<td>0.070</td>
<td>0.204</td>
<td>0.146</td>
<td>0.425</td>
</tr>
<tr>
<td>S15</td>
<td>58450001</td>
<td>61950000</td>
<td>0.0096</td>
<td>97</td>
<td>38924.46</td>
<td>429</td>
<td>0.065</td>
<td>0.188</td>
<td>0.075</td>
<td>0.218</td>
</tr>
<tr>
<td>S16</td>
<td>61950001</td>
<td>62900000</td>
<td>0.0093</td>
<td>54</td>
<td>22126.82</td>
<td>346</td>
<td>0.062</td>
<td>0.182</td>
<td>0.102</td>
<td>0.297</td>
</tr>
</tbody>
</table>
**Supplementary table 2.3.** Number of genes within the X' supergene sequence with each type of predicted pseudogenising mutation.

<table>
<thead>
<tr>
<th>Mutation type</th>
<th>N genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stop codon gained</td>
<td>37</td>
</tr>
<tr>
<td>Stop codon lost</td>
<td>10</td>
</tr>
<tr>
<td>Start codon lost</td>
<td>10</td>
</tr>
<tr>
<td>Frameshift</td>
<td>113</td>
</tr>
<tr>
<td>Frameshift and stop codon gained</td>
<td>15</td>
</tr>
<tr>
<td>Frameshift and stop codon lost</td>
<td>6</td>
</tr>
<tr>
<td>Frameshift and start codon lost</td>
<td>8</td>
</tr>
<tr>
<td>Frameshift and stop codon gained and stop codon lost</td>
<td>2</td>
</tr>
<tr>
<td>Frameshift and stop codon gained and start codon lost</td>
<td>1</td>
</tr>
</tbody>
</table>
**Supplementary table 2.4.** Functions of pseudogenised genes of interest.

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Functionality</th>
<th>Annotation(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>jg4815</td>
<td>nonfunctional_disrupted</td>
<td>Similar to how Protein held out wings</td>
</tr>
<tr>
<td>jg6281</td>
<td>nonfunctional_silenced</td>
<td>Similar to su(Hw) Protein suppressor of hairy wing</td>
</tr>
<tr>
<td>jg5727</td>
<td>nonfunctional_disrupted</td>
<td>Similar to su(Hw) Protein suppressor of hairy wing</td>
</tr>
<tr>
<td>jg7913</td>
<td>nonfunctional_disrupted</td>
<td>Similar to su(Hw) Protein suppressor of hairy wing</td>
</tr>
<tr>
<td>jg7400</td>
<td>nonfunctional_disrupted</td>
<td>Similar to stan Protocadherin-like wing polarity protein</td>
</tr>
<tr>
<td>jg6904</td>
<td>nonfunctional_disrupted</td>
<td>Similar to Etl1 SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A containing DEAD/H box 1 homolog; Similar to Marca1 SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A-like protein 1</td>
</tr>
<tr>
<td>jg6344</td>
<td>nonfunctional_disrupted</td>
<td>Similar to Etl1 SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A containing DEAD/H box 1 homolog; Similar to Marca1 SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A-like protein 1</td>
</tr>
<tr>
<td>jg7872</td>
<td>nonfunctional_silenced</td>
<td>Similar to dsx Protein doublesex</td>
</tr>
<tr>
<td>jg6244</td>
<td>nonfunctional_disrupted</td>
<td>Similar to Crocc Rootletin; Ciliary rootlet component, centrosome cohesion</td>
</tr>
<tr>
<td>jg6043</td>
<td>nonfunctional_silenced</td>
<td>Similar to TRAF3IP1 TRAF3-interacting protein 1; Microtubule-binding protein MIP-T3 C-terminal region</td>
</tr>
<tr>
<td>jg5804</td>
<td>nonfunctional_disrupted</td>
<td>Similar to ncd Protein claret segregational</td>
</tr>
</tbody>
</table>
**Supplementary table 2.5.** Upregulated X-linked genes with single-copy X'-linked homologs in X'X females.

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Annotation(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>jg4857</td>
<td>Similar to PARP3 Protein mono-ADP-ribosyltransferase PARP3</td>
</tr>
<tr>
<td>Jg5988</td>
<td>Similar to srfbp1 Serum response factor-binding protein 1; Similar to F52C9.6 Putative uncharacterised transposon-derived protein F52C9.6</td>
</tr>
<tr>
<td>Jg6324</td>
<td>Protein of unknown function</td>
</tr>
<tr>
<td>Jg7315</td>
<td>Similar to eff Ubiquitin-conjugating enzyme E2-17 kDa</td>
</tr>
</tbody>
</table>
Supplementary Table 2.6. SV signal breakpoints associated with annotated repeats. 75% of the SV signals from discordant paired-end short read alignments had one or more breakpoints inside annotated repeats (compared to 59% at random), suggesting repeats (e.g. transposon sequences) are differentially distributed across the X’ compared to the X. These discordant paired-end mappings are not simply a problem of multi-mapping reads since (i) the Smoove pipeline filtered out anything with MAPQ < 20 and required multiple paired-end or split reads to support each breakpoint, and (ii) the XO and XX controls did not have nearly as many SV signals using the Smoove pipeline, showing that the widespread discordance is a property of the X’X sample alone (see Figures 2A and S1). Overall, up to 75% of the SV signals may involve transpositions and/or repeat-involved rearrangements unique to the X’, placing repeat sequences in contexts they are not found in on the X. That still leaves a minimum of 25% of SV signals involved strictly with “unique sequences” (not annotated as repeats) putatively arising from other structural rearrangements such as inversions.

