UNIVERSITY OF CALIFORNIA, MERCED

Comparative genomics of exceptional genetic phenomena challenge foundational ideas in molecular evolution

A dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Quantitative and Systems Biology

by

Noelle Marie Anderson

2023

Committee in charge:

Professor Chris Amemiya, Chair
Professor Emily Jane McTavish
Professor Scott Roy, Advisor

The Dissertation of Noelle Marie Anderson is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Professor Scott Roy

Professor Emily Jane McTavish

Professor Chris Amemiya (Chair)

University of California, Merced
2023

Dedication

I would like to dedicate this work to my mother Beverly and my partner Robert, the two best scientists I know and whom I have been lucky to learn from and love and could not have done this without.

I would also like to dedicate my work to my ~3.5 billion year old lineage of life and the ecosystems that have allowed me to ponder on it all professionally.

Table of Contents

List of Tables

List of Figures

Acknowledgments

I would like to acknowledge and thank all my communities and supports.

I would like to thank my advisor Dr. Scott Roy for his support, for getting me into bioinformatics through the National Science Foundation Research Experience for Undergraduates program, and for being a model of curiosity, elegant scientific thinking, balanced rigor, and excitement for evolutionary weirdness.

I would also like to acknowledge and thank my thesis committee Chris Amemiya and Emily Jane McTavish, for their involvement in committee meetings and beyond, and general support and guidance in shaping my graduate work. I also thank the QSB program and UC Merced for the support through the Graduate Dean's Dissertation Fellowship.

Chapter 2

I would like to acknowledge and thank my collaborators Kamil Jaron, Christina Hodson, Laura Ross, Jan Ševčík, Brian Couger, Stacy Pirro, and my labmate Brooke Weinstein and advisor Scott Roy for their contributions to this work. Additionally, I would like to thank and acknowledge Scott Fitzgerald, Nikola Burdíková, Jacintha Ellers for their work collecting specimens.

Genome sequencing of all dipteran specimens besides *B. coprophila* was funded by Iridian Genomes. N.A. and B.W. were supported by National Science Foundation Award 1616878 to S.R. Springtail sequencing and K.S.J. were supported by a European Research Council Starting Grant (PGErepro, to L.R.). C.N.H. was supported by the National Sciences and Engineering Research Council of Canada postgraduate scholarship and the Darwin Trust of Edinburgh. Springtail collection and L.R. were supported by a Natural Environment Research Council Independent Research Fellowship (NE/K009516/1) and Dorothy Hodgkin Fellowship (DHF\R1\180120).

Chapter 3

I would like to acknowledge and thank my colleague Steve Sun and advisor Scott Roy for their contributions to this work. I would like to thank Aarti Raj for her work and company on an adjacent unpublished project.

N.A. and S.R. were supported for this work on National Scientific Foundation awards 1616878 and 1751372.

Curriculum Vita

EDUCATION

University of California, Merced *Fall 2015 - Summer 2023*
Doctor of Philosophy, Quantitative and Systems Biology
Master of Science, Quantitative and Systems Biology
GPA: 4.0

Middle Tennessee State University *Fall 2011 - Summer 2015*
Bachelor of Science in Biology
Microbiology Concentration, minor in English Literature
Cumulative GPA: 3.8 Major GPA: 4.0

TEACHING EXPERIENCE

Bioinformatics and Genome Annotation– Faculty Lecturer *Fall 2022, Spring 2023, Fall 2023*
San Francisco State University

(All below at the University of California, Merced)
Introduction to Genetics– Teaching Assistant - Discussion
Instructors of Record: Kamal Dulai, Liza Gómez Daglio *Spring 2021*
Instructors of Record: Aaron Hernday, Mark Sistrom *Fall 2016*

Cell Biology – Teaching Assistant - Discussion *Spring 2020*
Instructors of Record: Stephanie Woo, Aleksandra Stamenov

Evolution – Teaching Assistant - Lab
Instructor of Record: Dan Edwards *Spring 2018*
Instructor of Record: Emily Jane McTavish *Spring 2017*

Introduction to Molecular Biology – Teaching Assistant - Discussion *Fall 2016*
Instructor of Record: Kamal Dulai

Biostatistics – Teaching Assistant - Discussion - Guest lecturer *Spring 2016*
Instructor of Record: Emilia Huerta-Sánchez

Microbiology – Teaching Assistant - Lab *Fall 2015*
Instructor of Record: Marcos García-Ojeda

RESEARCH EXPERIENCE

The Joint Genome Institute
JGI-UC Merced Genomics Internship
Advisor: Guohong Wu *Summer 2019*

- Analyzed structural variation in the model grass species *Brachypodium distachyon*

University of California, Merced

Ph.D. Advisor: Scott Roy *Spring 2016 – Summer 2023*

- Use genomic data to identify sex determination turnover in diverse arthropods and mammals
- Employ phylogenetic methods to identify horizontally transferred gene fusions of genes in the same metabolic pathways across all available fungal genomes

Rotation Advisor: Emilia Huerta-Sánchez *Fall 2015*

- Generated haplotype structure for windows in the 1000 Human Genomes Phase 3 Variation dataset to provide insight into patterns of human evolution (introgression, demography, etc.)

Rotation Advisor: David Ardell *Fall 2015*

- Worked on a linguistic evolution project, based in ideas about the evolution of genetic code robustness, investigating how speaking population heterogeneity contributes to the structuring of language over generations and how languages remain robust to transmission noise

Middle Tennessee State University

Department of Biology

Supervisor: Rebecca Seipelt-Thiemann *Spring 2015*

- Compared patterns of exon conservation within human gene families

Middle Tennessee State University

Supervisor: Sarah Bergemann *Fall 2012 - Fall 2014*

- Completed an independent research project aimed at delineating species boundaries within a cryptic species complex of Entolomataceae fungi
- Assisted teaching Mycology class and performed lab preparation

San Francisco State University

The National Science Foundation's Research Experience for Undergraduates Program

Supervisor: Scott Roy *Summer 2014*

- Analyzed expressed sequence tag data of Ascomycete fungi for an evolutionarily significant motif indicating alternative splicing caused dual subcellular localization of metabolic proteins

PUBLICATIONS AND PRESENTATIONS

Couger, M.B.*, Roy, S.W.*, **Anderson, N.**, Pirro, S., Epps, C., Dizney, L., Linntribell, M., Ruedas L., Campbell, P. Sex chromosome transformation and the origin of a male-specific X chromosome in the creeping vole. *Science*. 2021; 372: 592-600. doi:10.1126/science.abg7019

Anderson, N., Jaron, K.S., Hodson, C.N., Couger, M.B., Ševčík, J., Pirro, S., Ross, L., Roy, S.W. Gene-rich X chromosomes implicate intragenomic conflict in the evolution of bizarre genetic systems. *Proceedings of the National Academy of Sciences.* 2022; 119 (23). doi:10.1073/pnas.2122580119

Anderson, N., "X chromosome biology simplified by a complex sex chromosome system in two dipteran families." *Evolutionary Genomics of Sex.* Arizona State University. November 2016. (Talk)

Anderson, N., Roy, S.W. Better living through HGT: Evidence for metabolic optimization through horizontal transfer of a fused gene. *Evolution 2016 Conference*. Austin, TX. June 2016. (Poster)

Ardell, D, **Anderson, N**, Winter, B. Noise in Phonology Affects Encoding Strategies in Morphology. *11th International Conference on the Evolution of Language.* March 2016. (Accepted abstract)

Largent, D., Kluting, K., **Anderson, N.**, Bergemann, S. Entolomatoid species in the subgenus *Leptonia* clade from New South Wales and northeastern Queensland, Australia. *Mycologia*. 2016; 131:153-176.

Anderson, N. "Alternative splicing determines the subcellular address of conserved metabolic enzyme in diverse fungi." *Summer Research Symposium in the Biological Research in Ecological and Evolutionary Developmental Biology Program* (as a part of the NSF REU Program). San Francisco State University. August 2014. (Talk)

Abstract

Non-paradigmatic genetic processes and systems in fungi and arthropods illuminate and challenge foundational ideas in molecular evolution

Noelle Marie Anderson
Doctor of Philosophy in Quantitative and Systems Biology
University of California, Merced
2023

Exceptions to paradigms are important in at least two ways, including highlighting limits of existing paradigms and in providing novel opportunities to test general hypotheses. In the context of genetic processes and systems, non-paradigmatic processes such as complex patterns of inheritance can reveal what may be overlooked in conventional systems and even inform us about established general ideas in evolutionary and population genetics. This thesis explores the potential of non-paradigmatic genetic systems as valuable tools for understanding evolutionary processes. In Chapters 1 and 2, I focus on an atypical genetic system in arthropods with genetic peculiarities that allow for less biased natural experiments for testing hypotheses regarding several fundamental questions regarding selection, mutation, and drift, and about sex determination system turnover. To highlight this system's use for several fundamental evolutionary questions, chapter 1 presents the power of X chromosomes for evolutionary analysis, the limitations of this approach, and how the atypical system resolves these limitations. I first present both the value in comparing sex chromosomes to autosomes, and the challenges in interpreting data from these studies due to the co-occurrence of confounding differences between X chromosomes and autosomes— primarily the hemizygosity of the X chromosome in XY males, and the unequal transmission of X chromosomes through the sexes. I describe ways in which this confounded problem has been previously addressed, then introduce an atypical arthropod genetic system and explain how the system allows for the isolation of hemizygosity as a driver of X chromosome and autosomal difference. Finally, I discuss the implications of potential findings in evolutionary rate, sequence diversity, and codon bias. In Chapter 2, I employ the same exceptional genetic system in arthropods in comparison to relatives with typical genetic systems to test a very different question, namely the origins of haplodiploidy. In this chapter, I specifically investigate the origins of male-haploid systems and the relationship of gene-rich X chromosomes with sex-determination system turnover and find evidence to support intragenomic conflict as a driver of turnover. Intragenomic conflict is generally understudied and its effects on evolutionary processes, potentially substantial, remain largely untested. Thus, I offer some of the first empirical testing of its contribution to major genetic and evolutionary shifts. In Chapter 3, I study the history of a gene fusion of two enzymes involved in subsequent steps of fungal leucine biosynthesis, an essential metabolic process in fungi, through phylogenetic reconstruction on the fusion domains. I find a very

complex history, featuring ancestral fusion, multiple instances of gene fission and loss, and abundant gain of the fused construct through horizontal gene transfer. This result is somewhat surprising given the core nature of these enzymes and longstanding paradigms of eukaryotic metabolic and genic evolution. Altogether, this doctoral body of work aims to demonstrate the value in studying exceptional genetic systems and evolutionary processes and how they are revealing for a broad set of important ideas in molecular evolution and genetics.

Chapter 1: A genetic system without X-Autosome transmissional differences removes confounding contributing force to resolve diverse biological questions

1.2 Abstract

Sex chromosomes evolve under different conditions than autosomes. Therefore, comparison of sex chromosomes with autosomes can reveal key insights into various foundational evolutionary processes. Many influential conclusions about mutational dominance and forces affecting selection have been gleaned from such comparisons. However, these studies are confounded by the presence of multiple potential causal forces for certain differential patterns, due to the multiple important ways that X chromosomes differ from autosomes. These may be broken down into two, namely hemizygosity of the X chromosome in males and differences in how X chromosomes and autosomes are transmitted through the sexes. The co-occurrence of these factors has complicated the interpretation of observed molecular differences, leading to very different evolutionary conclusions. The aim of this chapter is to present and explicate a strategy for overcoming this essential confound through focus on an atypical genetic system. I first discuss the importance of X chromosome and autosome comparisons in evolutionary research and the ways in which various potential underlying contributing factors are obscured by the dual presence of X hemizygosity and transmissional differences. Then, I discuss previous attempts to address this difficulty in study. Finally, I propose a novel approach to disentangle these factors and thus also the conclusions we draw from them using an atypical genetic transmission system found in three hexapod families. The proposed model offers a promising framework to resolve these complexities and gain a better understanding of the fundamental questions in evolutionary biology and population genetics.

1.3 X-Autosome differences, why they matter for key biological questions, and why identifying causal forces is challenging

The two sex chromosomes in species with differentiated sex chromosomes (X/Y or Z/W) tend to exhibit strikingly different genic contents, reflecting very different evolutionary conditions since their evolution from an ancestral autosome pair. While Y and W chromosomes typically have few and functionally-biased genes, reflecting massive gene loss and movement, X and Z chromosomes largely retain the ancestral gene complement. Given that autosomes typically contain quasi-random collections of genes, this implies that X and Z chromosomes, at least

initially, also largely contain a quasi-random collection of genes. Therefore, differences in gene content and sequence patterns between X chromosomes and autosomes are largely the result of the evolutionary conditions Xs now find themselves in. Because they evolved from autosomes and now are under different conditions within the genome, sex chromosomes are an excellent tool for studying evolutionary biology and have been used to make some of the most valuable inferences in population genetics and molecular evolution.

A wide variety of studies have exploited this natural experiment to test an equally wide set of evolutionary hypotheses. Indeed, many important conclusions about evolution have been drawn from the comparison of sex chromosomes and autosomes— that recessive beneficial mutations play a significant role in adaptation, that effective population size is an important determinant in genetic diversity, that sexual antagonism is a contributor to chromosomal gene content and to recombination suppression, that deleterious mutations affecting codon bias are largely recessive, and more. These conclusions have been drawn from the different patterns in genomic sequence observed between Xs and autosomes, patterns such as faster adaptive evolution of the X, greater female- and male-biased expression on the X, more gene traffic on and off the X chromosome, reduced X-linked nucleotide diversity, and greater codon usage bias on the X (Schaffner 2004; Vicoso and Charlesworth 2006; Ellegren 2011; Abbott et al. 2017).

However, in many cases, there is reason to be concerned that the observed X chromosomal patterns that have been taken as evidence for various evolutionary conclusions could actually be due to different forces— that is, that the experiments suffer from a confoundment. While the differences between X chromosomes and autosomes are exactly what makes them such useful tools in comparative genomics, that there are *multiple* differences can complicate their use. Specifically, this limitation derives from the coexistence of two key differences between X chromosomes and autosomes— (i) the hemizygosity of the X in males and (ii) the differences in the ways that Xs and autosomes are transmitted through males and females. Importantly, the various experiments testing fundamental hypotheses have emphasized one of these two differences; however, the co-occurrence of the two differences are inherently convoluted. Put another way: while each observed molecular difference between X chromosomes and autosomes is typically assumed to be due to one of these differences (transmission or hemizygosity), it has been impossible in most cases to fully exclude the possibility that the molecular difference is instead due to the other. Under both hemizygosity and transmissional differences, the relative contributions of parameters such as sexual antagonism, effective population size (N_e), mutation rate, recombination rate, and zygosity to the observed differences between autosomes and sex chromosomes remain difficult to disentangle (Vicoso and Charlesworth 2006). Thus, experiments are needed wherein the effects of hemizygosity and transmissional differences can be differentiated.

The central purpose of this work is to propose and explain a new model for disentangling these questions, namely an atypical system of genetic transmission that has evolved at least three times in hexapods (only dipterans are discussed here, other hexapods in Chapter 2). I first explain various distinctive features of X chromosomes that have been observed, highlighting both the potential implications for large biological questions and the ways that confoundedness between transmission and hemizygosity leave these implications as provisional. I then explain previous attempts to unravel transmission and hemizygosity in typical and atypical genetic systems. Finally, I introduce a new atypical system and explain how it allows for an improved experiment for disentangling transmission and hemizygosity to make cleaner conclusions about phenomena central to molecular evolution and population genetics.

1.4 The problem and the stakes: Observed X-A patterns with various evolutionary implications and potential causes

1.4.1 Faster-X either implies many adaptive mutations are recessive or are sexually antagonistic

One major observed difference between X chromosomes and autosomes is in observed faster protein evolution of X chromosomes (Meisel and Connallon 2013). This phenomenon has two major proposed causes. On the one hand, faster-X could be due to differences in transmission if sexually antagonistic mutations are common (Charlesworth et al. 1987; Vicoso and Charlesworth 2009). Faster-X could also be due to differences in effective population size (N_e; because for each mating pair, there are 4 autosomes and 3 X chromosomes, thus $N_eX = 3/4N_eA$, see Box 1) if many mutations are slightly deleterious and accumulate on the X via enhanced drift under lowered N_e (Charlesworth et al. 1987; Vicoso and Charlesworth 2009). On the other hand, faster-X could be due to X hemizygosity through increased efficiency of selection on the haploid X in males if many newly arising adaptive mutations are recessive (Haldane 1924; Charlesworth et al. 1987; see Box 1). The results on faster-X (or its ZW parallel, faster-Z) in a diversity of species, from dozens of empirical studies over the past 20 years, have failed to paint a cohesive picture of what forces are predominantly responsible for increased evolutionary rate. Unraveling this difference between hemizygosity and transmissional and N_e differences is important because these varied possibilities have different implications for fundamental questions in evolution, particularly regarding the nature of new mutations and their dominance, and general and sex-specific fitness effects. I now explain these dynamics and implications.

The degree of dominance/recessivity of new mutations, particularly beneficial mutations, is of central importance for many key evolutionary/genetic questions including the evolution of sexual reproduction, recombination, the evolution of selfing, evolution within small populations, hybrid inviability, inbreeding

depression, and the maintenance of genetic variability (Connallon and Hall 2018). While experimental evolution studies have made headway in estimating the distribution of mutant fitness effects and dominance, these studies mostly allow for the study of newly-arising deleterious mutations, and are limited in their ability to assess new beneficial mutations(Orr 2010). Instead, much modern work on the dominance of beneficial mutations has focused on the comparison of the evolutionary rate between X chromosomes and autosomes, drawing upon long-established theory.

Box One: Charlesworth's Faster-X Theory

While Haldane (1924) was first to point out that selection would be inefficient on rare recessive mutations except when sex-linked, when selection would be more efficient in the hemizygous sex, Charlesworth et al. (1987) formalized this theory of what is now known as "Faster-X". Faster-X is faster adaptive evolution on the X chromosome relative to autosomes and in this initial formulation is dependent on new beneficial mutations being at least partially recessive ($h < {}^1/_2$).

The rate of evolution (R) is equal to the population level mutation rate (M) multiplied by the probability that a mutation fixes (p). M is equal to twice the generational mutation rate (μ) and from Haldane, the probability that a beneficial mutation fixes is roughly twice the selection coefficient (s) (1927). When comparing R between Xs and autosomes, the differences in effective population size (N_e) must be considered, since assuming an equal sex ratio, there are 3 copies of the X for each 4 autosome copies. There will also be differences in sex-averaged selection coefficients, since males transmit the X only ${}^1/_3$ of the time, and females ${}^2/_3$. Below, I show the **expected evolutionary rates between X chromosomes (R_X) and autosomes (R_A).** A and X subscripts indicate chromosome, s and f indicate sex.

Autosomes	X Chromosomes
$M_A = 2N_{eA}\mu$	$M_X = 2(\frac{3}{4}N_{eA})\mu = \frac{3}{2}N_{eA}\mu$
$s_A = \frac{s_m h_m + s_f h_f}{2}$	$s_X = \frac{s_m + 2s_f h_f}{3}$
$\boldsymbol{R_A} = 2N_{eA}\mu 2(\frac{s_m h_m + s_f h_f}{2})$	$\boldsymbol{R_X} = N_{eA}\mu(s_m + 2s_f h_f)$

When selection is equal in males and females ($s_m = s_f$),

$\boldsymbol{R_A} = 4N_{eA}\mu sh$ and $\boldsymbol{R_X} = N_{eA}\mu s(1+2h)$, so that $\frac{\boldsymbol{R_X}}{\boldsymbol{R_A}} = \frac{1+2h}{4h}$

Faster-X occurs when $\frac{\boldsymbol{R_X}}{\boldsymbol{R_A}} > 1$, thus, faster-X is expected to be seen when $h < {}^1/_2$, when mutations are at least partially recessive.

Theoretical considerations indicate that X chromosomes are expected to evolve faster than autosomes when many new beneficial mutations are recessive and are

thus better exposed to selection in the hemizygous male (Haldane 1924; Charlesworth et al. 1987; see Box 1). While the rate of substitution on the X is most often discussed in terms of the effect of greater positive selection, more efficient selection in hemizygous males could also drive more removal of deleterious recessive mutations, which would be expected to lead to a slower-X effect (Charlesworth et al. 1987; Vicoso and Charlesworth 2009). Notably this theory only applies to newly arising mutations, as adaptation on standing variation is expected to drive a slower-X effect as well (Charlesworth et al. 1987; Orr and Betancourt 2001).

When faster-X evolution has been observed, it has been largely attributed to positive selection (Baines et al. 2008; Hvilsom et al. 2012; Langley et al. 2012; Meisel and Connallon 2013; Kousathanas et al. 2014; Sackton et al. 2014; Dean et al. 2015). A lack of faster-X is sometimes attributed to increased purifying selection, rather than positive selection, due to hemizygosity in the heterogametic sex, as seen in some butterflies (Rousselle et al. 2016).

An interesting and informative parallel to X chromosomal evolution is evolution on the Z chromosome. Despite the similarities between faster-X and faster-Z, faster-Z is complicated by a reduction of N_eZ/N_eA below ¾ due to greater variance in male reproductive success and the reversed heterogamety in the sexes. This reduction in N_e would be expected to cause fast-Z largely through relaxed selection allowing greater fixation of deleterious mutations and has been seen in some ZW species (Mank et al. 2010; Wright et al. 2015; Hayes et al. 2020). However, faster-Z putatively due to positive selection has also been observed in silkmoths, turtles, and some birds (Sackton et al. 2014; Dean et al. 2015; Radhakrishnan and Valenzuela 2017). One possible explanation is that the Z is more frequently transmitted through spermatogenesis, which is often more mutagenic than is oogenesis due to more mitotic divisions, potentially providing more variation for selection to act on (Li et al. 2002; Kirkpatrick and Hall 2004; Ellegren 2007).

Another issue relevant to faster-X and other X-A patterns is differences in recombination rate on autosomes compared to X chromosomes. Differences in recombination rate could play a role in the efficiency of selection on the X and thus contribute to or impede faster-X. However, the case is not so simple. For instance, in *Drosophila*, Xs have a higher recombination rate than autosomes because of the lack of any recombination in males and thus would be expected to be less subject to consequences of linked selection. The fact that faster-X has been observed extensively in *Drosophila* species (Meisel and Connallon 2013; Charlesworth et al. 2018), as well as in organisms where autosomes have a higher recombination rate than Xs, makes it unclear if recombination rate plays a major role in evolutionary rate differences.

In short, the conflicting results of studies on the evolutionary rate of the X/Z chromosome are difficult to interpret and there is currently little definitive

consensus about the distribution of dominance and fitness effects of mutations and how they contribute to the evolution of the X chromosome. Understanding the faster-X phenomenon is essential to understanding adaptive evolution and the effects of hemizygous selection.

1.4.2 Reduced nucleotide diversity may have implications for purifying selection, effective population size, or male-driven mutation

Reduced nucleotide diversity on the X chromosome compared to autosomes has been observed in several mammals, birds (Z), plants, nematodes, and some *Drosophila* species (Sundström et al. 2004; Mank et al. 2007; Ellegren 2009; Leffler et al. 2012; Wilson Sayres 2018). This could be due to hemizygosity of the X/Z chromosome, due to stronger purifying selection on the haploid X in hemizygous males (Ellegren 2009). Alternatively, this may be due to the differential transmission of Xs and autosomes through the sexes. As discussed above in the context of faster-Z evolution, sex chromosomes and autosomes may experience different rates of mutation based on sex-biased transmission. In XY systems, the dynamics are reversed and X chromosomes are transmitted through females twice as often as in males, potentially reducing the mutational load of the X chromosome due to fewer rounds of mitotic division in oogenesis compared to spermatogenesis. Thus, from male-biased mutation, we may expect reduced mutation on the X (Ellegren 2007). Lower X chromosomal sequence diversity could also be attributed to reduced N_e of the X relative to autosomes in most systems, where increased drift may cause loss of diversity on the X. On the other hand, sexual antagonism on the X chromosome is expected to increase diversity via balancing selection (Mank 2017).

Depending on the primary cause of the reduced nucleotide diversity on X chromosomes, we may conclude different things about sequence evolution. If Xs undergo more purifying selection due to hemizygosity, it lends evidence to the idea that many beneficial mutations are recessive. If the X experiences less diversity from relaxed selection due to drift, we may conclude that N_e plays a large role in selectional efficiency on the X. If the observed patterns of lower X diversity are due to lower mutation rates, this would give support to the notion that oogenesis is indeed less mutagenic than spermatogenesis, at least within that taxon. Thus, we can arrive at several different conclusions highly relevant for key evolutionary questions, and understanding the different true drivers of this pattern is important.

1.4.3 Greater X-linked codon usage bias may be driven by enhanced hemizygous selection or by male-driven mutation

The third pattern observed to be different between X chromosomes and autosomes is increased codon usage bias on X chromosomes. Higher codon usage bias (CUB) on the X/Z chromosome has been seen in multiple *Drosophila* species, *C. elegans*, chicken, and in some plants, but is not found in silkmoth or in

several mammalian species (Singh et al. 2005; Haddrill et al. 2010; Qiu et al. 2011; Rao et al. 2011; Kessler and Dean 2014; Sackton et al. 2014). The preferential usage of some synonymous codons over others for the same amino acid is thought to arise primarily from selection for translational efficiency and/or accuracy (Reviewed in Hershberg and Petrov 2008). Like other patterns described here, greater codon bias on the X may indeed be due to improved selection on the X in hemizygous males. However, increased codon bias on the X could also be due to a reduction in mutation rate on the X relative to autosomes due to the X's less frequent passage through the more mutagenic spermatogenesis. Alternatively, codon bias has been attributed to differences in N_e and GC-biased gene conversion (Galtier et al. 2018).

Similar to arguments for previous patterns discussed here, elevated X-linked codon bias due to hemizygosity would support the idea that many new beneficial mutations are recessive or that purifying selection is enhanced under hemizygosity. Primarily transmissional causes would support the conclusion of unequal gametogenic mutation rates between sexes in the species of study, other biased mutational processes, or strong effects of N_e. Codon usage bias is a complex and still misunderstood phenomenon, thus understanding what creates these patterns is important for fundamentally understanding sequence evolution.

1.4.4 The evolution of haplodiploidy is allowed by male haploid viability or facilitated by genomic conflict

The confounding presence of both transmissional differences and differences in zygosity also has implications for the evolution of unique genetic systems such as haplodiploidy. Haplodiploidy may be attributed to increased viability for newly haploid males due to more efficient selection in hemizygous males (Bull 1979; Goldstein 1994; Blackmon et al. 2015), or to genetic conflict between Xs and autosomes that destabilize sex ratios, leading to sex determination turnover (Haig 1993a; Werren and Beukeboom 1998; Normark and Ross 2014). To understand the complex forces driving sex determination system turnover and the evolution of atypical genetic systems, it is important to distinguish these potential causes. These hypotheses are the topic of the second chapter of this thesis, where they are more fully developed and tested empirically.

1.5 Previous attempts to disentangle causes of X-A differences

Various groups have taken different approaches to untangle the effects of hemizygosity and transmission and differences in effective population size (N_e) between sex chromosomes and autosomes. Here, I describe three approaches commonly taken to better understand the separate contributions of shaping forces.

1.5.1 Studying genes with sex-specific expression removes potential sexual antagonism

Differences in expression between sexes can be driven by differential selection in males and females due to sexual antagonism, when traits possess different fitness optima in different sexes (Rice 1984). Sex-biased or sex-specific gene expression is often attributed to historical sexual antagonism, where biased or private expression essentially resolves the problem of sexual antagonistic mutations (Connallon and Knowles 2005; Mank 2017; Wright et al. 2018). While female-biased transmission is expected to favor female-benefiting sexually antagonistic mutations and potentially drive demasculinization or feminization of the X, hemizygosity of the X chromosome is expected to favor recessive male-benefitting mutations. Differential N_e, recombination, and mutation rates are also expected to contribute to patterns of sex-bias on the X chromosome. Indeed, many conflicting patterns of sex-biased gene expression on the sex chromosomes have been observed in a diversity of organisms and could be due to a number of evolutionary forces (reviewed in Grath and Parsch 2016). By focusing on genes with very different sex-specific expression patterns where current sexual antagonism is not expected to have a significant effect, the sex-biased transmissional forces are largely nullified, allowing tests of the effects of other potential variables. Thus, many groups attempted to remove the inherent confounds of the X chromosome by studying X-autosomal differences in genes with very different sex-specific expression patterns.

However, empirical results even for genes with sex-biased expression do not reveal a clear story. In Lepidopterans, both faster-Z for the expected female-biased genes and a lack of this phenomenon have been observed, as well as evidence for relaxed selection in male-biased Z-linked genes, indicating signatures of both adaptive evolution through hemizygosity and of drift due to differences in Ne, and in some cases a surprising faster-Z result in male biased genes has been found (Sackton et al. 2014; Rousselle et al. 2016; Pinharanda et al. 2019; Mongue et al. 2022). Similarly in birds, both presence and absence of faster-Z for female-biased genes have been observed (Mank et al. 2010; Dean et al. 2015). In *Drosophila*, strong faster-X in male-biased genes as expected through hemizygous adaptive evolution has been confirmed through many studies, though faster-X in female-biased genes has also been observed (female-biased findings from Grath and Parsch 2012; Müller et al. 2012; other results reviewed in Meisel and Connallon 2013 and Charlesworth et al. 2018). In mammals, evidence for faster-X in male-biased genes and overall reduced X chromosomal diversity has been shown, both indicating increased selection due to hemizygosity, but is generally less explored (Torgerson and Singh 2005; Baines and Harr 2007; Kousathanas et al. 2014). Given that sex-biased genes, while biased and less likely to be subject to sexual antagonism, are still typically expressed in both sexes at some level through varying lineage-specific levels of dosage compensation, and are affected by differing recombination and mutation rates, this approach is informative but not uncomplicated.

1.5.2 Undifferentiated sex chromosomes or regions remove typical X-A differences

Another strategy that has been used to address the confounding variables of hemizygosity and transmissional differences is taking advantage of entire sex chromosomes or regions of sex chromosomes that have not yet differentiated or are in various stages of differentiation. This is most commonly associated with newly-evolved sex chromosomes ("neo-sex chromosomes") but can also occur in long-standing "homomorphic" sex chromosomes (i.e., cases in which sex chromosomes continue to recombine outside of a relatively small section holding the sex-determining gene(s)), when homomorphic sex chromosomes belatedly cease recombination and start to differentiate.

When sex chromosomes or regions of them have not yet ceased recombination, individuals of the heterogametic sex (XY males, ZW females) still have both complements of the sex chromosomal genes that are both transmitted through males and females at equal rates, rendering them essentially identical to autosomal genes and thus subject to autosomal dynamics. By contrast, at times shortly after recombination has ceased, X/Z-linked genes experience characteristic transmissional differences (bias through one sex), but not hemizygosity, since they are coupled to functioning Y/W-linked gene copies (since these genes have not yet had time to degrade).

These considerations give rise to the strategy of comparing X- and Z-linked genes that ceased recombination at different times in the past. Results from studies focusing on comparing newly formed neo-X/Z and ancient sex chromosomes, different sex chromosome strata (a stratum is a group of genes that ceased recombination at the same time), and pseudoautosomal regions of sex chromosomes have supported this strategy, though have lead to findings being attributed to various causes. In *Drosophila*, comparing hemizygous X/Z loci to diploid X/Z loci showed faster adaptive evolution of hemizygous regions, suggesting hemizgosity alone is sufficient to cause faster adaptive evolution (Zhou and Bachtrog 2012). Poeciliid fish (guppies) show a high diversity of differentiated sex chromosomes, and faster-X is found in species with highly differentiated Xs, absent in guppies with homomorphic chromosomes, and in species with intermediately differentiated sex chromosomes, faster-X is seen only in the oldest strata (Darolti et al. 2023). The correlation of faster-X with the degree of sex chromosome degeneration seen here is consistent with the effect of beneficial recessive mutations being better acted on by selection in hemizygous X-linked genes. While still an effective comparison, these studies are complicated due to the difficulty in demarcating evolutionary strata and estimating X-Y/Z-W divergence, along with varied rates of differentiation and dosage compensation across lineages (Charlesworth 2021).

1.5.3 Unique pea aphid genetic system removes transmissional and N_e differences

A third approach for distinguishing confounding forces acting on sex chromosome evolution has been the use of a specific atypical genetic system in pea aphids. Pea aphids are facultatively sexual: they reproduce parthenogenetically to yield XX females across most of their life cycle, except for a single annual generation in which they reproduce sexually with XO males (Davis 2012). For this generation, XX females produce XO males through programmed loss of one X chromosome (Wilson et al. 1997). Conversely, matings between XX females and XO males produce only XX female progeny because of elimination of X-lacking sperm, yielding all X-bearing sperm in males. Because AO sperm are eliminated, males pass on one X in every gamete like females, unlike typical systems where males transmit the X in half of sperm and females in all. In aphids, this leads to equivalent transmission of Xs through both sexes, an autosomal-like inheritance which eliminates A-X differences in transmission and N_e (Jaquiéry et al. 2012). In this system, both equivalent and reduced X-linked diversity has been observed, producing mixed evidence toward increased purifying selection due to hemizygosity (Brisson et al. 2009; Jaquiéry et al. 2012). Faster-X evolution has been observed in pea aphids and was initially attributed to an increase in adaptive evolution on the hemizygous X (Jaquiéry et al. 2012). However, in follow-up work using expression and polymorphism data, Jaquiéry and colleagues conclude that the fast-X effect in aphids is primarily due to relaxed selection on many lowly-expressed X-linked genes (Jaquiéry et al. 2018)They do conclude a minor role for positive selection however, due to the presence of stronger faster-X and low Tajima's D in male-biased X-linked genes.