<table>
<thead>
<tr>
<th>SV Classification</th>
<th>Count on X</th>
<th>Count on X with one breakpoint located in a repeat</th>
<th>Count on X with both breakpoints located in repeats</th>
<th>Percent with one breakpoint located in a repeat</th>
<th>Percent with both breakpoints located in repeats</th>
<th>Percent with one or both breakpoints located inside a repeat</th>
</tr>
</thead>
<tbody>
<tr>
<td>INV</td>
<td>29</td>
<td>10</td>
<td>8</td>
<td>18</td>
<td>34.4828</td>
<td>27.5862</td>
</tr>
<tr>
<td>DEL</td>
<td>2697</td>
<td>987</td>
<td>1044</td>
<td>2031</td>
<td>36.5962</td>
<td>38.7097</td>
</tr>
<tr>
<td>DUP</td>
<td>56</td>
<td>21</td>
<td>14</td>
<td>35</td>
<td>37.5</td>
<td>25</td>
</tr>
<tr>
<td>TOTAL</td>
<td>2782</td>
<td>1018</td>
<td>1066</td>
<td>2084</td>
<td>36.5924</td>
<td>38.3178</td>
</tr>
</tbody>
</table>
**Supplementary table 2.7.** Classification of unmapped reads for each genotype.

<table>
<thead>
<tr>
<th>Genotype</th>
<th>X0</th>
<th>XX</th>
<th>X'X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean proportion of unmapped reads (%)</td>
<td>6.46</td>
<td>3.71</td>
<td>4.76</td>
</tr>
<tr>
<td>Mean proportion of unmapped read pairs (neither mapped, %)</td>
<td>6.07</td>
<td>2.91</td>
<td>3.19</td>
</tr>
<tr>
<td>Read pairs classified as contaminants by Kraken2 (% of total reads)</td>
<td>2.93</td>
<td>1.30</td>
<td>1.44</td>
</tr>
<tr>
<td>Read pairs classified as bacterial by Kraken2 (% of total reads)</td>
<td>2.89</td>
<td>1.27</td>
<td>1.41</td>
</tr>
<tr>
<td>Read pairs classified as chordate by Kraken2 (% of total reads)</td>
<td>0.0029</td>
<td>0.0020</td>
<td>0.0017</td>
</tr>
<tr>
<td>Read pairs classified as other (viral, fungal, protozoan) by Kraken2 (% of total reads)</td>
<td>0.0058</td>
<td>0.0047</td>
<td>0.0044</td>
</tr>
<tr>
<td>Mean proportion of unmapped-unclassified pairs identified as contaminants by BLAST (% of total reads)</td>
<td>2.51</td>
<td>1.25</td>
<td>1.29</td>
</tr>
<tr>
<td>Read pairs identified as contamination by Kraken2 and BLAST (% of total reads)</td>
<td>5.44</td>
<td>2.54</td>
<td>2.73</td>
</tr>
<tr>
<td>Read pairs identified as GRC (% of total reads)</td>
<td>0.30</td>
<td>0.09</td>
<td>0.12</td>
</tr>
<tr>
<td>Read pairs identified as contamination or GRC (% of total reads)</td>
<td>5.73</td>
<td>2.64</td>
<td>2.85</td>
</tr>
<tr>
<td>Proportion of unmapped read pairs explained (%)</td>
<td>94.45</td>
<td>90.56</td>
<td>89.41</td>
</tr>
</tbody>
</table>
Supplementary figures

**Supplementary figure 2.1.** Structural variant calls from the XX genotype from Illumina 75bp paired-end short reads. These calls serve as an extra control along with calls from the X0 genotype, which, when contrasted with calls from the X'X genotype, provide further support that the X' is enriched for complex rearrangements.

**Supplementary figure 2.2.** Expected (A) versus observed (B) differences in genomic coverage across the genome assembled *de novo* from PacBio reads from X'X individuals, indicating that the vast majority of reads from the X and X' chromosomes collapsed together upon assembly. Note that the assembly is 291.63 Mb in length but appears shorter in the figure above because some scaffolds were shorter than the 20 Kb windows across which coverage was calculated, so such scaffolds were not included.
Supplementary figure 2.3. Distribution of contig lengths from the raw, short read SPAdes assembly of the X’. (A) All contigs; (B) contigs equal to or over 1Kb in length; (C) contigs less than 1Kb in length.

Supplementary figure 2.4. PacBio long reads from X’X individuals mapped to the X’X assembly. Now that the X and X’ inversion are separately assembled, PacBio reads map to the correct chromosomes with approximately expected coverage and could thus be used to fill some remaining gaps in the assembly of the X’ inversion.
Supplementary figure 2.5. Differential expression smear plot for X-linked genes with single-copy X' homologs between the two types of females (XX versus X'X). Negative fold change (FC) represents upregulation of the gene copy in X'X females; positive FC represents upregulation of the gene in XX females. Transcript counts have been normalised such that overall X in X'X females expression equals that of XX females. Genes with significant fold change and significant adjusted P values are coloured red.

References


**Supplementary materials for Chapter III.** Why put all your eggs in one basket? Evolutionary perspectives on the origins of monogenic reproduction

Supplementary Tables

**Supplementary table 3.1.** Reported reproductive strategies across dark-winged fungus gnats (Sciaridae). *Brady sia, Phytosciara, Scatopsciara* and *Ctenosciara* belong to the monophyletic group Megalosphynae; *Lycoriella* belong to the non-monophyletic group Pseudolycoriella; *Corynoptera* belong to either Pseudolycoriella or the non-monophyletic Cratyninae (Shin et al. 2013). The positions of the *Cosmosciara, Hyperlasion* and *Rhynchosciara* genera are unclear.

<table>
<thead>
<tr>
<th>Genus</th>
<th>Species</th>
<th>Reported strategy</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Brady sia</em></td>
<td><em>coprophila / tilicola</em></td>
<td>Monogenic</td>
<td>(Metz 1938)</td>
</tr>
<tr>
<td></td>
<td><em>impatiens / difformis</em></td>
<td>Monogenic</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>ocellaris / tritici</em></td>
<td>Mixed</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>varians</em></td>
<td>Monogenic</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>reynoldsi</em></td>
<td>Digenic</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>prolica</em></td>
<td>Digenic</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>fenestralis</em></td>
<td>Mixed</td>
<td>(McCarmthy 1945a)</td>
</tr>
<tr>
<td></td>
<td><em>spatitergum</em></td>
<td>Monogenic</td>
<td>(Stefftan 1974)</td>
</tr>
<tr>
<td></td>
<td><em>molokaiensis</em></td>
<td>Monogenic</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>bishopi</em></td>
<td>Digenic</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>matrogressensis</em></td>
<td>Mixed</td>
<td>(Rocha and Perondini 2000)</td>
</tr>
<tr>
<td>Species</td>
<td>Type</td>
<td>Classification</td>
<td>Reference</td>
</tr>
<tr>
<td>-------------------------</td>
<td>----------</td>
<td>----------------</td>
<td>----------------</td>
</tr>
<tr>
<td>paupera</td>
<td>Monogenic</td>
<td>(Liu 2007)</td>
<td></td>
</tr>
<tr>
<td>odoriphaga</td>
<td>Mixed</td>
<td>(Cheng et al. 2017)</td>
<td></td>
</tr>
<tr>
<td>Phytosciara vulcanata</td>
<td>Digenic</td>
<td>(Steffan 1974)</td>
<td></td>
</tr>
<tr>
<td>Scatopsicara nacta</td>
<td>Monogenic</td>
<td>(McCarthy 1945a)</td>
<td></td>
</tr>
<tr>
<td>nigrita</td>
<td>Digenic</td>
<td>(Steffan 1974)</td>
<td></td>
</tr>
<tr>
<td>cunicularius</td>
<td>Mixed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ctenosciara hawaiensis</td>
<td>Monogenic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lycoriella similans</td>
<td>Mixed</td>
<td>(Metz 1938)</td>
<td></td>
</tr>
<tr>
<td>ingenua / mali / pauciseta</td>
<td>Digenic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>agraria</td>
<td>Digenic</td>
<td>(McCarthy 1945b)</td>
<td></td>
</tr>
<tr>
<td>solispina</td>
<td>Monogenic</td>
<td>(Steffan 1974)</td>
<td></td>
</tr>
<tr>
<td>hoyti</td>
<td>Digenic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>aurpilia</td>
<td>Mixed</td>
<td>(Binns 1980)</td>
<td></td>
</tr>
<tr>
<td>Corynoptera subtrivialis</td>
<td>Monogenic</td>
<td>(Metz 1938)</td>
<td></td>
</tr>
<tr>
<td>brevipalpis</td>
<td>Digenic</td>
<td>(Steffan 1974)</td>
<td></td>
</tr>
<tr>
<td>Cosmosciara perniososa</td>
<td>Digenic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rhynchosciara hollaenderi</td>
<td>Monogenic</td>
<td>(Mattingly and Dumont 1971)</td>
<td></td>
</tr>
<tr>
<td>americana / angelae</td>
<td>Monogenic</td>
<td>(Lara et al. 1965)</td>
<td></td>
</tr>
<tr>
<td>Hyperlasion wasmanni</td>
<td>Mixed</td>
<td>(Mohrig et al. 2019)</td>
<td></td>
</tr>
</tbody>
</table>
**Supplementary table 3.2.** Reported reproductive strategies across gall midges (Cecidomyiidae).

<table>
<thead>
<tr>
<th>Genus</th>
<th>Species</th>
<th>Reported strategy</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Mayetiola</em></td>
<td><em>destructor</em></td>
<td>Mixed</td>
<td>(Stuart and Hatchett 1991)</td>
</tr>
<tr>
<td><em>Rabdophaga</em></td>
<td><em>heterobia</em></td>
<td>Monogenic</td>
<td>(Barnes 1931)</td>
</tr>
<tr>
<td></td>
<td><em>saliciperda</em></td>
<td>Monogenic</td>
<td>(Kraczkiewicz 1966)</td>
</tr>
<tr>
<td></td>
<td><em>salisbatatas</em> / <em>batatas</em></td>
<td>Monogenic</td>
<td>(Geyer-Duszyńska 1961)</td>
</tr>
<tr>
<td><em>Dasineura</em></td>
<td><em>folliculi</em></td>
<td>Monogenic</td>
<td>(Dorchin et al. 2007)</td>
</tr>
<tr>
<td></td>
<td><em>carbonaria</em></td>
<td>Monogenic</td>
<td></td>
</tr>
<tr>
<td><em>Solidago</em></td>
<td><em>gigantea</em></td>
<td>Mixed</td>
<td></td>
</tr>
<tr>
<td><em>Izeniola</em></td>
<td><em>obesula</em></td>
<td>Monogenic</td>
<td>(Dorchin and Freidberg 2004)</td>
</tr>
<tr>
<td><em>Orseolia</em></td>
<td><em>Oryzivora</em></td>
<td>Monogenic</td>
<td>(Omoloye 2006)</td>
</tr>
<tr>
<td><em>Aphidoletes</em></td>
<td><em>aphidimyza</em></td>
<td>Monogenic</td>
<td>(Tabadkani et al. 2012)</td>
</tr>
<tr>
<td><em>Sitodiplosis</em></td>
<td><em>mosellana</em></td>
<td>Monogenic</td>
<td>(Smith et al. 2004)</td>
</tr>
<tr>
<td><em>Cystiphora</em></td>
<td><em>sonchi</em></td>
<td>Mixed</td>
<td>(Mcclay 1996)</td>
</tr>
</tbody>
</table>