While the unusual pea aphid inheritance pattern makes for an excellent model system, it is also complicated in its use by the existence of multiple female morphs in pea aphids. Only one annual generation of sexual morphs occur, and the rest of the 10-20 annual reproductive events are through asexual females (Jaquiéry et al. 2012). It was recently discovered that nearly half of the pea aphid genome is expressed in morph-biased ways, and that morph-biased genes have an increased rate of evolution compared to unbiased genes (Purandare et al. 2014). These same authors also showed a significant decrease in codon usage bias with how morph-specific genes are and a higher Tajima's D for rarer-morph-biased genes over asexual biased genes, both potentially indicating more relaxed selection driving accelerated evolution in these genes. Thus, this situation appears to be more complex than was initially hoped. Depending on the distribution of these morph-biased genes across the genome, estimates of X-linked versus autosomal rates of sequence evolution, diversity, and codon usage may be biased, muddying the clarity of expectations and implications drawn from results in the pea aphid system.

1.6 A novel tool for the problem of confounding forces: The genetic systems of Cecidomyiid and Sciarid flies

In this chapter and the next, I discuss how the atypical genetic system of Cecidomyiidae and Sciaridae, two dipteran families of gall midges and fungus gnats, allows for the study of not only another route to understanding how sex determination systems originate, but also how their abnormalities can be used to test fundamental hypotheses in molecular evolution.

Sciaridae and Cecidomyiidae, both families in the large and diverse cosmopolitan superfamily Sciaroidea, were once thought to be sister groups because of the genetic peculiarities described below. However, recent analyses strongly support that Sciaridae is more closely related several other families than to Cecidomyiidae (Ševčík et al. 2016). What makes these flies unique and valuable are two genetic peculiarities in males— somatic paternal X chromosome elimination (PXE) and germline paternal genome elimination (PGE). I will refer to this system in this chapter as "PXE-PGE", and in Chapter 2 as simply "gPGE" to emphasize its contrast to other genetic male haploid systems, explained more in Chapter 2. (Note that neither family has Y chromosomes, however this is not surprising since XX/XO systems are relatively common in dipterans more broadly, unlike in mammals.)

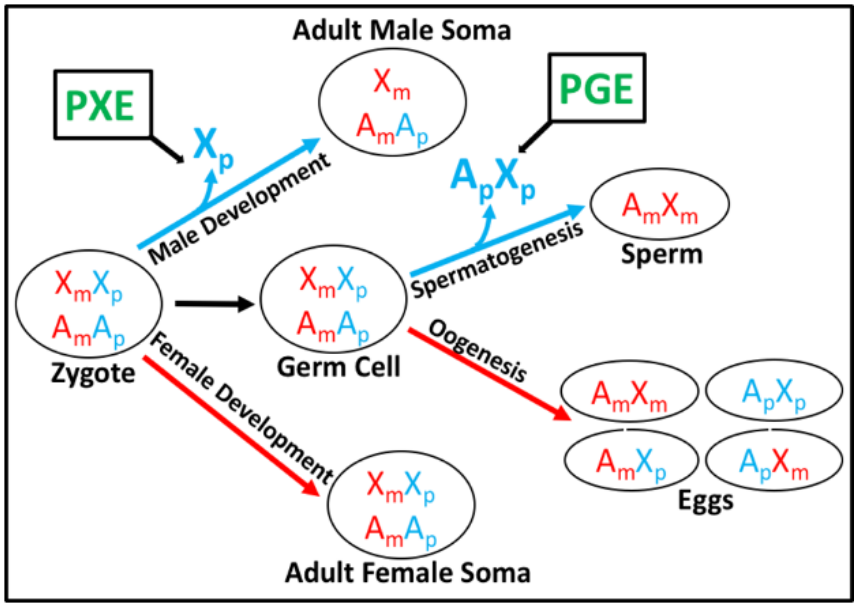


Figure 1-1. Simplified schematic of the Cecidomyid/Sciarid chromosomal cycle, showing paternal X chromosome elimination (PXE) and paternal genome elimination (PGE). "A" indicates autosomes while "X" represents X chromosomes. Red letters with subscript "m" indicate maternally-derived chromosomes, while blue letters with subscript "p" are paternally-derived. A different depiction of this system is shown in Figure 2-1A of Chapter 2 as "Germline-specific PGE" to highlight its contrasts to other genetic systems.

In typical male heterogametic systems, (and as in all *non*-PXE-PGE families in the Sciaroidea superfamily), the X is present in two copies in adult females and each

egg carries a single X, while adult males are hemizygous for the X chromosome and pass it on half as often as females (transmitting the Y in the other half of sperm). Because of this, each sperm has a 50-50 chance of transmitting the maternally-derived X chromosome. However, under PXE-PGE, males begin life as XX zygotes and are determined by elimination of the paternal X chromosome (PXE) early in development to become X0 adults (Fig. 1-1). Next, males undergo elimination of the whole paternal genome (PGE) during meiosis I in spermatogenesis. This PGE event causes all paternally-derived chromosomes to be subsequently excluded from sperm (Fig. 1-1; reviewed in Sánchez 2014)
Because of these two features, all PXE-PGE sperm contain an X chromosome (like eggs) and all X chromosomes and autosomes passed on by the male are maternally-derived (Fig. 1-1). Ironically, this complex system leads to a simplified set of differences between the evolutionary forces acting on Xs and autosomes. Specifically, under PXE-PGE, Xs and autosomes differ in adult hemizygosity as in typical diplodiploid systems, but unlike most systems, X chromosomes and autosomes do not experience transmissional differences, being transmitted through the sexes equivalently (Fig. 1-1).

This equivalency in the transmission of Xs and autosomes has profound effects. Differences in chromosomal N_e, mutation rate, recombination rate, reproductive success between sexes, and demographic effects are *eliminated* as factors that could responsible for any observed difference between Xs and autosomes in these species. Effects observed in PXE-PGE species thus cannot be due to these transmissional differences between the X and autosomes, and thus can be attributed to the hemizygosity of the X, making X-autosome comparisons in these species better controlled natural experiments than those in other species. PXE-PGE is not the only known system offering some of these advantages, as this effect is quite similar to the outcome of the genetic system in aphids, discussed above, but notably is without the complications of alternating modes of inheritance and multiple morphs that may bias results and conclusions (Jaquiéry et al. 2012).

Box Two: Expectations for X-A evolutionary rates in PXE-PGE

In PXE-PGE, females and males transmit a set of autosomes and a set of Xs through every gamete, unlike in normal systems where males transmit an X in only half of sperm. This results in equal Ne between Xs and autosomes, which makes the population mutation rate for the X, M_X, equal to that of autosomes and changes the sex-averaged selection coefficient for the X chromosomes: $s_X = \frac{s_m + s_f h_f}{2}$

These changes make $\boldsymbol{R_X} = 2N_{eA}\mu(s_m + s_f h_f)$ and when $s_m = s_f$, $\boldsymbol{R_X} = 2N_{eA}\mu s(1+h)$.

Thus, the expected X/A rate of evolution under PXE-PGE is $\frac{\boldsymbol{R_X}}{\boldsymbol{R_A}} = \frac{2+2h}{4h}$

In PXE-PGE, unlike in typical systems (Box 1), faster-X is expected when $h < 1$, i.e. for beneficial mutations at any dominance, instead of only for recessive mutations $h < 0.5$. Therefore, if new beneficial mutations significantly contribute to increased X-linked evolutionary rate, I would expect to see a faster-X effect in PXE-PGE.

In contrast to typical systems, drawing conclusions from the PXE-PGE system is much clearer, as hemizygosity of the X in adult males is the only potential driving force of X-A differences when transmission is equivalent. Table 1-1 below shows observed patterns of X-A difference discussed here and their interpretations depending on if the phenomenon is observed or absent in PXE-PGE systems. If we observe faster-X in a PXE-PGE system, we may conclude that hemizygosity plays a large role in the general faster-X phenomenon in enhancing selection on new beneficial mutations and in contrast to non-PXE-PGE systems, beneficial mutations at any dominance level are more likely fix on Xs over autosomes (notably, in one instance slower-X evolution is predicted in PXE-PGE, namely dominant female-beneficial sexually antagonistic mutations; See Box 2). By contrast, absence of faster-X in PXE-PGE would suggest a stronger role for sexually antagonistic mutations, N_e, mutation rate, or recombination in evolutionary rate. Without the unequal transmission through male versus female gametogenesis, male-biased mutation is not a factor in PXE-PGE, nor is sexually antagonistic mutation, thus these cannot contribute to reduced X-linked diversity or increased codon bias. Similarly this is the case with differences in N_e and mutation and recombination rates. Thus, if a reduction of X-linked diversity is observed in PXE-PGE, it would be expected to be due to increased purifying selection on the hemizygous X in males alone. In the absence of reduced diversity on the X chromosome, we may conclude that where this pattern is observed in other lineages it is due to differences in N_e, mutation rates and type, and recombination rate, and the same with absence of increased X-linked codon bias. Presence of increased codon usage bias on the X chromosome in PXE-PGE systems can only be due to increased selectional efficiency on the X, though may be both due to increased adaptive or purifying selection.

Phenomenon	Conclusion if in PXE-PGE	Conclusion if absent in PXE-PGE
Faster-X	More efficient hemizygous selection on new beneficial mutations	Among potentially beneficial newly arising mutations that fix, many are sexually antagonistic. N_e, mutation rate, or recombination rate may also contribute
Reduced X diversity	Hemizygosity increased efficiency of purifying selection	N_e, mutation rate, or recombination rate contribute more to diversity reduction
Increased X-linked Codon bias	Underscores the importance of selective efficiency for codon bias	Underscores the importance of mutation for disrupting codon bias

Table 1-1. Observed differences between X chromosomes and autosomes and their evolutionary implications when present or absent in the PXE-PGE system.

Because of the simplification of observations and causal forces in the PXE-PGE systems, I propose the use of the atypical genetic systems of Cecidomyiid and Sciarid flies for cleaner study of some of the most complex and difficult to study phenomena in molecular evolution and population genetics. In the next chapter, I demonstrate their utility in distinguishing the confounding evolutionary forces that lead to systems like haplodiploidy.

Chapter 2: Gene-rich X chromosomes implicate intragenomic conflict in the evolution of bizarre genetic systems

Noelle Anderson, Kamil S. Jaron, Christina N. Hodson, Matthew B. Couger, Jan Ševčík, Brooke Weinstein, Stacy Pirro, Laura Ross, Scott William Roy

Proceedings of the National Academy of Sciences 2022

2.1 Abstract

Haplodiploidy and paternal genome elimination (HD/PGE) are common in invertebrates, having evolved at least two dozen times, all from male heterogamety (i.e., systems with X chromosomes). However, why X chromosomes are important for the evolution of HD/PGE remains debated. The Haploid Viability Hypothesis posits that X-linked genes promote the evolution of male haploidy by facilitating purging recessive deleterious mutations. The Intragenomic Conflict Hypothesis holds that conflict between genes drives genetic system turnover; under this model, X-linked genes could promote the evolution of male haploidy due to conflicts with autosomes over sex ratios and genetic transmission. We studied lineages where we can distinguish these hypotheses: species with germline PGE that retain an XX/X0 sex determination system (gPGE+X). Because evolving PGE in these cases involves changes in transmission without increases in male hemizygosity, a high degree of X linkage in these systems is predicted by the Intragenomic Conflict Hypothesis but not the Haploid Viability Hypothesis. To quantify the degree of X linkage, we sequenced and compared 7 gPGE+X species' genomes with 11 related species with typical XX/XY or XX/X0 genetic systems, representing three transitions to gPGE. We find highly increased X linkage in both modern and ancestral genomes of gPGE+X species compared to non-gPGE relatives, and recover a significant positive correlation between percent X linkage and the evolution of gPGE. These are among the first empirical results suggesting a role for intragenomic conflict in the evolution of novel genetic systems like HD/PGE.

Keywords: Sex chromosomes, sex determination, haplodiploidy, genome elimination, genomic conflict, insects, Diptera, springtail

2.2 Significance Statement

Sex determination systems such as haplodiploidy, in which males' gene transmission is haploid, are surprisingly common, however, the evolutionary paths to these systems are poorly understood. X chromosomes may play a particularly important role, either by increasing survival of males with only maternal genomes, or due to conflicts between X-chromosomal and autosomal genes. We studied X-chromosome gene richness in three arthropod lineages in which males are diploid as adults but only transmit their maternally-inherited haploid genome. We find that species with such atypical systems have far more X chromosomal genes than related diploid species. These results suggest that conflict between genetic elements within the genome drives the evolution of unusual sex determination systems.

2.3 Introduction

Many animal lineages have evolved genetic systems in which females are diploid but males are haploid or effectively haploid, with each male creating genetically identical sperm carrying the single haploid genome originally inherited from his mother (Ashman et al. 2014). Such systems range from haplodiploidy (HD), in which males are produced from unfertilized eggs; to embryonic paternal genome elimination, in which diploid males eliminate their paternal genome early in development; to forms of germline-specific PGE (gPGE), where the paternal genome is present in male diploid cells but excluded during male meiosis (Figure 2-1a).

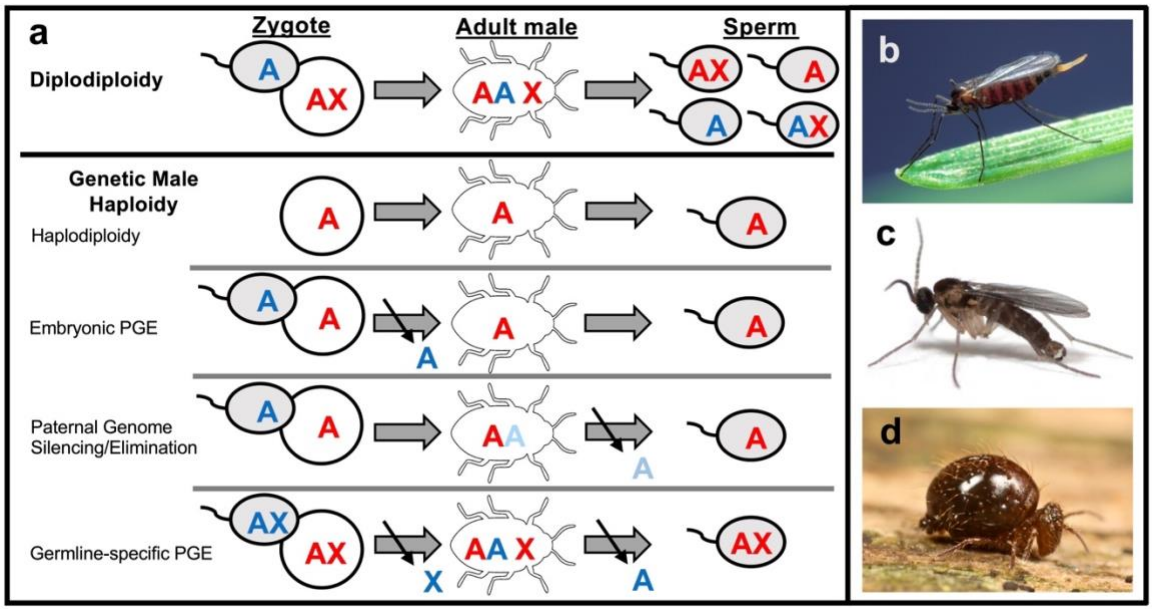


Figure 2-1. Schematic of different genetic systems discussed. a) Male production and spermatogenesis under diplodiploidy, and various forms of male genetic haploidy are shown. Blue and red letters indicate paternal and maternally-derived material respectively. A and X represent autosomes and X chromosomes, respectively. Shown are haplodiploidy (HD), where males develop from unfertilized eggs; embryonic paternal genome elimination, where males eliminate their paternally-inherited genome early in development; paternal genome silencing/elimination, a form of germline PGE where males silence paternal autosomes in somatic cells (indicated by the light blue "A" in males) and eliminate these chromosomes during meiosis; and germline-specific PGE (gPGE), as observed in Sciarids, Cecidomyiids and Symphypleonan springtails, wherein males are produced by somatic loss of the paternal X chromosome(s), and the paternal genome is eliminated in spermatogenesis. b-d) representative species with gPGE: b) the Hessian fly gall midge *Mayetiola destructor* (Cecidiomyiidae), image by Scott Bauer and publicly available via the U.S. Department of Agriculture; c) the fungus gnat *Bradysia coprophila* (= *B. tilicola;* Sciaridae), image by Mike Palmer, used with permission; and d) the springtail *Allacma fusca* (Sminthuridae), image publicly available and taken by Andy Murray. Figure 1-1 in Chapter 1 represents the gPGE genetic system (there called "PXE-PGE") in a different schematic to emphasize the elimination events.

HD/PGE is widespread, seen in ~12% of arthropods and having evolved roughly two dozen times (Ashman et al. 2014). This recurrent evolution perhaps reflects the various advantages of HD/PGE, particularly to mothers, who can increase the transmission of their genes over paternally-inherited genes, control the sex ratio, ensure reproductive success without a mate (in HD), and, under monogamy, reduce conflict between gregarious offspring (Brown 1964; Hamilton 1967; Hartl and Brown 1970; Bull 1979; Normark 2006). Given these general benefits, why does HD/PGE evolve in some lineages and not in others? An important clue comes from the finding that HD/PGE evolves from ancestral male heterogamety (XX/XY or XX/X0) (Gardner and Ross 2014; Blackmon et al. 2015). The most influential hypothesis for this association is the Haploid Viability hypothesis. This hypothesis emphasizes that, starting from an ancestral standard diploid system, newly-evolved haploid males are expected to have markedly lowered fitness due to uncovered recessive deleterious mutations. However, because hemizygosity of X-linked genes facilitates purging of recessive deleterious mutations, an ancestral increase in the proportion of genes on the X chromosome is expected to lead to a decrease in the total number of segregating recessive deleterious mutations, reducing the fitness burden of deleterious mutations for newly-evolved haploid males (Bull 1979; Goldstein 1994; Normark 2004; Blackmon et al. 2015).