**References**


Supplementary materials for Chapter IV. Molecular control of the sex ratio in dark-winged fungus gnats

Supplementary Figure 4.1. Log10 transcripts per million (TPM) for X and X’ alleles in each of the three 4-8h female embryo replicates. That distributions are similar and there are no obvious outliers with particularly high expression in replicate 2 suggests that the difference in the number of reads assigned to each allele was likely due to differences in developmental timing at the point of collection, rather than, for example, particularly high expression of only a few genes in this replicate. If this sample was collected at a later stage, after the onset of zygotic genome activation, then the amount of expression from X alleles relative to X’ alleles should be greater (note that this will not be visible in the above plots as X and X’ alleles are being compared with no autosomal reference). This is because the mother has the genotype X’X whereas half her offspring are XX and the other half are X’X. Also see Supplementary Table 4.1.

Supplementary Table 4.1. Number of genes with allele-specific expression in the three female 4-8h embryo RNAseq replicates. The fact that many more genes have X-specific expression in the outlying sample (replicate 2) suggests that the difference in the number of reads assigned for this sample was likely due an increase in expression of the whole X (and less expression of the X’), rather than upregulation of a handful of X-linked genes.

<table>
<thead>
<tr>
<th>Replicate</th>
<th>Genes with X-specific expression</th>
<th>Genes with X’-specific expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female 4-8h replicate 1</td>
<td>328 (9.62%)</td>
<td>133 (3.90%)</td>
</tr>
<tr>
<td>Female 4-8h replicate 2</td>
<td>411 (12.05%)</td>
<td>100 (2.93%)</td>
</tr>
<tr>
<td>Female 4-8h replicate 3</td>
<td>350 (10.26%)</td>
<td>108 (3.17%)</td>
</tr>
</tbody>
</table>
Supplementary materials for Chapter V. Faster adaptation but slower divergence of X chromosomes under paternal genome elimination

Supplementary figures

**Supplementary Figure 5.1.** Untransformed phylogeny of species analysed. Rooted with the *Dilophus* clade (Bibionidae). Branch labels represent branch length (genetic distance). Bootstrap values are 100 for every node. Plotted with FigTree v1.4.4 (Rambaut, 2010).
Supplementary Figure 5.2. Alpha (α) values on the X versus autosomes with all genes (including those with negative values) shown. X-linked genes have significantly lower α compared to autosomal genes in B. coprophila (Wilcoxon W = 7235209, P < 10^-5). In L. ingenua there is no significant difference (Wilcoxon W = 2599712, P = 0.8717).

BUSCO Assessment Results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Aphidoletes aphidimyza</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bradysia confinis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bradysia coprophila</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bradysia desolata</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bradysia odorphaga</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bradysia pectoralis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contarinia nasturtii</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contarinia rumicis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dilophus febrilis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dilophus femoratus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lycoriella agraria</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lycoriella ingenua</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mayetiola destructor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mayetiola hordei</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phyosciara flavipes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sitodiplosis mosellana</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Supplementary Figure 5.3. BUSCO scores for all genomes used in this study.

Supplementary Figure 5.4. Dissection diagram for RNA extractions. Female (A) and male (B) reproductive tracts are shown. (1) Hindgut; (2) accessory gland; (3) spermatheca; (4) ovipositor; (5) ovaries containing eggs; (6) testes; (7) ejaculatory sac; (8) genital clasper. The somatic reproductive tissues (2, 3, 4, 7), germline reproductive tissues (5, 6), and carcass (1, 8 + remainder or carcass) were separated and sequenced separately.
Supplementary Figure 5.5. RNAseq data quality control. M = male; F = female; G = gonads; R = somatic reproductive tissues; B = body. Principle component analyses (PCA) for all RNAseq libraries used in this study for *B. coprophila* (A) and *L. ingenua* (B) are shown. Transcript per million (TPM) distributions for *B. coprophila* (C-E) and *L. ingenua* (F-H) are also shown. Distributions and variances are similar within a tissue and species, however there are some differences between the two species that are potentially noteworthy. In the *B. coprophila* PCA, the gonads cluster with the reproductive tract whereas in the *L. ingenua* PCA the gonads cluster with the body. TPM distributions for the gonads and reproductive tissues are also slightly different between the species. In terms of sampling methodology, the only notable difference was that *L. ingenua* females are digenic and virgins are
therefore difficult to collect. As a result, some individuals had already mated and some females had already laid eggs at the time of dissection.
Supplementary tables

**Supplementary Table 5.1.** Non-synonymous \((dN)\), synonymous \((dS)\), and scaled divergence \((dN/dS)\), calculated across all sites for each species pair studied.