However other hypotheses are possible. In particular, the Intragenomic Conflict hypothesis, instead, is a more general hypothesis that sees conflicts between genes within individuals as forces that can destabilize genetic systems and thus

promote the origins of novel systems including, but not limited to, HD/PGE (Haig 1993b; Werren and Beukeboom 1998; Normark and Ross 2014)**.** X chromosomes seem to be more often associated with intragenomic conflict compared to autosomes (Burt and Trivers 2006; Mank et al. 2014; Bachtrog 2020). In particular, X-linked genes can evolve X chromosome drive (>50% transmission of the X in sperm), which can lead to female-biased population sex ratios. Under such sex biases, males have higher average fitness, thus driving selection for new means of producing males (Haig 1993a; Ross et al. 2010). This generally increased male fitness could select for production of haploid males. Moreover, silencing or foregoing the paternal genomic contribution (and in particular the paternal X) could be selectively advantageous insofar as selfish driving X alleles are expected to disproportionately act in males. Although more theoretical work is needed, according to the most developed model, HD/PGE in particular could evolve under the Intragenomic Conflict hypothesis through the exploitation of X chromosome drive by maternal autosomes that increase their transmission by becoming effectively X-linked (Haig 1993b). According to this model, the more genes are X-linked, the more genes will be selected to promote X chromosome drive (and the fewer will be selected to suppress drive), increasing the chance of the evolution of male haploidy.

These two hypotheses differ in whether they predict an association between X linkage and the origins of gPGE in those gPGE systems in which paternal chromosomes are expressed in the soma but are eliminated during male meiosis (Fig. 1a). Given diploid expression of autosomes in the male soma, gPGE, unlike other types of HD/PGE, does not uncover deleterious recessive alleles. Thus, the Haploid Viability hypothesis does not predict an association between X linkage and the evolution of gPGE. However, the notion that X-autosome conflict drives novel systems equally applies to gPGE and other HD/PGE systems, thus the Intragenomic Conflict hypothesis predicts an association between X linkage and the evolution of gPGE. (Notably, in most characterized gPGE systems including those studied here, the paternal genome remains present and expressed through the diploid pre-meiotic stages of spermatogenesis and is only eliminated during meiosis.)

To our knowledge, this differential prediction has not been noted or tested. gPGE systems that retain sex chromosomes and diploid expression of somatic autosomes are known from three lineages: flies in Sciaridae (Metz 1938) and Cecidomyiidae (White 1950) (fungus gnats and gall midges respectively, two families in the diverse dipteran superfamily Sciaroidea) and springtails in the order Symphypleona (Dallai et al. 2000). Sciaridae and Cecidomyiidae represent a substantial fraction of worldwide biodiversity and are some of the most abundant species of flying insects found in tropical rainforests and in temperate ecosystems, with many new species in these groups continuing to be described (Hebert et al. 2016; Ševčík et al. 2022; de Souza Amorim et al. 2022). Sciaridae, Cecidomyiidae,

and Symphypleona have independently evolved similar variants of gPGE, in which males are produced through somatic elimination of paternal X chromosomes, while the remainder of the paternal genome is retained until its elimination during meiosis (Fig. 1-1) (C.W. Metz 1938; Gallun and Hatchett 1969; Stuart and Hatchett 1988a; Stuart and Hatchett 1988b; Dallai et al. 2000; Goday and Rosario Esteban 2001; Burt and Trivers 2006; Gardner and Ross 2014; Jaron et al. 2022). These clades offer a powerful opportunity to disentangle whether the origin of HD/PGE is better explained by the Haploid Viability hypothesis or the Intragenomic Conflict hypothesis.

To test these two hypotheses for the origins of HD/PGE, we performed whole genome sequencing and comparative analysis of 17 genomes from species with gPGE and their non-gPGE relatives. We developed methods to estimate genome-wide X chromosomal linkage using additional 35 dipteran species for validation, and then used these methods to estimate X linkage across the 17 studied species. We find evidence for ancestral gene-rich X chromosomes coincident with three independent origins of gPGE. These results provide the first empirical evidence suggesting a role for intragenomic conflict in the origins of atypical genetic systems.

2.4 Materials and Methods

2.4.1 Specimens and sequencing

In order to compare X chromosomes of gPGE species to their diplodiploid relatives, we collected and sequenced males of 18 species, 14 belonging to the superfamily Sciaroidea spanning nearly all families within, two outgroup species in the dipteran families Anisopodidae and Bibionidae (Sciaroidea and these families are both in the infraorder Bibionomorpha), and two springtail species, *Allacma fusca* and *Orchesella cincta.* Eleven Bibionomorphan specimens were collected and provided by Jan Ševčík, *Catotricha subobsoleta* by Scott Fitzgerald, *Bolitophila hybrida* by Nikola Burdíková, and *Bradysia coprophila* (= *B. tilicola*) was cultured at the University of Edinburgh. Springtails were provided by Jacintha Ellers. Both specimens were flash-frozen and stored at -80°C.

For 15 dipteran species, DNA extractions (Qiagen DNAeasy Blood & Tissue kit), library preparation (Illumina TruSeq kit), and sequencing (Illumina Hi-Seq) were performed by Iridian Genomes. Genomes were assembled using Megahit 1.13 (Li et al. 2016). For the two springtail species and the Sciarid *Bradysia coprophila*, DNA was extracted from male heads using a modified extraction protocol from DNAeasy Blood & Tissue kit (Qiagen, The Netherlands) and Wizard Genomic DNA Purification kit (Promega). TruSeq DNA Nano gel free libraries (350 bp insert) were generated by Edinburgh Genomics (UK) and sequenced on the Illumina HiSeq X (for springtails) or NovaSeq S1 (for *B. coprophila*) generating short reads (150 bp

paired-end). The genome for *B. coprophila* was assembled using Megahit 1.2.9 (Li et al. 2016). The genome of springtail *A. fusca* was assembled using SPAdes v3.13.1 (Bankevich et al. 2012). Both genomes of *B. coprophila* and *A. fusca* assemblies were decontaminated with BlobTools (Laetsch and Blaxter 2017). The assembly of *A. fusca* was annotated using BRAKER 2.1.5 (Hoff et al. 2019). We assessed the quality of all genomes using BUSCO (Simão et al. 2015), to determine the proportion of single copy orthologs expected to be present in either insects (insecta_odb10 for fungus gnat species) or arthropods (for springtails) in the genome assemblies (Fig. 2-S1). *Lestremia cinerea* was excluded from downstream analysis due to irregular genome coverage patterns and a low number of complete BUSCO genes present, indicating likely issues with the genome quality for this species (Fig. 2-S1 and 2-S2). We used publicly available genome assemblies for the Cecidomyiid *Mayetiola destructor* (GCA_000149195.1) and for the springtail *Orchesella cincta* (GCA_001718145.1). For *M. destructor,* we used publicly available male (SRR1738190) and female reads (SRR1738189), and for *O. cincta*, we additionally used available female reads (SRR2222657).

2.4.2 Assigning ancestral linkage groups

The X chromosome in each fly species was identified using two strategies— Muller group linkage and genomic read coverage, similar to strategies implemented in Vicoso and Bachtrog 2015 (Vicoso and Bachtrog 2015). Muller elements are six chromosomal elements first characterized in *Drosophila* that are regarded as being informative about chromosomal linkage (Muller 1940). The *D. melanogaster* proteome (flybase r6.32) (Marygold et al. 2016) was searched against each assessed genome translated into 6 frames using TBLASTN. Top hits for each *D. melanogaster* gene were identified and corresponding genes were classified by the Muller element of their closest *D. melanogaster* ortholog. The X chromosomes in springtails were identified using the coverage approach only.

2.4.3 Identifying X linkage via coverage

Our second strategy implemented DNA coverage levels to characterize autosomal and X-linked sequence, as we expect the single copy X chromosome in males to cause X-linked sequence to be found at half the coverage level of autosomes. Male DNA reads were mapped to their respective genome assemblies and repetitive sequence that could not be singly mapped was accounted for when calculating an adjusted coverage (See 2.6 Supplemental Methods and Figures below). For species in which female read data was available, *M. destructor* and the two springtails, the relative coverage of male to female was used. In the case of *A. fusca,* we used median coverage of two males and 11 females available (Jaron et al. 2022). To classify genes by coverage as either autosomal or X-linked, we used a multi-step protocol relying on the full genome and per-Muller male DNA coverage distributions (See detail in 2.6 Supplemental Methods and Figures). We also assessed 35 other dipteran genomes outside Bibionormorpha using publicly

available data and the same methods of analysis (Fig. 2-S5, Table 2-S1).

2.4.4 Statistical analysis and phylogenetic correction

To test the association between X linkage and the evolution of PGE, we estimated a Bayesian generalized linear mixed "threshold" model (Hadfield and Nakagawa 2010) and a likelihood-based phylogenetic logistic regression described in (Tung Ho and Ané 2014). Both methods attempt to control for the phylogenetic relatedness of the species. For further detail, see 2.6 Supplemental Methods and Figures below.

2.4.5 Testing for ancestral Muller group linkage

To test for evidence of ancestral X linkage, we compared various pairs of species. We studied each Muller element for which both compared species had partial X linkage, in which the ancestral linkage groups have broken up and are now partially X-linked and partially autosomal. Genomes of each species pair were reciprocally blasted to defined putative pairwise orthologs using TBLASTX. Only best reciprocal hits and orthologs that blasted to the same *D. melanogaster* gene were included in further analysis. Each ortholog pair was then assigned based on its inferred X/autosomal linkage for both species (X-linked/X-linked, X-linked/autosomal, autosomal/X-linked, or autosomal/autosomal). Association between X linkage across between-species orthologs was tested by a Chi square test.

2.5 Results and Discussion

2.5.1 Development and testing of an improved method to estimate genome-wide X chromosomal linkage

Illumina genome sequencing and assembly was performed for males of each studied Sciaroidea species, and average read coverage was calculated for each contig. For the dipteran species, putative orthologs of *D. melanogaster* genes were identified via TBLASTN searches of each genome. Each ortholog was then assigned to one of the so-called Muller elements, *D. melanogaster* chromosomal linkage groups that have persisted over long evolutionary time in some fly lineages (Muller 1940; Sved et al. 2016) (though not all (Vicoso and Bachtrog 2015)). For each Muller group in each species, the fraction of X-linked genes was estimated from read coverage distributions using improved methods based on Vicoso and Bachtrog's previous work demonstrating the lability of sex-linked Muller elements across Diptera (Vicoso and Bachtrog 2015) (see 2.6 Supplemental Methods and Figures for full detail). To validate our method, we ran our assignment on 35 dipterans outside of Sciaroidea with a variety of DNA coverage distributions, recapitulating and expanding what was found about Muller element linkage across Diptera by Vicoso and Bachtrog (Fig. 2-S5, Table 2-S1) (Vicoso and Bachtrog 2015). Publicly available chromosome level assemblies for *B. coprophila, A. gambiae, T. dalmanni*, and several Drosophilid species allowed for direct comparison to our assignment and for each we found our estimation of X linkage to be within 1% of previous estimates, allowing us to be confident in our assignment of X-linked and autosomal genes (Fig. 2-S5a).

2.5.2 Increased numbers of X-linked genes in gPGE species relative to related species

To test whether the evolution of gPGE is associated with gene-rich X chromosomes, we estimated the proportion of the genome that is X-linked for 17 species of Sciaroidea flies and two species of springtails. We sampled the flies across seven families spanning the root of Sciaroidea, including two families with gPGE and two outgroup species within Bibionomorpha. We used the publicly available genome assembly and annotation supported by physical mapping for the Hessian fly, Cecidomyiid *Mayetiola destructor* (Richards and Stuart; Aggarwal et al. 2009), and also used available female read data to estimate relative male to female coverage. For the springtails, we performed genomic sequencing of males and females from one species from the gPGE order Symphypleona, *Allacma fusca* (Fig. 2-1d), and of *Orchesella cincta*, from Entomobryidae, the closest relative springtail order with standard XX/X0 sex determination. In Springtails, instead of orthologs, we used genome annotations to estimate the gene density and used both male and female read coverage data. Our assignment methods provided clear estimates of X linkage for nearly all our species, with exceptions in one

species, the Cecidomyiid *Lestremia cinerea*, which showed three distinct peaks in genome coverage rather than two, as well as a low number of complete BUSCO genes present (Fig. 2-S1, 2-S2).

Among all non-gPGE fly species of Bibionomorpha, we found very few X-linked genes, with the X chromosome in all species comprised mostly of genes from the diminutive F Muller element (<1% of all genes), consistent with the previous inference for the ancestral dipteran X chromosome (Fig. 2-2a) (Vicoso and Bachtrog 2015). Interestingly, no Muller elements exhibited clear X-linked peaks in coverage in *Platyura marginata* and *Symmerus nobilis*, the latter of which is sister to all other Sciaroidean species, suggesting either homomorphic sex chromosomes, a lack of an X chromosome, or neo-X chromosomes too recently evolved to be distinguished via coverage as previously observed (Vicoso and Bachtrog 2015).

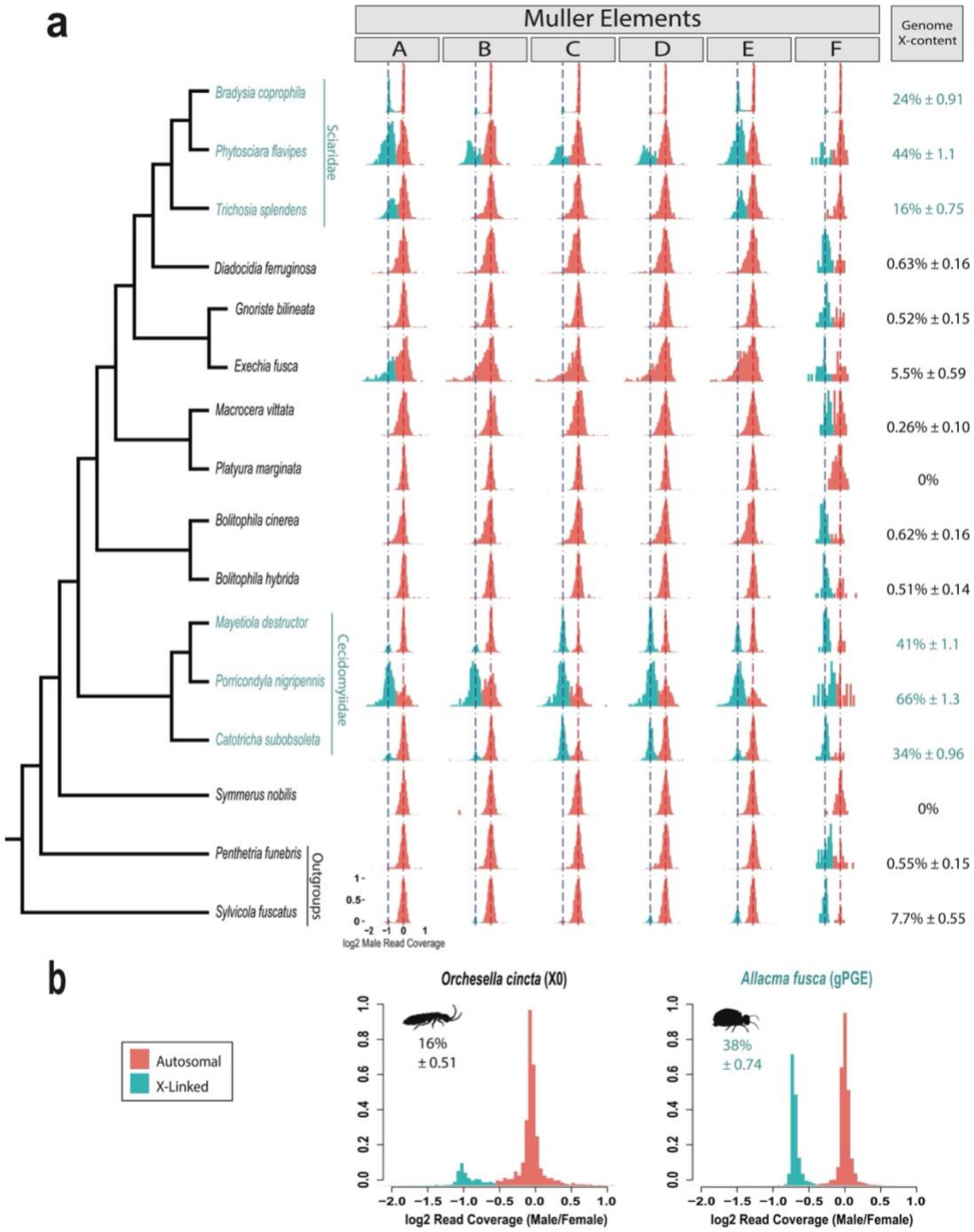


Figure 2-2. Frequency of X-linked and autosomal genes in gPGE species and related diplodiploid species, assessed by DNA read coverage. a) Sciaroidea and outgroups within Bibionormorpha; topology based on Ševčík et al. 2016. Plots for each Muller element show log2 male read coverage normalized by putative median autosomal coverage, with assigned X linkage (blue) and autosomal linkage (red) indicated. The y-axis represents gene frequency scaled to the maximum in each

distribution. Red dashed vertical lines at 0 indicate the expected autosomal coverage peak, blue dashed lines at -1 indicate the expected position of the X-linked peak, at half the coverage of the autosomes. Blue and black species names and genome-wide estimates represent gPGE and diplodiploid species, respectively. Percent estimates represent percent X linkage for each Muller and across each full genome, with error represented by 2SD. As with the two springtails, for the Cecidomyiid *Mayetiola destructor*, female read data was available and thus male/female coverage is shown. In *M. destructor* genes are only included if assignments agree with previous physical mapping placements or were previously unassigned (Aggarwal et al. 2009). b) Whole genome autosomal and X linkage distributions for springtails diplodiploid *Orchesella cincta* and gPGE *Allacma fusca* showing relative male/female coverage.

By contrast, for all six studied gPGE species in both the Sciaridae and Cecidomyiidae clades, genome-wide, we found large fractions of genes to be X-linked, including genes from all six Muller elements (Fig. 2-2a, 2-3). Notably, our results agree with previous results for *M. destructor*, identifying Muller elements C, D, F, and E as partially X-linked (Vicoso and Bachtrog 2015), and our methods additionally detect a small minority of X-linked genes for elements A and B. We also found a clear contrast between the two studied springtail genes: while only 16% of genes in the genome of non-gPGE *Orchesella cincta* are X-linked, for the gPGE springtail *Allacma fusca*, 38% of annotated genes are X-linked (Fig. 2-2b).

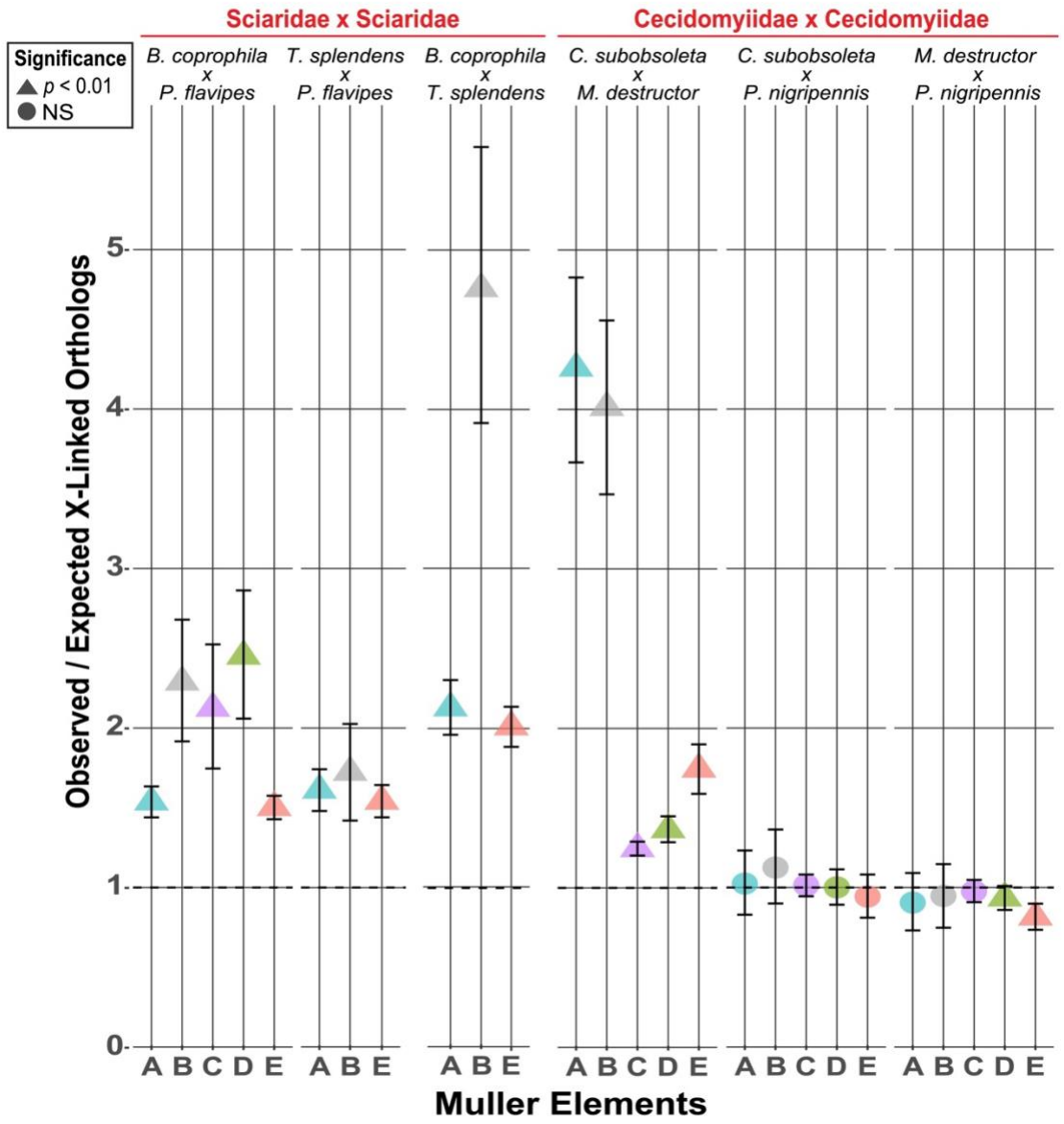


Figure 2-3. Number of ortholog pairs in which both genes are X-linked, compared to the null expectation, for pairs of gPGE species from the same family. Within-family comparisons are shown, between-family comparisons in Fig. 2-S3. Color indicates Muller element. Muller elements for which species do not share X-linked orthologs are excluded, as is the F element. Shapes indicate significance via Chi square. Error bars represent 95% CIs computed from 10,000 bootstrap replicates. Observed/Expected value if no association between X-linked orthologs is 1.

2.5.3 Statistical tests support the relationship between gPGE and X linkage

To test the association between percent X chromosome linkage and the evolution of gPGE, we used multiple statistical methods. While the number of transitions to

gPGE with X chromosomes is small and all current phylogenetic methods with a binary response variable are prone to inflated Type 1 error rates with small sample sizes (Ives and Garland Jr. 2014), we are only aware of those transitions represented here and took several approaches to mitigate this issue. First, we used two different phylogenetically informed methods, a Bayesian probit model and a likelihood model using logistic regression, which gave similar results. In addition, we performed stringent diagnostics of our models, including bootstrapping, checks for Markov chain convergence, and autocorrelation of samples in the posterior distribution.

Our non-phylogenetically informed generalized linear model (glm) inferred a positive and significant effect of the degree of X linkage on the evolution of gPGE, with every one standard deviation increase in percent X linkage increasing the log odds of evolving gPGE by a factor of 3.34. However, occurrences of gPGE group in two tight clusters on the phylogeny, and we detected a strong phylogenetic signal in the amount of X-linkage (Blomberg's K = 0.946; p = 0.012), meaning this correlation could largely result from phylogenetic proximity. Therefore, we used Bayesian estimation with a mixed phylogenetic model and found a robust, positive effect that does not overlap zero (mean = 1.87, 95% 95% CI = 0.366 - 3.73, pMCMC = 0.005; MCMC effective sample size = 304,300). In addition, we performed Ives and Garland's binary phylogenetic logistic regression (Ives and Garland Jr. 2010) with 10,000 bootstrap replicates and found an even stronger standardized effect (mean = 3.37, 95% bootstrap CI = 3.36 - 3.40, p-value = 2.382e-08) which is similar to the estimate with the non-phylogenetic glm. Thus, all three methods support transitions to gPGE being more prevalent in lineages with a higher proportion of X-linked genes (Fig. 2-S3; see 2.6 Supplemental Methods and Figures for full details).

2.5.4 Correspondence between X-linked genes within families indicates ancestrally gene-rich X chromosomes

Although we found an association between gene-rich X chromosomes and gPGE in all three independent origins of this genetic system, the observed association could be explained by either X linkage facilitating the evolution of gPGE or vice versa. Consistent with the former, we see the same patterns of Muller group X linkage within families (E>A>B in Sciaridae species; C>D>E>A>B in Cecidomyiidae). In addition, we found an association between X-linked gene subsets within individual Muller elements, as expected from ancestral linkage. For instance, the subsets of Muller B genes that are X-linked in the Sciaridae species *B. coprophila* and *T. splendens* significantly overlap, and the same is true for all partially X-linked Mullers in both Sciaridae (Fig. 2-3). By contrast, X-linked genes between Sciaridae and Cecidomyiidae do not significantly overlap, supporting independent origins of the large X in these two families (Fig. 2-S4).

Examination of Cecidomyiidae reveals an intriguing pattern. The deeply-diverged species *C. subobsoleta* and *M. destructor* show high correspondence between X-linked gene subsets, indicating substantial ancestral X linkage. However, *P. nigripennis* shows divergent X linkage, with no significant pattern seen in shared X linkage with other Cecidomyiids, and a relative increase in X linkage on Muller elements A and B. This pattern suggests turnover and increases in X linkage in this lineage since the divergence from *M. destructor* (or, less parsimoniously, parallel loss of A/B linkage in the other lineages) (Fig. 2-2a, 2-3).

2.5.5 X Chromosome lability and partial Muller linkage

Our data attest to substantial dynamism of the X chromosome and Muller linkage in both gPGE families within Diptera. This is in contrast to the dominant model of Dipteran sex chromosome evolution where sex linked Muller elements are expected to remain stable over long evolutionary periods. Some notable cases indicate remarkable conservation, such as the X chromosome of the German cockroach which has remained conserved with the ancestral dipteran X chromosome (Muller element F) despite 400 million years of divergence (Meisel et al. 2019). On the other hand, even within *Drosophila* this pattern is disrupted, with fusions of ancestral drosophilid X-linked element A and typically autosomal element D in the obscura clade into the X chromosome, as well as in *D. willistoni* (Fig. 2-S5b) (Schaeffer et al. 2008; Vicoso and Bachtrog 2013). Vicoso and Bachtrog demonstrated abundant sex chromosome turnover across Diptera, broadly challenging the established paradigm of sex linked Muller element stability (Vicoso and Bachtrog 2015).

In addition to demonstrating cases of lost and replaced sex chromosomes, Vicoso and Bachtrog also showed cases of partial linkage, where parts of multiple Muller elements are incorporated into sex chromosomes (Vicoso and Bachtrog 2015). Specifically, they find partial linkage for the B element of *Holcocephala fusca* and for the E element of *M. destructor*, both of which our methods also identified as partially X-linked (35% and 40% of genes, respectively). We additionally find minor partial linkage of Muller elements A and B in *M. destructor* (Fig. 2-2a). In *Anopheles gambiae*, element A is typically discussed as if fully X-linked, however the X chromosome has been previously shown to be only partially composed of element A and parts of other Muller elements, while the rest of ancestrally Muller A genes are now found on autosomes (Keller Valsecchi et al. 2021), consistent with our results (Fig. 2-S5, Supplemental table 2-S1). Additionally, minor partial X linkage of *A. gambiae* elements E (11%) and F (33%) has been previously identified (M. et al. 2002) and is consistent with our findings of 11% and 29% X linkage respectively (Table 2-S1). Our methods demonstrate the resolution to detect low levels of X linkage and suggest partial linkage and general Muller element breakdown may be more common than is generally appreciated.

2.5.6 Concluding remarks

We find that species in the gPGE groups Cecidomyiidae and Sciaridae have, on average, X chromosomes 24 times more gene-rich than non-gPGE Sciaroidea species, with a more than doubling of the X chromosome gene content of the gPGE springtail species compared to the diploid outgroup (Fig. 2-2a and b). Furthermore, we recovered a robust positive correlation between the percent X linkage in the genome and the evolution of gPGE (Fig. 2-S3). While having additional independent origins of X chromosome-containing gPGE would add strength to our conclusions, we are only aware of those studied here.

Notably, while previous similar reports of an association between the extent of X linkage and atypical sex determination are consistent with either the Haploid Viability hypothesis or the Intragenomic Conflict hypothesis (Blackmon et al. 2015), these findings represent the first empirical evidence that suggests Intragenomic Conflict as a strong driver of the evolution of unconventional sex determining systems such as gPGE and haplodiploidy. Given the widespread and repeated evolution of male haploidy, and its association with many unique ecological and life history strategies, our findings point to an important role for intragenomic conflict in shaping biology at all levels from molecule to organism to community.

2.6 Supplemental Methods and Figures

2.6.1 Identifying X linkage via coverage

We used expected DNA coverage levels to distinguish X-linked and autosomal sequence: Because the X chromosome is present in a single copy in males, in males, sequence that is X-linked is expected to be at half coverage compared to autosomal sequence. Trimmed to 50 nucleotides, Male DNA reads for each Bibionomorphan were mapped to their respective genome assemblies using Bowtie with default parameters except for the addition of the -m1 flag to discard reads that mapped to multiple locations in the genome. Because some Bibionomorphan contigs contained large amounts of repetitive sequence that prevented reads from mapping singly, we corrected coverage estimates to only account for singularly mappable positions on the contigs. To do this, we simulated 50nt reads from every mappable position on each contig, mapped them back to the genome from which they were generated using Bowtie, and subtracted the number of reads from each contig that were unable to map singularly from the contig length. This provided us with an adjusted contig length that excluded sequence content that could not be mapped to singularly to use for adjusting coverage estimates; contigs with less than 1000 mappable bases were excluded. Coverage was calculated as: (Read count x read length) / (Contig length - number of multiply mapping reads for that contig + 1). Because male and female DNA

sequence for *M. destructor* is available, the comparison of male to female read coverage was used as with the springtails, in addition to using linkage information previously established by physical mapping to more stringently classify X linkage (Aggarwal et al. 2009)**.**

To assign genes as autosomal or X-linked via coverage, we used a multi-step protocol. First, we used standard methods to (i) identify the highest peak in the full genome coverage distribution and (ii) identify the highest peak in each Muller element distribution that falls closest to the full genome distribution highest peak. The highest peaks were assigned as autosomal centers in all species except *Porricondyla nigripennis* and *Drosophila willistoni,* where the X chromosomal peak was larger. Next, to detect a second coverage peak, we counted genes in coverage bins outward from the highest peak to the left or right, depending on whether the search for the secondary peak was for the X-chromosomal peak, expected to be at half coverage, or autosomal peak, expected to be twice the X-linked coverage. Each bin count was compared via Chi square to an expected bin count calculated as the average between the bin count and the adjacent bin count, such that only significant rises in coverage like a true peak could be identified. Next, we searched for the second highest peak center within bins with the most significant increase and identified those as potentially X-linked (with the exception of the two species listed above where the secondary peaks were autosomal). Distributions with candidate secondary peaks were labeled as bimodal if the secondary peak was at least one tenth of the height of the highest peak, otherwise unimodal. Genes per Muller element distribution were assigned as X-linked or autosomal via k-means clustering, using the Muller-specific X and autosomal peaks as initial cluster centers. The doubled standard deviation of the proportion of X-linked genes in each distribution was estimated as a proxy for the 95% confidence interval.

2.6.2 Statistical analysis and phylogenetic correction

To test the statistical significance of the association between gPGE and the X-linked proportion of the genome within Bibionomorpha, we used Bayesian estimation with a mixed phylogenetic model. The distribution of the predictor variable, the percent of the genome which is assigned as X-linked, includes several zeros and is overdispersed. We therefore transformed it using a Box-Cox power transformation (shifting parameter lambda2 = 0.1; maximum likelihood estimated transformation parameter lambda.hat = -0.1; R package AID v. 2.6 (Asar et al. 2017)), confirming the normality of the data using a Shapiro-Wilk test ($p = 0.0565$). Before proceeding, percent X linkage was also centered at zero and scaled to improve model specification and ease the interpretation of priors and parameters.

Binomial logistic regression using R's generalized linear model function (glm(gPGE~%X-1, family=binomial(link="logit"))) predicts a positive and significant

effect of the degree of X linkage on the evolution of gPGE, with every one increase in the standard deviation of percent X linkage increasing the log odds of evolving gPGE by 3.343. We then used the wald.test function of the aod package v 1.3.1 (Lesnoff and Lancelot 2010) and calculated a chi-squared test statistic of 5.4, with one degree of freedom, and a p-value of 0.02, indicating that the effect of percent X linkage in the genome is statistically significant on the evolution of gPGE. To see how well this non-phylogenetically informed model performs, we tested whether it fits the data better than a null model (with just an intercept) by calculating the difference in deviance (14.586). Next, we performed a Chi-square test with one degree of freedom to obtain an associated p-value of 0.00013, telling us the model with percent X as a predictor of gPGE fits significantly better than a null model. Finally, we calculated various diagnostic statistics with the R package modEvA v. 2.0 (Márcia Barbosa et al. 2013), which all showed a strong correlation between percent X linkage and the evolution of gPGE (D2 = 0.658; R2 Cox-Snell = 0.572; R2 Tjur = 0.756; R2 McFadden = 0.641; R2 Nagelkerke = 0.78).