<table>
<thead>
<tr>
<th>Species pair</th>
<th>Partition</th>
<th>(dN)</th>
<th>(dS)</th>
<th>(dN/dS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(P. flavipes/)</td>
<td>Autosomes</td>
<td>0.01407534</td>
<td>0.1455235</td>
<td>0.09672209</td>
</tr>
<tr>
<td>(B. pectoralis)</td>
<td>X</td>
<td>0.01064850</td>
<td>0.1371652</td>
<td>0.07763265</td>
</tr>
<tr>
<td>(B. coprophila/)</td>
<td>Autosomes</td>
<td>0.01711715</td>
<td>0.1224654</td>
<td>0.1397713</td>
</tr>
<tr>
<td>(B. odoriphaga)</td>
<td>X</td>
<td>0.01309876</td>
<td>0.1050874</td>
<td>0.1246464</td>
</tr>
<tr>
<td>(B. confinis/)</td>
<td>Autosomes</td>
<td>0.01445257</td>
<td>0.02967786</td>
<td>0.4869814</td>
</tr>
<tr>
<td>(B. desolata)</td>
<td>X</td>
<td>0.01305075</td>
<td>0.03759486</td>
<td>0.3471419</td>
</tr>
<tr>
<td>(L. ingenua/)</td>
<td>Autosomes</td>
<td>0.01264615</td>
<td>0.03784004</td>
<td>0.3342001</td>
</tr>
<tr>
<td>(L. agraria)</td>
<td>X</td>
<td>0.01224157</td>
<td>0.0362275</td>
<td>0.3379069</td>
</tr>
<tr>
<td>(C. nasturtii/)</td>
<td>Autosomes</td>
<td>0.01099144</td>
<td>0.2056144</td>
<td>0.05345657</td>
</tr>
<tr>
<td>(C. rumicis)</td>
<td>X</td>
<td>0.00965438</td>
<td>0.1860239</td>
<td>0.05189860</td>
</tr>
<tr>
<td>(S. mosellana/)</td>
<td>Autosomes</td>
<td>0.010461580</td>
<td>0.1364135</td>
<td>0.07669024</td>
</tr>
<tr>
<td>(A. aphidimyza)</td>
<td>X</td>
<td>0.009302891</td>
<td>0.1277974</td>
<td>0.07279405</td>
</tr>
<tr>
<td>(M. destructor/)</td>
<td>Autosomes</td>
<td>0.02581267</td>
<td>0.1386329</td>
<td>0.1861944</td>
</tr>
<tr>
<td>(M. hordei)</td>
<td>X</td>
<td>0.02238566</td>
<td>0.1225223</td>
<td>0.1827068</td>
</tr>
<tr>
<td>(D. febrilis/)</td>
<td>Autosomes</td>
<td>0.02486015</td>
<td>0.12172080</td>
<td>0.2042391</td>
</tr>
<tr>
<td>(D. femoratus)</td>
<td>X</td>
<td>0.01399283</td>
<td>0.07519602</td>
<td>0.1860847</td>
</tr>
</tbody>
</table>

**Supplementary Table 5.2.** Statistical test results for comparisons of \(dN/dS\) and \(pN/pS\) between differentially expressed (DE) X-linked and autosomal genes. Autosomes are significantly higher than X chromosomes if \(P\) values are italicised; X chromosomes are significantly higher than autosomes if \(P\) values are bold.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Species</th>
<th>Tissue</th>
<th>DE gene category</th>
<th>Wilcoxon W</th>
<th>(P) value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(dN/dS)</td>
<td>(B. coprophila)</td>
<td>Body</td>
<td>Strongly male-biased</td>
<td>72878</td>
<td><strong>0.01969</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Male-biased</td>
<td>175408</td>
<td><strong>0.0008208</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Unbiased</td>
<td>64638</td>
<td><strong>0.009477</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Female-biased</td>
<td>9869</td>
<td>0.08963</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Strongly female-biased</td>
<td>50321</td>
<td><strong>0.005037</strong></td>
</tr>
<tr>
<td></td>
<td>Gonads</td>
<td>Strongly male-biased</td>
<td>Male-biased</td>
<td>Unbiased</td>
<td>Female-biased</td>
</tr>
<tr>
<td>-------------------</td>
<td>-------------------------------------------</td>
<td>----------------------</td>
<td>-------------</td>
<td>-----------</td>
<td>---------------</td>
</tr>
<tr>
<td>L. ingenua Body</td>
<td>Strongly male-biased</td>
<td>89978</td>
<td>0.9811</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Male-biased</td>
<td>31926</td>
<td>0.1131</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unbiased</td>
<td>17968</td>
<td>0.07267</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Female-biased</td>
<td>1778.5</td>
<td>0.5443</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Strongly female-biased</td>
<td>1083</td>
<td>0.8151</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pN/pS B. coprophila</td>
<td>Strongly male-biased</td>
<td>53858</td>
<td>0.6701</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Male-biased</td>
<td>124295</td>
<td>0.4865</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unbiased</td>
<td>48418</td>
<td>0.273</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Female-biased</td>
<td>7769.5</td>
<td>0.5175</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Strongly female-biased</td>
<td>37594</td>
<td>0.6986</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L. ingenua Body</td>
<td>Strongly male-biased</td>
<td>69363</td>
<td>0.3124</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Male-biased</td>
<td>21767</td>
<td>0.1572</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unbiased</td>
<td>15039</td>
<td>0.6937</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Female-biased</td>
<td>1306</td>
<td>0.6947</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Strongly female-biased</td>
<td>818</td>
<td>0.7047</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Supplementary Table 5.3. Numbers (and proportions in parantheses) of X-linked versus autosomal genes expressed in each species.

<table>
<thead>
<tr>
<th>Species</th>
<th>Sex</th>
<th>Tissue</th>
<th>X-linked genes with TPM &gt; 0</th>
<th>Autosomal genes with TPM &gt; 0</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>B. coprophila</strong></td>
<td>Male</td>
<td>Body</td>
<td>3524 (91%)</td>
<td>14594 (88%)</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td></td>
<td>3484 (90%)</td>
<td>14382 (87%)</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>Germ line reproductive tissues</td>
<td>3563 (92%)</td>
<td>14707 (89%)</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td></td>
<td>3149 (82%)</td>
<td>12726 (77%)</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>Somatic reproductive tissues</td>
<td>3233 (84%)</td>
<td>12885 (78%)</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td></td>
<td>3369 (87%)</td>
<td>13843 (84%)</td>
</tr>
<tr>
<td><strong>L. ingenua</strong></td>
<td>Male</td>
<td>Body</td>
<td>3458 (85%)</td>
<td>14520 (92%)</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td></td>
<td>3466 (86%)</td>
<td>14445 (92%)</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>Germ line reproductive tissues</td>
<td>2266 (56%)</td>
<td>9678 (61%)</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td></td>
<td>3068 (76%)</td>
<td>12696 (80%)</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>Somatic reproductive tissues</td>
<td>2801 (69%)</td>
<td>11925 (76%)</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td></td>
<td>3522 (87%)</td>
<td>14339 (91%)</td>
</tr>
</tbody>
</table>

### Supplementary Table 5.4. Accessions for all publicly available genomes and data used in this study.

Note that some genomes and WGS libraries used to not yet have an associated publication.

<table>
<thead>
<tr>
<th>Species</th>
<th>Data component/description</th>
<th>NCBI Accession / Bioprioject</th>
<th>Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aphidoletes aphidimyza</td>
<td>Genome</td>
<td>PRJNA634691</td>
<td>No citation available</td>
</tr>
<tr>
<td>Bradysia coprophila</td>
<td></td>
<td>PRJNA953429, PRJNA291918</td>
<td>(Baird et al., 2023; Urban et al., 2022)</td>
</tr>
<tr>
<td>Bradysia odoriphaga</td>
<td></td>
<td>PRJNA612767</td>
<td>(Li et al., 2022)</td>
</tr>
<tr>
<td>Species</td>
<td>Project Accession Number</td>
<td>Citation</td>
<td></td>
</tr>
<tr>
<td>------------------------------</td>
<td>--------------------------</td>
<td>---------------------------</td>
<td></td>
</tr>
<tr>
<td><em>Contarinia nasturtii</em></td>
<td>PRJNA565389</td>
<td>(Mori et al., 2021)</td>
<td></td>
</tr>
<tr>
<td><em>Dilophus febrilis</em></td>
<td>PRJEB64321</td>
<td>No citation available</td>
<td></td>
</tr>
<tr>
<td><em>Mayetiola destructor</em></td>
<td>PRJNA45867</td>
<td>(Zhao et al., 2015)</td>
<td></td>
</tr>
<tr>
<td><em>Phytosciara flavipes</em></td>
<td>PRJNA369255</td>
<td>(Anderson et al., 2022)</td>
<td></td>
</tr>
<tr>
<td><em>Sitodiplosis mosellana</em></td>
<td>PRJNA720212</td>
<td>(Gong et al., 2022)</td>
<td></td>
</tr>
<tr>
<td><em>Aphidoletes aphidimya</em></td>
<td>SRR14087947</td>
<td>(Xian et al., 2021)</td>
<td></td>
</tr>
<tr>
<td><em>Bradyisia odoriphaga</em></td>
<td>SRR13756224</td>
<td>No citation available</td>
<td></td>
</tr>
<tr>
<td><em>Contarinia nasturtii</em></td>
<td>SRR11470102</td>
<td>(Mori et al., 2021)</td>
<td></td>
</tr>
<tr>
<td><em>Contarinia rumicis</em></td>
<td>SRR10116830</td>
<td>(Margaryan et al., 2021)</td>
<td></td>
</tr>
<tr>
<td><em>Dilophus febrilis</em></td>
<td>ERR12148423</td>
<td>No citation available</td>
<td></td>
</tr>
<tr>
<td><em>Dilophus femoratus</em></td>
<td>ERR12148424</td>
<td>No citation available</td>
<td></td>
</tr>
<tr>
<td><em>Mayetiola destructor</em></td>
<td>SRR1738190</td>
<td>(Vicoso and Bachtrog, 2015)</td>
<td></td>
</tr>
<tr>
<td><em>Mayetiola hordei</em></td>
<td>SRR136493</td>
<td>No citation available</td>
<td></td>
</tr>
<tr>
<td><em>Sitodiplosis mosellana</em></td>
<td>SRR14251190</td>
<td>(Gong et al., 2022)</td>
<td></td>
</tr>
</tbody>
</table>