However, this correlation could result from increased genomic conflict in lineages with greater X linkage, as predicted by the Intragenomic Conflict Hypothesis, or simply from shared phylogenetic history. The latter is especially of concern because occurrences of gPGE occur in two tight clusters on the phylogeny. Therefore, to incorporate phylogenetic contrasts, we first obtained a tree with branch lengths based on Ševčík et al. (Ševčík et al. 2016). Branch lengths for two species missing from the Ševčík tree, *P. marginata* and *B. hybrida* were calculated as the average of their clades and branch lengths for three other absent species, *Symmerus nobilis, Trichosia splendens, Macrocera vittata,* and *Exechia fusca*, were assigned by proxy as branch lengths from congenitor species in the tree. We then used the chronos function in ape v. 5.5 (Paradis and Schliep 2019) to obtain time-calibrated branch lengths using penalized likelihood with a "correlated" evolution model. We did this over absolute time by calibrating the rootage to 160 mya, the divergence time for the family Bibionomorpha in the dipteran Timetree of Life (Bertone and Brian M. Wiegmann 2009). For comparison, we also generated a tree with Grafen's (Grafen 1989) branch lengths (power = 1).

We first calculated the phylogenetic signal of the continuous predictor variable, percent X linkage, using Pagel's lambda and Blomberg's K with the phylosig function in the R package phytools v. 0.7-90 (Revell 2012). Lambda ~ 1 using both the chronogram (1.03; p = 0.010) and the Grafen's branch length tree (0.965; p = 0.009), meaning the evolution of percent X linkage in the genome along the phylogeny roughly corresponds to a Brownian motion expectation. Testing Blomberg's K with our chronogram, percent X linkage in our 16 Bibionomorphan taxa is again as expected under Brownian motion with K = 0.946 (p = 0.012). Using Grafen's branch lengths results in a somewhat lower phylogenetic signal, but still highly significant (K=0.706, p=0.001).

Since percent X linkage and gPGE could be correlated largely due to phylogenetic proximity, we analyzed the evolution of gPGE with a Bayesian generalized linear mixed model using MCMCglmmRAM (Hadfield and Nakagawa 2010), fitting a reduced version of Wright's threshold model (Wright 1934; Felsenstein 2005; Hadfield 2015). The results of this model were generally robust to prior specification but, to improve precision and the efficiency of the MCMC sampling, we used a parameter expanded prior for the variance of the covariance matrix that is strongly skewed toward low variances of phylogenetic signal (prior = list(R = list(V = 1,fix = 1),G = list(G1 = list(V = 1, nu = 1000, alpha.mu = 0, alpha.V = 1)))). We ran the model for 50 million iterations with a burn-in period of 50,000, saving every 100th iteration and using slice sampling to update the latent variables. We inspected MCMC chains visually, and all fixed and random effect variances passed the Heidelberg convergence and half-width diagnostics in the R package coda v. 0.19-4 (Plummer et al. 2005), which tests the null hypothesis that the sampled values come from a stationary distribution. Moreover, autocorrelation was essentially nonexistent for the fixed and random effects variances. The positive correlation between percent X linkage and gPGE described in our results uses our chronogram but is also robust with Grafen's branch lengths (mean = 1.613, 95% CI = 0.2221 - 3.3602, pMCMC = 0.0097; effective sample size = 195,625). To better understand how our model is performing, we used the predict2 function of the postMCMCglmm package in R v. 0.1-2 (Wiley 2013) to calculate the average marginal, out-of-sample predicted probability of evolving PGE using all of the posterior samples. Unfortunately, post-MCMCglmm only supports ordinal models. Fortunately, the threshold model is identical to an ordinal model, except the residual variance now refers to the variance of the link function rather than the variance of the non-identified residuals. Therefore, we multiplied our threshold model's location effects by sqrt(2) and the variance components by 2, making the predictions equivalent.

Finally, for comparison, we used the phyloglm function of the R package "phylolm" (Tung Ho and Ané 2014) to fit the phylogenetic logistic regression described in Ives and Garland (Ives and Garland Jr. 2010) using Firth's penalized likelihood with 10,000 parametric bootstrap replicates. For lower X linkage, the predictions from the three models mostly overlap. However, the Bayesian GLMM threshold model made lower predictions for the evolution of gPGE at high X linkage percentages; in contrast, the phylogenetic logistic regression was in line with the non-phylogenetic GLM (Fig. S3). Nonetheless, all three methods support an association between a higher proportion of X-linked genes and transitions to gPGE.

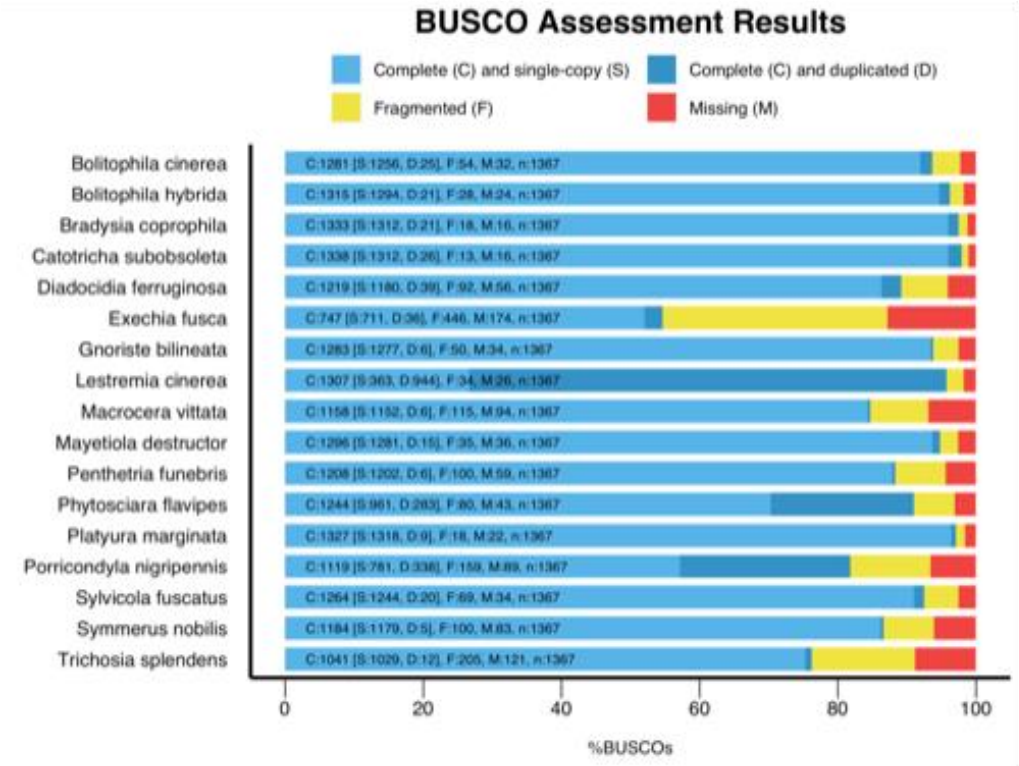


Figure 2-S1. Genome assembly completeness by BUSCO analysis. Counts for each BUSCO category are shown with abbreviations in the bars.

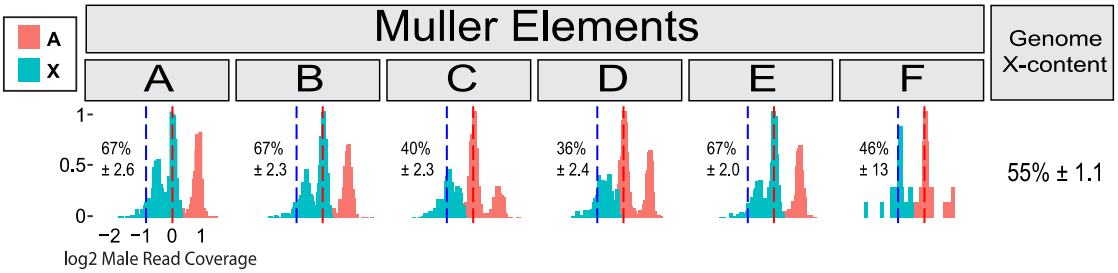


Figure 2-S2. Male DNA coverage distributions for the Cecidomyid *Lestremia cinerea*. Three distinct read coverage peaks were found in *L. cinerea*, present across all Muller elements. Our method is not designed to accommodate such multi-peak situations, and unsurprisingly breaks down in *L. cinerea*. Our current methods assign a majority of the *L. cinerea* genome as X-linked (55%), though because of the unknown identity of the far left peak, this estimate may be unreliable. This unusual distribution could be indicative of partial genome duplication, as suggested by the BUSCO results, but more investigation is needed.

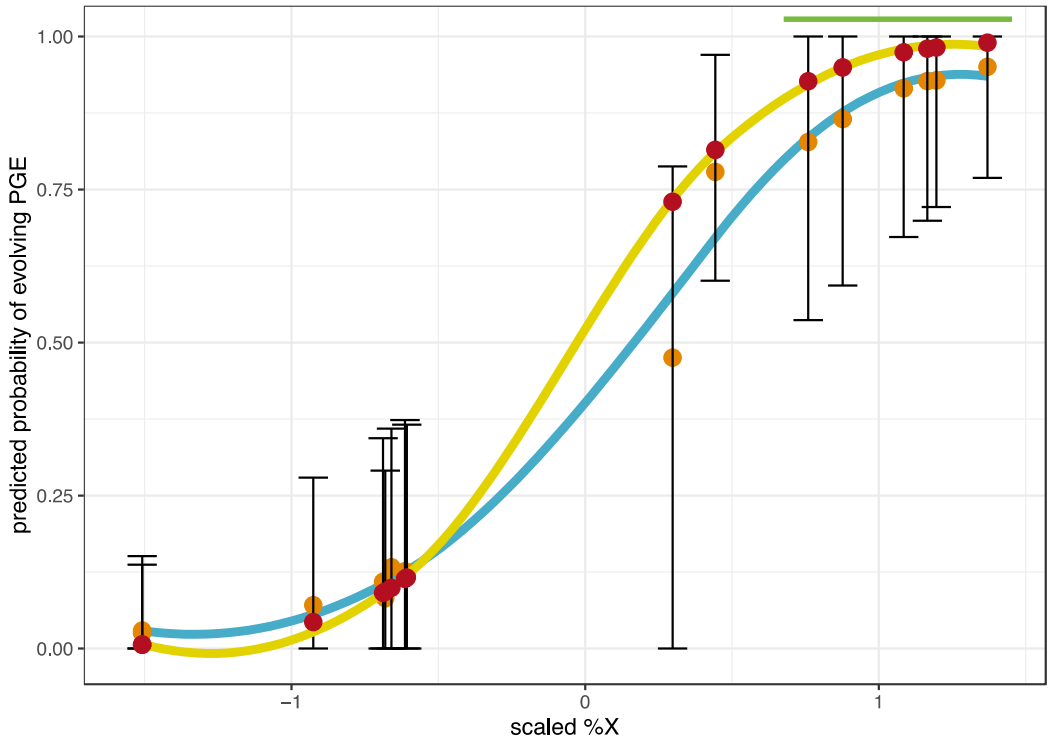


Figure 2-S3. Predicted probability of evolving paternal genome elimination (gPGE) based on percent X linkage, as estimated from the data. Orange dots are the average, out of sample, predicted probability of gPGE with the MCMCglmmRAM threshold model, and black bars are the 95% confidence interval. The blue line is the average predicted probability made with the geom_smooth function in R (method = “loess”). The yellow line is the logistic regression curve made with the fitted coefficients from the phyloglm model using the plogis function in R. Red dots are the predicted fit from the non-phylogenetic GLM, and the green bar spans the six gPGE species.