**Supplementary Table 5.5.** Coverage for all DNAseq datasets generated in this study. Coverage estimates are based off an expected genome size of 300Mb.
<table>
<thead>
<tr>
<th>Species</th>
<th>Description</th>
<th>Estimated coverage for sequencing library</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bradysia confinis</td>
<td>1 x male sample</td>
<td>300x</td>
</tr>
<tr>
<td>Bradysia coprophila</td>
<td>11 x female resequencing samples</td>
<td>20x</td>
</tr>
<tr>
<td>Bradysia desolata</td>
<td>1 x male sample</td>
<td>300x</td>
</tr>
<tr>
<td>Bradysia pectoralis</td>
<td>1 x male sample</td>
<td>300x</td>
</tr>
<tr>
<td>Lycoriella agraria</td>
<td>1 x female sample</td>
<td>20x</td>
</tr>
<tr>
<td>Lycoriella ingenua</td>
<td>5 x female samples, 1 x male sample</td>
<td>20x females, 300x male</td>
</tr>
</tbody>
</table>

**Supplementary Table 5.6.** Assembly statistics for genomes used in this study. Note that numbers of predicted genes are likely too high for many species due to them having highly fragmented genomes.

<table>
<thead>
<tr>
<th>Species</th>
<th>Genome size (bp)</th>
<th>Number of scaffolds</th>
<th>N50 (bp)</th>
<th>Number of predicted genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aphidoletes aphidimyza</td>
<td>191431736</td>
<td>69</td>
<td>46847481</td>
<td>17655</td>
</tr>
<tr>
<td>Bradysia confinis</td>
<td>338817951</td>
<td>182010</td>
<td>4231</td>
<td>28337</td>
</tr>
<tr>
<td>Bradysia coprophila</td>
<td>296980291</td>
<td>4</td>
<td>71047972</td>
<td>20498</td>
</tr>
<tr>
<td>Bradysia desolata</td>
<td>275897695</td>
<td>277134</td>
<td>1735</td>
<td>37825</td>
</tr>
<tr>
<td>Bradysia odoriphaga</td>
<td>362665999</td>
<td>1698</td>
<td>2683481</td>
<td>16218</td>
</tr>
<tr>
<td>Bradysia pectoralis</td>
<td>255331210</td>
<td>205721</td>
<td>2183</td>
<td>28537</td>
</tr>
<tr>
<td>Contarinia nasturtii</td>
<td>185885008</td>
<td>5545</td>
<td>4652070</td>
<td>26752</td>
</tr>
<tr>
<td>Contarinia rumicis</td>
<td>162540589</td>
<td>82390</td>
<td>3141</td>
<td>26791</td>
</tr>
<tr>
<td>Dilophus febrilis</td>
<td>305577324</td>
<td>107</td>
<td>47804581</td>
<td>17096</td>
</tr>
</tbody>
</table>
**Supplementary Table 5.7.** Comparison between sex-biased assignments using the specificity metric (SPM) and DESeq2. Note that, importantly, there were no conflicts where a gene was assigned as sex-biased in opposite directions (e.g. male-biased by SPM but female-biased by DESeq2).

<table>
<thead>
<tr>
<th>Species/tissue</th>
<th>Consistent between DESeq2 and SPM</th>
<th>Conflicting between DESeq2 and SPM</th>
<th>Conflicting because not significant in DESeq2</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>B. coprophila</em> body</td>
<td>8285</td>
<td>6982</td>
<td>5614</td>
</tr>
<tr>
<td><em>B. coprophila</em> reproductive tissues</td>
<td>8960</td>
<td>5567</td>
<td>4371</td>
</tr>
<tr>
<td><em>L. ingenua</em> body</td>
<td>9943</td>
<td>5703</td>
<td>4353</td>
</tr>
<tr>
<td><em>L. ingenua</em> reproductive tissues</td>
<td>9025</td>
<td>5561</td>
<td>4638</td>
</tr>
</tbody>
</table>
Where possible, we performed sex linkage assignment of genome scaffolds for all 16 species used in this study. For four species: *Aphidoletes aphidimyza*, *Bradysia coprophila*, *Dilophus febrilis*, and *Sitodiplosis mosellana*, we used chromosomal assemblies. Autosomal and X-linked scaffolds have been previously identified for *B. coprophila* and *D. febrilis* and are listed in the NCBI genome legacy pages. The two species *A. aphidimyza* and *S. mosellana* are gall midges, which all have two autosomes and two nonhomologous X chromosomes that are all similar in size (White, 1973). We identified the two X chromosomes through alignment of male (XO) reads. The *A. aphidimyza* male reads were produced for another study (Baird RB, Ross L, in prep), and the *S. mosellana* reads used were from pooled individuals from Gong et al. 2022). The methodology for alignments and calculation of coverage are described in the main text.

For the non-chromosome-level genomes, we also used alignment of male reads where possible. This was the case for *B. confinis*, *B. desolata*, *B. pectoralis*, *L. ingenua*, *M. destructor*, and *P. flavipes*. For the latter two species we used publicly available reads (see Supplementary Table 5.4); male reads for the other four species were generated as part of the present study. For each species, we identified the ‘haploid’ (1n) and ‘diploid’ (2n) peaks from the bimodal coverage distributions and assigned scaffolds with a mean coverage of 1n +/- 1n*025, i.e. 25% either side of each peak (Supplementary Figure 5.7).

For some species, neither chromosome-level genomes nor male reads were available. For *D. femoratus*, we assumed that orthologs identified as autosomal and X-linked in *D. febrilis* would most likely be conserved in their linkage. For the species pair *Contarinia nasturtii* and *C. rumicis*, only female reads were available. We used alignments of the species with the more contiguous genome, *C. nasturtii*, to the closest outgroup, *S. mosellana*, to assign linkage to scaffolds (see main text for methods). Using this method for linkage assignments assumes that there is minimal turnover in sex linked sequences between *Sitodiplosis* and *Contarinia*. That all gall midges that have been studied before have the same karyotype with similarly sized chromosomes suggests that there has not been extensive turnover (Benatti et al., 2010; Gong et al., 2022; White, 1973; also see *A. aphidimyza* and *Obodilposis robiniae* genomes on NCBI under project numbers PRJNA634691 and PRJNA883195, respectively). Furthermore, we performed an alignment between the relatively contiguous *C.
nasturtii assembly and the chromosome-level S. mosellana assembly using MUMmer v4 (Marçais et al., 2018) and Dot (https://github.com/marianattestad/dot) and found that the two genomes are likely highly syntenic, with some level of rearrangement within large scaffolds but very little translocations between them (Supplementary Figure 5.8A). We also performed outgroup alignment-based assignments for B. odoriphaga using B. coprophila. Karyotypes are also consistent amongst fungus gnats (McCarthy, 1945a, 1945b; Metz, 1938), and we also aligned the two genomes to check synteny (Supplementary Figure 5.8B).

The total amount of sequence assigned to the X chromosomes and autosomes of each species is shown in Supplementary Table 5.8.

Supplementary Figure 5.7. Male read coverage distributions for short-read genomes. Scaffolds within the haploid region (shaded red) were assigned as X-linked; those within the diploid region (shaded blue) were assigned as autosomal.
Supplementary Figure 5.8. Alignment between the genomes of Contarinia nasturtii and Sitodiplosis moselllana (A) and B. odoriphaga and B. coprophila (B).

Supplementary Table 5.7. Final sex-linkage assignments for each species

<table>
<thead>
<tr>
<th>Species</th>
<th>X-linked sequence (bp)</th>
<th>Autosomal sequence (bp)</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aphidoletes aphidimyza</td>
<td>92607069</td>
<td>97390422</td>
<td>Male read coverage, chromosome-level</td>
</tr>
<tr>
<td>Bradysia confinis</td>
<td>51572387</td>
<td>130844112</td>
<td>Male read coverage, histogram</td>
</tr>
<tr>
<td>Bradysia coprophila</td>
<td>70507861</td>
<td>226472426</td>
<td>Male read coverage, chromosome-level</td>
</tr>
<tr>
<td>Bradysia desolata</td>
<td>64733756</td>
<td>98206588</td>
<td>Male read coverage, histogram</td>
</tr>
<tr>
<td>Bradysia odoriphaga</td>
<td>78764338</td>
<td>263514572</td>
<td>Alignment to outgroup</td>
</tr>
<tr>
<td>Bradysia pectoralis</td>
<td>51139535</td>
<td>108253696</td>
<td>Male read coverage, histogram</td>
</tr>
<tr>
<td>Contarinia nasturtii</td>
<td>72629392</td>
<td>79820263</td>
<td>Alignment to outgroup</td>
</tr>
<tr>
<td>Dilophus febrilis</td>
<td>30953041</td>
<td>265192474</td>
<td>Male read coverage, chromosome-level</td>
</tr>
<tr>
<td>Lycoriella ingenua</td>
<td>48699441</td>
<td>107774135</td>
<td>Male read coverage, histogram</td>
</tr>
</tbody>
</table>
**Supplementary Text 5.2. Assignments of female genotypes in Bradysia coprophila.**

We generated WGS Illumina libraries from 11 wild-caught female *B. coprophila* individuals for population sequencing. Since *B. coprophila* is monogenic (i.e. females produce single-sex broods), females can be either XX (male-producing) or X’X (female-producing). The X’ chromosome is distinct from the ‘normal’ X in that it carries a large inversion which suppresses recombination between the two homologs (Baird et al., 2023; Crouse, 1979). Since wild-caught females are phenotypically identical, we analysed heterozygosity for each of these libraries and compared them to known XX and X’X females from the lab stock. To this end, we mapped reads to the core genome (containing just the X, Urban et al., 2022, 2021) with BWA-MEM (Li, 2013), and followed the GATK-4 best practices pipeline (DePristo et al., 2011; McKenna et al., 2010). Alignment files were sorted, duplicates marked and removed and read groups assigned using PicardTools (“Picard toolkit,” 2019). Variants were called with GATK-4 and filtered for QD > 2, FS < 60 and MQ > 40. When running HaplotypeCaller and GenotypeGVCFs, we used the options EMIT_ALL_CONFIDENT_SITES and -include-non-variant-sites, respectively, to call genotypes (including invariant sites) at every position. We then used the scripts parseVCF.py and popgenWindows.py (https://github.com/simonhmartin/genomics_general) to compute the density of heterozygous sites within 100Kb windows across the genome. Heterozygosity distributions revealed that 8 of the 11 resequenced individuals had the genotype X’X (Supplementary Figure 5.9). The remaining 3 were XX.
**Supplementary Figure 5.9.** Heterozygosity on the X chromosome in female *B. coprophila* resequenced individuals compared to the lab line (upper two panels). Note that some small regions in the putative XX individuals show heterozygosity. The reason for this is unclear, but the regions may correspond to segregating polymorphic inversions or, they may arise from recombination events between the X and X’ (it is currently unknown whether such events occur).

**References**


genomes of Danish vertebrate species generated for the national DNA reference database, DNAmark. Environ. DNA 3, 472–480.


Rambaut, A., 2010. FigTree v1.3.1.