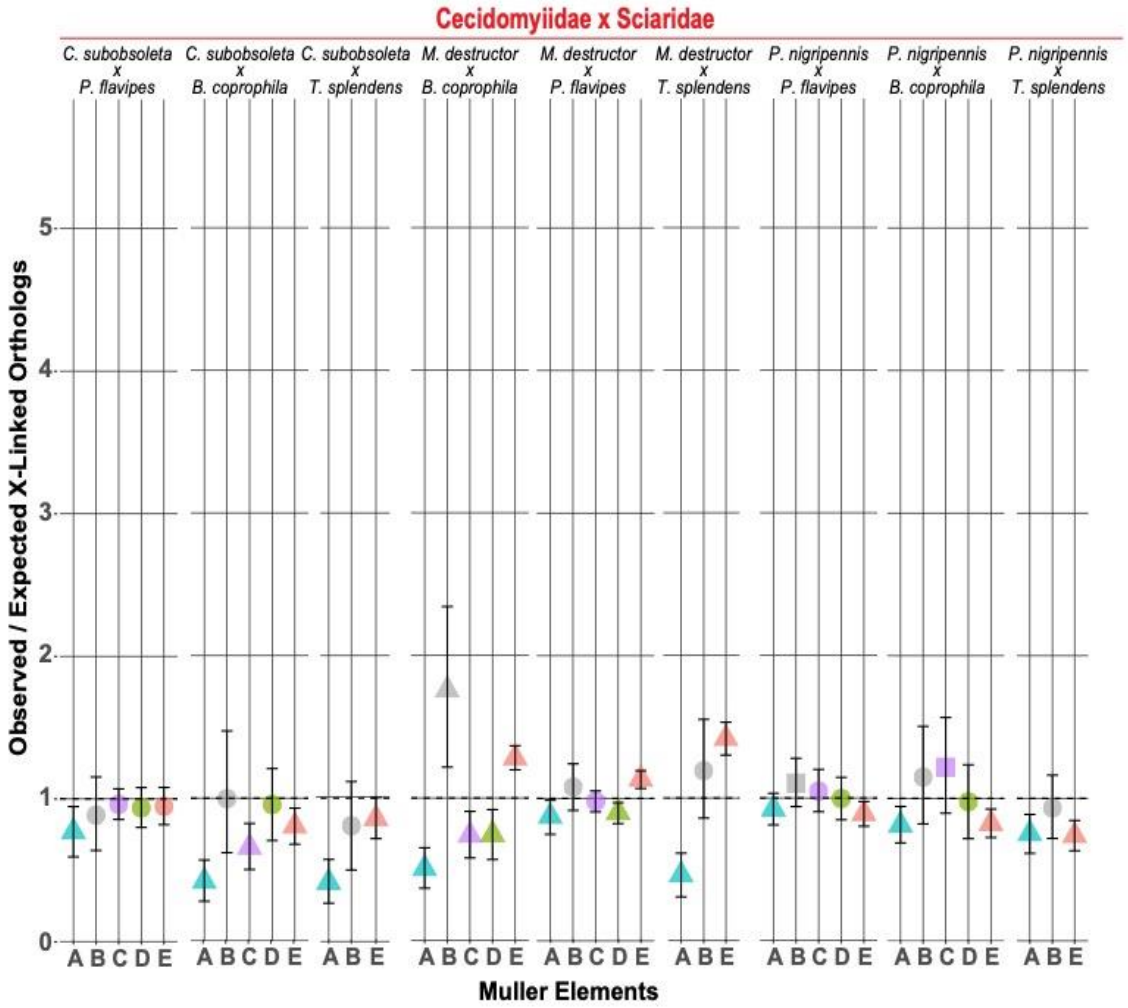
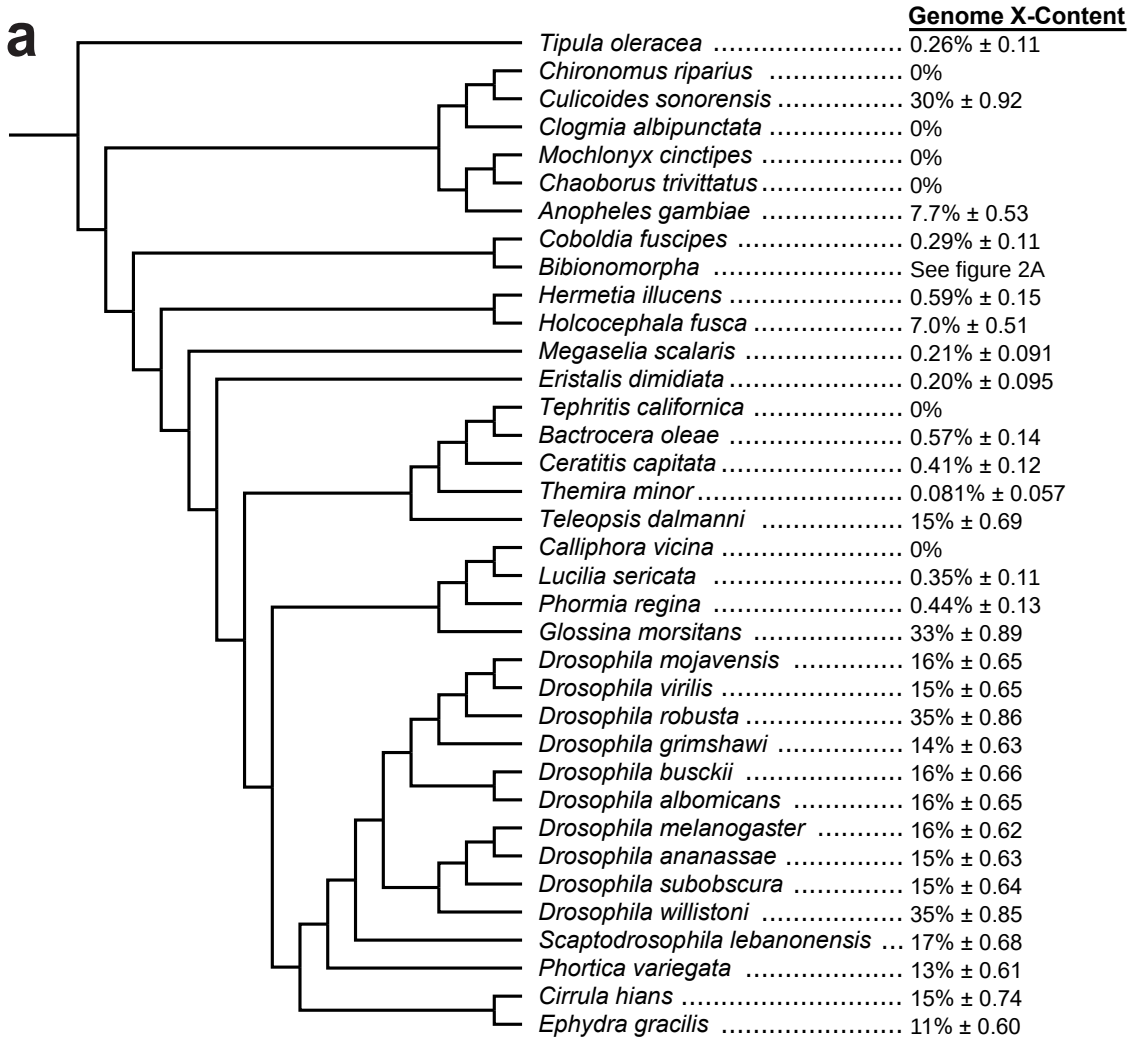
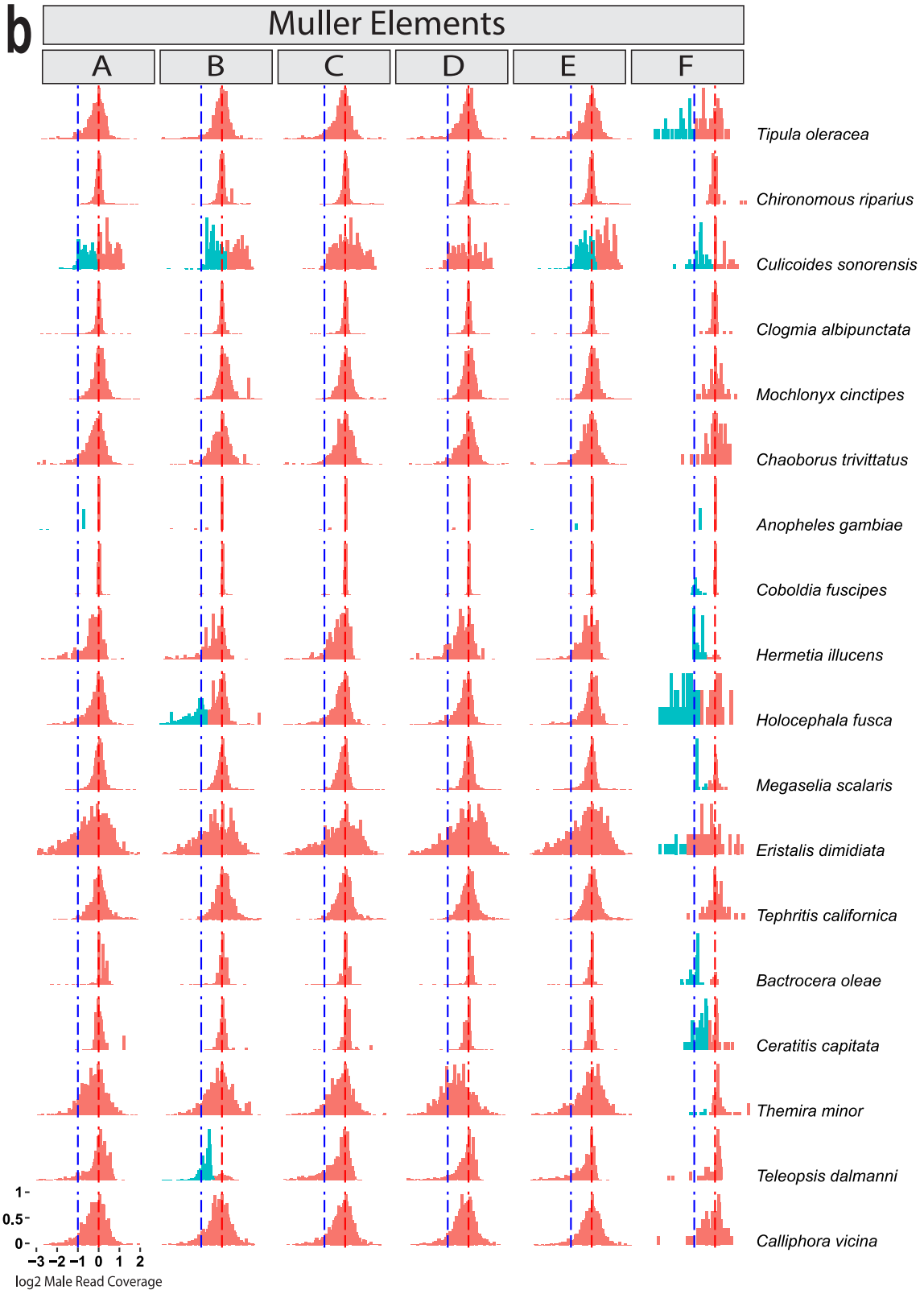


Figure 2-S4. Number of ortholog pairs in which both genes are X-linked, compared to the null expectation, for pairs of gPGE species from different families. Between-family comparisons are shown here, while within-family comparisons are shown in main Figure 2-3. Color indicates Muller element. Muller elements for which species do not share X-linked orthologs are excluded, as is the F element. Shapes indicate significance via Chi square. Error bars represent 95% CIs computed from 10,000 bootstrap replicates. Expected value if no association between X-linked orthologs is 1.


a
Genome X-Content
Tipula oleracea ..... 0.26% ± 0.11
Chironomus riparius 0%
Culicoides sonorensis ..... 30% ± 0.92
Clogmia albipunctata 0%
Mochlonyx cinctipes 0%
Chaoborus trivittatus 0%
Anopheles gambiae ..... 7.7% ± 0.53
Coboldia fuscipes ..... 0.29% ± 0.11
Bibionomorpha ..... See figure 2A
Hermetia illucens ..... 0.59% ± 0.15
Holcocephala fusca ..... 7.0% ± 0.51
Megaselia scalaris ..... 0.21% ± 0.091
Eristalis dimidiata ..... 0.20% ± 0.095
Tephritis californica 0%
Bactrocera oleae ..... 0.57% ± 0.14
Ceratitis capitata ..... 0.41% ± 0.12
Themira minor ..... 0.081% ± 0.057
Teleopsis dalmanni ..... 15% ± 0.69
Calliphora vicina 0%
Lucilia sericata ..... 0.35% ± 0.11
Phormia regina ..... 0.44% ± 0.13
Glossina morsitans ..... 33% ± 0.89
Drosophila mojavensis ..... 16% ± 0.65
Drosophila virilis ..... 15% ± 0.65
Drosophila robusta ..... 35% ± 0.86
Drosophila grimshawi ..... 14% ± 0.63
Drosophila busckii ..... 16% ± 0.66
Drosophila albomicans ..... 16% ± 0.65
Drosophila melanogaster ..... 16% ± 0.62
Drosophila ananassae ..... 15% ± 0.63
Drosophila subobscura ..... 15% ± 0.64
Drosophila willistoni ..... 35% ± 0.85
Scaptodrosophila lebanonensis ... 17% ± 0.68
Phortica variegata ..... 13% ± 0.61
Cirrula hians ..... 15% ± 0.74
Ephydra gracilis ..... 11% ± 0.60


b
Muller Elements
A
B
C
D
E
F
Tipula oleracea
Chironomous riparius
Culicoides sonorensis
Clogmia albipunctata
Mochlonyx cinctipes
Chaoborus trivittatus
Anopheles gambiae
Coboldia fuscipes
Hermetia illucens
Holocephala fusca
Megaselia scalaris
Eristalis dimidiata
Tephritis californica
Bactrocera oleae
Ceratitis capitata
Themira minor
Teleopsis dalmanni
Calliphora vicina
1
0.5
0
−3 −2 −1 0 1 2
log2 Male Read Coverage

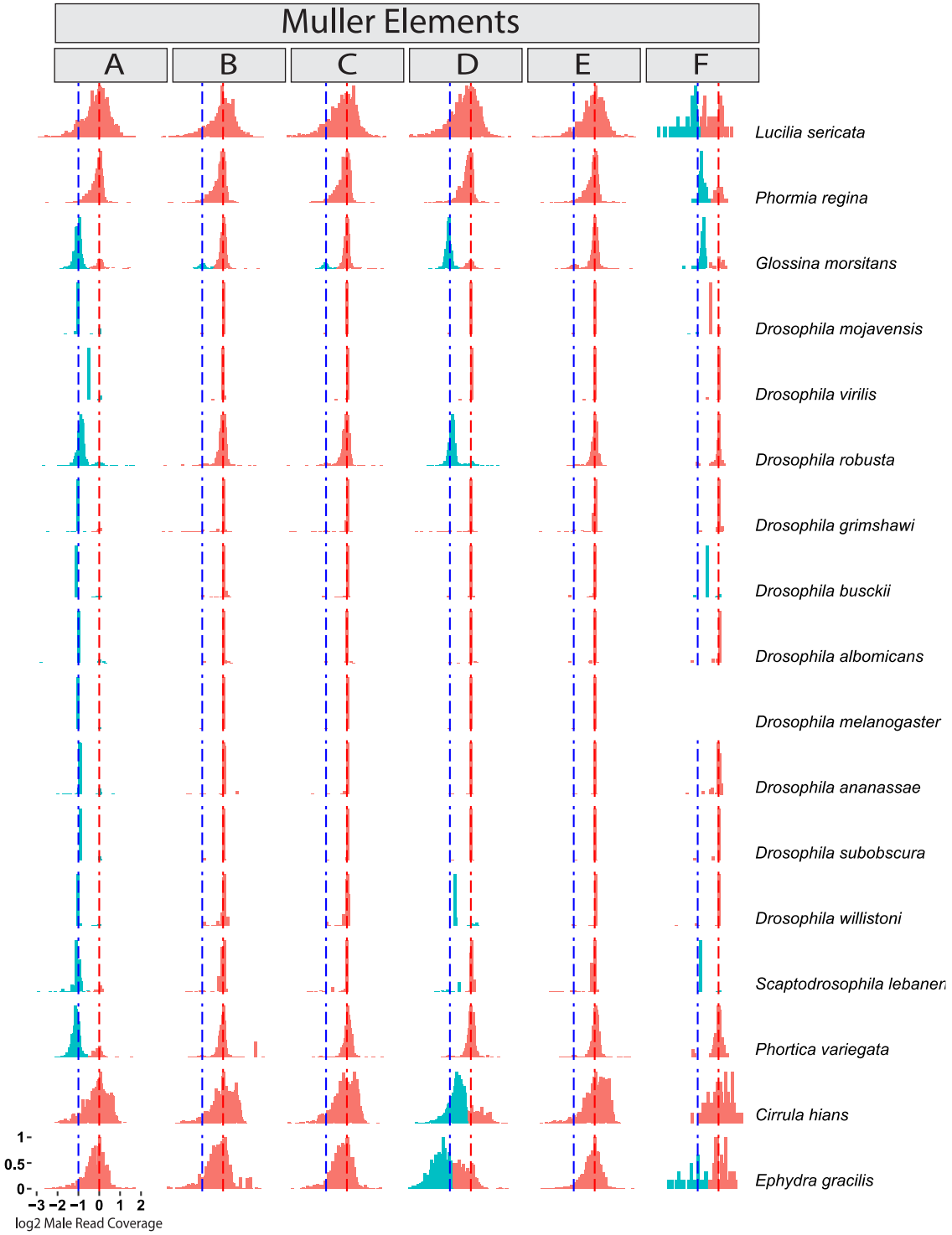


Figure 2-S5. Distribution of X-linked genome content across diverse dipterans, assessed by male DNA read coverage. a) Phylogenetic tree showing the percent of X-linked genes estimated from the whole genome. b) Frequency of X-linked genes across Muller elements, showing log2 male read coverage normalized by

putative median autosomal coverage, with assigned X linkage (blue) and autosomal linkage (red) indicated. The y-axis represents gene frequency scaled to the maximum in each distribution. Red dashed vertical lines at 0 indicate the expected autosomal coverage peak, blue dashed lines at -1 indicate the expected position of the X-linked peak, at half the coverage of the autosomes. Percent estimates represent percent X linkage for each Muller and across each full genome, with error represented by 2SD. *Drosophila melanogaster* F element, the dot chromosome, is excluded for insufficient gene number.

Table 2-S1. The estimated percentage of the full genome that is X-linked and the per-Muller element estimates for Bibionormorphan and non-Bibionormorphan dipterans. Error is represented as twice the standard deviation.

Species	**X-linked percent of genome**	**A**	**B**	**C**	**D**	**E**	**F**
Bibionomorphan species included in main text figures							
Sciara coprophila	24% ± 0.91	43% ± 2.6	13% ± 1.7	12% ± 1.6	0%	44% ± 2	8.6% ± 6.7
Phytosciara flavipes	44% ± 1.1	57% ± 2.6	35% ± 2.4	33% ± 2.3	35% ± 2.3	58% ± 2.1	25% ± 11
Trichosia splendens	16% ± 0.75	38% ± 2.6	0%	0%	0%	39% ± 2	0%
Diadocidia ferruginosa	0.63% ± 0.16	0%	0%	0%	0%	0%	78% ± 9.6
Gnoriste bilineata	0.52% ± 0.15	0%	0%	0%	0%	0%	70% ± 11
Exechia fusca	5.5% ± 0.59	33% ± 3.2	0%	0%	0%	0%	55% ± 14
Macrocera vittata	0.26% ± 0.1	0%	0%	0%	0%	0%	46% ± 14
Platyura marginata	0%	0%	0%	0%	0%	0%	0%
Bolitophila cinerea	0.62% ± 0.16	0%	0%	0%	0%	0%	79% ± 9.5
Bolitophila hybrida	0.51% ± 0.14	0%	0%	0%	0%	0%	68% ± 11
Mayetiola destructor	41% ± 1.1	17% ± 2.1	17% ± 1.9	70% ± 2.3	59% ± 2.6	40% ± 2.2	75% ± 11
Porricondyla nigripennis	66% ± 1.3	67% ± 3.1	59% ± 3.1	68% ± 2.7	64% ± 2.9	71% ± 2.3	66% ± 14
Catotricha	34% ± 0.96	17% ±	16% ±	67% ±	42% ±	23% ±	79% ±

subobsoleta		1.9	1.8	2.1	2.3	1.7	9.2
Lestremia cinerea	55% ± 1.1	67% ± 2.6	67% ± 2.3	40% ± 2.3	36% ± 2.4	67% ± 2	46% ± 13
Symmerus nobilis	0%	0%	0%	0%	0%	0%	0%
Penthetria funebris	0.55% ± 0.15	0%	0%	0%	0%	0%	72% ± 10
Sylvicola fuscatus	7.7% ± 0.55	0%	4.6% ± 1	0%	8.8% ± 1.4	17% ± 1.5	82% ± 8.8
Non-Bibio-nomorphan dipterans	***X-linked percent of genome***	***A***	***B***	***C***	***D***	***E***	***F***
Tipula oleracea	0.26% ± 0.11	0%	0%	0%	0%	0%	36% ± 12
Chironomus riparius	0%	0%	0%	0%	0%	0%	0%
Culicoides sonorensis	30% ± 0.92	46% ± 2.5	56% ± 2.3	0%	0%	48% ± 2	57% ± 11
Clogmia albipuntata	0%	0%	0%	0%	0%	0%	0%
Mochlonyx cinctipes	0%	0%	0%	0%	0%	0%	0%
Chaoborus trivittatus	0%	0%	0%	0%	0%	0%	0%
Anopheles gambiae	7.7% ± 0.53	29% ± 2.3	0%	0%	0%	11% ± 1.3	29% ± 11
Coboldia fuscipes	0.29% ± 0.11	0%	0%	0%	0%	0%	39% ± 11
Hermetia illucens	0.59% ± 0.15	0%	0%	0%	0%	0%	79% ± 9.1
Holcocephala fusca	7.0 % ± 0.51	0%	35% ± 2.2	0%	0%	0%	66% ± 11
Megaselia abdita	0.21% ± 0.091	0%	0%	0%	0%	0%	31% ± 11
Eristalis dimidiata	0.20% ± 0.095	0%	0%	0%	0%	0%	25% ± 10
Tephritis californica	0%	0%	0%	0%	0%	0%	0%
Bactrocera oleae	0.57% ± 0.14	0%	0%	0%	0%	0%	82% ± 8.7

Ceratitis capitata	0.41% ± 0.12	0%	0%	0%	0%	0%	59% ± 11
Themira minor	0.081% ± 0.057	0%	0%	0%	0%	0%	10% ± 6.9
Teleopsis dalmanni	15% ± 0.69	0%	81% ± 1.8	0%	0%	0%	0%
Calliphora vicina	0%	0%	0%	0%	0%	0%	0%
Lucilia sericata	0.35% ± 0.11	0%	0%	0%	0%	0%	51% ± 12
Phormia regina	0.44% ± 0.13	0%	0%	0%	0%	0%	66% ± 11
Glossina morsitans	33% ± 0.89	84% ± 1.8	11% ± 1.4	10% ± 1.3	81% ± 1.7	0%	79% ± 9.4
Drosophila mojavensis	16% ± 0.65	100% ± 0	0%	0%	0%	0%	5.1% ± 4.9
Drosophila virilis	15% ± 0.65	100% ± 0	0%	0%	0%	0%	0%
Drosophila robusta	35% ± 0.86	100% ± 0	0%	0%	100% ± 0	0%	0%
Drosophila grimshawi	14% ± 0.63	88% ± 1.5	0%	0%	0%	0%	0%
Drosophila busckii	16% ± 0.66	100% ± 0	0%	0%	0%	0%	100% ± 0
Drosophila albomicans	16% ± 0.65	100% ± 0	0%	0%	0%	0%	0%
Drosophila melanogaster	16% ± 0.62	100% ± 0	0%	0%	0%	0%	0%
Drosophila ananassae	15% ± 0.63	100% ± 0	0%	0%	0%	0%	0%
Drosophila subobscura	15% ± 0.64	100% ± 0	0%	0%	0%	0%	0%
Drosophila willistoni	35% ± 0.85	100% ± 0	0%	0%	100% ± 0	0%	0%
Scapto-drosophila lebanonensis	17% ± 0.68	91% ± 1.3	0%	0%	13% ± 1.4	0%	100% ± 0
Phortica variegata	13% ± 0.61	83% ± 1.8	0%	0%	0%	0%	0%
Cirrula hians	15% ± 0.74	0%	0%	0%	74% ± 2	0%	0%
Ephydra gracilis	11% ± 0.60	0%	0%	0%	59% ± 2.1	0%	38% ± 11

Chapter 3: A complex history of leucine biosynthesis genes in fungi: gene fusion, fission, loss and horizontal transfer

Submitted: Noelle Anderson, Steven Sun, Scott William Roy

3.1 Abstract

Fungi exhibit unique metabolic capabilities and are highly dependent on their metabolic adaptations to thrive in diverse environments, making them an excellent model for studying metabolic adaptation and evolution. Here, we trace the distribution and history across fungi of a gene fusion involving two essential enzymes that work in subsequent steps in fungal leucine biosynthesis, 3-isopropylmalate dehydrogenase (IPMDH) and 2-isopropylmalate isomerase (IPMI). Through phylogenetic reconstruction, we find evidence for a complex history involving ancestral IPMDH-IPMI fusion, secondary fission and loss of the fused genes, potential cases of secondary fusion, and multiple cases of horizontal gene transfer. While genes involved in the same metabolic pathways are often physically associated in ways thought to improve metabolic efficiency, suggesting adaptive significance for the fused gene, recurrent loss or fission of the fused gene complicates this narrative. The results presented here represent a remarkably intricate history for a pair of key enzymes, highlighting the complexities not captured by current dominant models of molecular evolution.

Three figures too large for the printed thesis format are included as supplemental files: 3-S1 and 3-S2 represent full uncollapsed gene trees for separate IPMDH and IPMI domains with NCBI accession numbers and JGI protein IDs with taxa strain identifiers. 3-S3 represents a species tree from TimeTree of Pezizomycotina species in our primary search database. Full legends for supplemental figures are included in section 3.6 after the discussion.

3.2 Introduction

Fungi, unlike most eukaryotes, rely entirely on osmotrophic nutrient uptake and take in diverse substrates. Because of this, they are highly dependent on their metabolic capabilities, including their ability to specialize their metabolism ((Watkinson 2016; Naranjo-Ortiz and Gabaldón 2019; Naranjo-Ortiz and Gabaldón 2020)**,** Naranjo-Ortiz and Gabaldón 2020). This fact, coupled with the availability of large genomic datasets, makes fungi an excellent model for studying metabolic adaptation and evolution.

Diffusion of products between enzymes can be a limiting step in metabolic pathways, a problem that can be mitigated in part by the physical association of enzymes that act in sequential steps of the pathway, often by means of increased substrate channeling efficiency, sometimes within enzyme complexes (so-called metabolons) (Sweetlove and Fernie 2018). The physical association of functionally related genes in gene clusters that has been frequently observed in fungal genomes and may be related to the increased importance of optimizing metabolism in fungi (Wisecaver et al. 2014; Wisecaver and Rokas 2015; Slot 2017; Nützmann et al. 2018; Rokas et al. 2018; Rokas et al. 2020).

While metabolic gene clusters have been better studied for their metabolic properties, gene fusions are a less appreciated path to improving metabolism by physically associating genes. Indeed, metabolically related gene fusions have demonstrated improved enzyme kinetics over their unfused counterparts and substrate channeling, notably with yeast mitochondrial TCA cycle enzymes in a biochemical reaction (Lindbladh et al. 1994; Elcock and Andrew McCammon 1996), though this claim has also been challenged on the basis of how kinetics are evaluated (Pettersson et al. 2000). In a study including a fusion of yeast glycerol pathway enzymes transformed in *E. coli*, multi-fold improvements in enzyme kinetics and glycerol biosynthesis were observed *in vivo* and in purified reactions (Meynial Salles et al. 2007). A single peptide with linked functional domains used in adjacent steps in a metabolic pathway may thus confer an advantage relative to unfused gene copies. However, in other cases no evidence has been found for increased metabolic flux in observed fusions of adjacent genes in a metabolic pathway (Castellana et al. 2014), and in some cases fusions show reduced enzyme activity compared to free enzymes (Kourtz et al. 2005).

Another important process in metabolic evolution in fungi is lateral gene transfer (Richards 2011; Richards and Talbot 2013; Naranjo-Ortiz and Gabaldón 2020). While much more frequent among prokaryotes, this mode of evolution has now been observed across all three domains of life, often even involving HGT from one domain of life to another (Zhaxybayeva and Doolittle 2011; Gabaldón 2020; Cote-L'Heureux et al. 2022). Horizontally transferred genes can have many positive effects on the genome, bringing in novel functions, replacing ancestral copies with equivalent functions, and even replacing ancestral copies with slightly (or more so) improved copies (Soucy et al. 2015). Genes that have been horizontally transferred in fungi are overrepresented in primary and secondary metabolic functions and gene clusters have been shown to be frequently horizontally transferred (Richards et al. 2011; Fitzpatrick 2012; Wisecaver et al. 2014; Wisecaver and Rokas 2015).Thus, fungi provide an abundance of opportunities for understanding the role of metabolically-related physically associated genes that may improve metabolic efficiency, and how HGT can be a path to propagating these metabolic improvements between divergent species.

Here, we report the complex evolution of fungal 3-isopropylmalate dehydrogenases (IPMDH, EC 1.1.1.85) and 2-isopropylmalate isomerases (IPMI, EC 4.2.1.33), two enzymes sequential in leucine biosynthesis and with confamilial variants in the TCA cycle (Fig. 3-1). In leucine biosynthesis, IPMI catalyzes the isomerization of 2-isopropylmalate into 3-isopropylmalate, while IPMDH decarboxylates the 3-isopropylmalate into α-ketoisocaproate. This fusion was previously identified in fungi and phylogenetically analyzed by Leonard and Richards, but too little fungal data was available at that time to reveal the more complex story we illustrate here (Leonard and Richards 2012). Consistent with roles for gene fusion and HGT in improving existing metabolic pathways, we find evidence for fused genes that encode both enzymes as separate domains, and for HGT of these fused constructs. At the same time, we also find evidence for recurrent fission and loss of these fused copies, complicating the simple story of directional evolution for increased metabolic efficiency. These results demonstrate the complex evolutionary histories within eukaryotes of even the most essential and well-understood genes.

Figure 3-1. IPMDH-IPMI gene fusion domain structure. Length ranges show 10-90th percentiles across fused genes observed across fungi.

3.3 Methods

Stimulated by a report of a gene fusion involving an aconitase gene in fission yeast (Jung et al. 2015), we used the 'Clusters' tool on the JGI's Mycocosm website to manually search for unknown additional aconitase fusions (Grigoriev et al. 2011; Grigoriev et al. 2014). Multiple fusions of IPMDH-IPMI were found and in each case, the IPMDH domain (PF00180/IPR024084) lay upstream of the IPMI domain (PF00330/IPR001030) with an intervening linker region mode of 95 residues (Fig. 3-1). Additional IPMDH-IPMI fusions not found in JGI's Mycocosm Clusters were identified using NCBI's Conserved Domain Architecture Retrieval Tool (Geer et al. 2002). IPMDH and IPMI were then searched separately against all available fungal proteomes from the 1000 Fungal Genomes dataset in December 2018 (Grigoriev et al. 2014), and Genbank's non-redundant (nr) protein database using BlastP 2.7.1 (Altschul et al. 1990; Camacho et al. 2009) and highly-significant results compiled (e.g., e-value < 10^{-20}). To compile the unfused homologs, 50 unfused versions of IPMDH and IPMI were randomly (but broadly taxonomically) chosen from JGI's database and pblast searches were performed against all the same two databases. Additional sampling of taxa within groups underrepresented in our initial blast search was done to ensure adequate sampling. Full non- genera or

bootstrap value-collapsed IPMDH and IPMI gene trees with taxa NCBI accessions or JGI protein IDs are available in supplemental materials.

Domain boundaries were identified HMMscan using HMMER version 3.1b2 (http://hmmer.org/) sequences were then split into the respective IPMDH and IPMI domains before aligning individually using MAFFT with alignment strategy FFT-NS-2 (Katoh et al. 2002). Model testing was performed in IQ-TREE version 1.6.10 (Nguyen et al. 2015) and for both alignments, the LG model with 6 rate categories was chosen. Gene trees were constructed using IQ-TREE and ultrafast bootstrapping approximation with 10,000 pseudoreplicates was performed (Minh et al. 2013).

Horizontal transfer events were identified by comparing the gene trees to the species tree. If the species with the identified fusions cluster while the rest of the phylogeny remains largely congruent with the species tree topology, the genes are likely to be derived from a common fusion event with multiple rounds of HGT, particularly when the taxonomic distribution of the fusion is punctate.

For searching the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) dataset (Keeling et al. 2014), the fused protein sequence JGI protein ID 441756 from the *Entophlyctis helioformis* JEL805 v1.0 genome was used as a BlastP query against all translated transcriptomes of the MMETSP project.

3.4 Results

3.4.1 Kingdom-wide search for IPMDH-IPMI fusions in fungi

In our manual search of JGI's Mycocosm, we identified genes containing IPMDH and IPMI (two enzymes in the isocitrate/isopropylmalate dehydrogenase and aconitase families, respectively) fused together in various Pezizomycotina species. Intrigued, we searched for the fusion across all available fungal protein sequences in Genbank and JGI's Mycocosm database. To our surprise, we found that fusion genes were present in nearly all major groups. However, despite this broad distribution, the overall phylogenetic distribution of the fusion is punctate, suggesting a complex history. Fused genes are absent from entire groups (e.g. Basidiomycota, Saccharomycotina, Monoblepharidomycota, Microsporidia and Rozellomycota), and are found in only some species within other groups (e.g., Pezizomycotina, Taphrinomycotina, Mucoromycotina, Blastocladiomycota, Chytridiomyceta, Aphelidiomycota) (See Fig. 3-2).

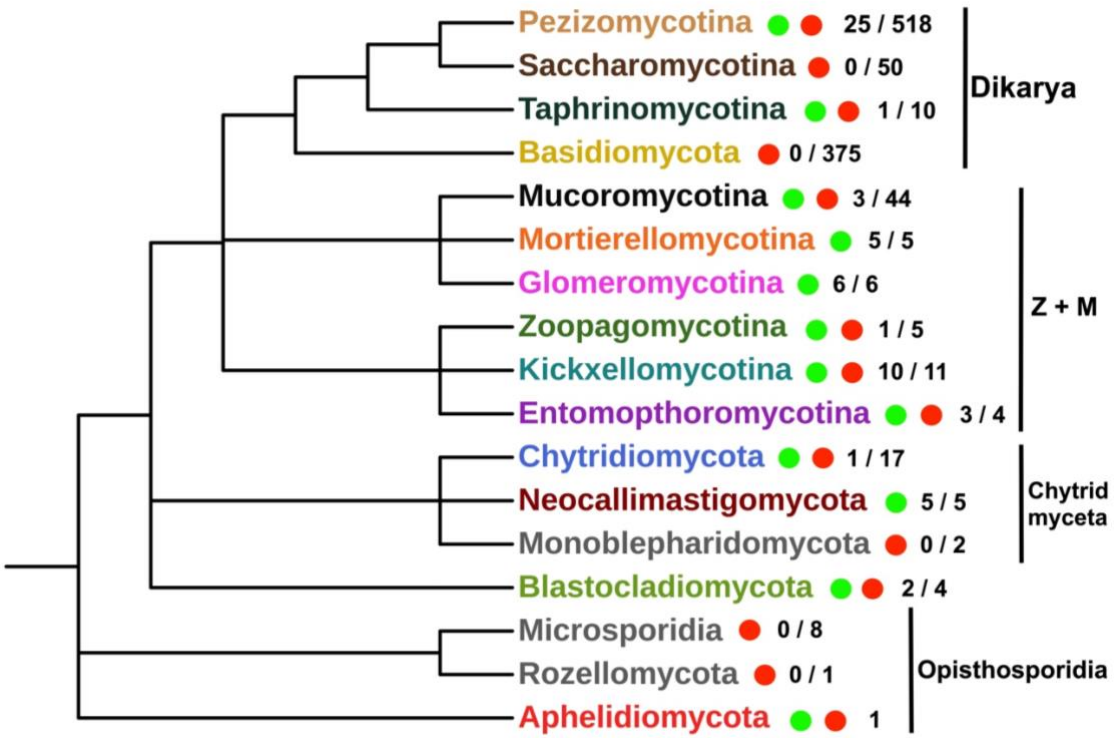


Figure 3-2. Cladogram indicating the phylogenetic distribution of the fusion gene across all major fungal groups, with topology based on Li et al. 2021. Groups in which all assessed species either contain the fusion gene (green circle) or lack the fusion gene (red circle), or in which diversity within the group is observed (green and red circles together). Values by each taxonomic group represent the number of species with confirmed fusions found within that group (left), out of the number in the search database from JGI's Mycocosm at the time of download, plus any fusions in species found only through NCBI (right). The single Aphelidiomycota fusion was located via NCBI.

3.4.2 Phylogenetic analysis of fused genes and unfused homologs

We used phylogenetic analysis to better understand the history of the IPMDH-IPMI fusion genes and their irregular distribution across fungi. We first separated fused genes into their two constituent domains, and then conducted extensive recursive BlastP searches of the nr database on NCBI and the JGI fungal genomes database to identify as many homologs as possible, both fused and unfused, from across available fungal genomes. We also performed searches across eukaryotes and prokaryotes more generally, and found a single fused gene outside of fungi in the Aphelidiomycota phylum (in the species *Amoeboaphelidium occidentale,* KAI3662018.1), which represents Opisthosporidia, a sister group to fungi containing Aphelidiomycota, Rozellomycota, and Microsporidia.

We separately performed phylogenetic reconstruction of gene trees, one each for the IPMDH- and IPMI domains, along with homologs from across fungal species and the aphelid *Amoeboaphelidium occidentale* (Fig. 3-2). Through much of the trees, there is a high degree of similarity between the two trees (described in detail below), as expected for multi-domain genes with a single shared history. However, other parts of the obtained trees revealed complex histories. For both IPMDH and IPMI trees, most fused domains fell within two major clades, both of which consisted exclusively or almost exclusively of fused domains. We now discuss these two major clades in detail.

3.4.3 A diverse clade of fused genes suggests the presence of a fused IPMDH-IPMI gene in early fungal ancestors

For both of the IPMDH and IPMI gene trees, the larger of the two clades contained species from two major fungal phyla (Zoopagomycota and Mucoromycota) with representatives from all subphyla within (Mortierellomycotina, Mucoromycotina, Glomeromycotina, Zoopagomycotina, Kickxellomycotina, Entomopthoromycotina) containing the fused gene. Thus we refer to this clade as Z+M. Note that while Zoopagomycota and Mucoromycota were formerly classified as a clade under the group Zygomycetes based on morphology and reproduction, later evidence supported Zygomycete paraphyly where either group forms a clade with Dikarya, but still more recent event supports the initial hypothesis Z+M monophyly – thus the true relationships are best regarded as unresolved (Li et al. 2021).

For both IPMDH and IPMI, within the Z+M clade, we found that the gene tree topology of the clades (see simplified tree summaries in Fig. 3-3) has clear similarities to known organismal relationships between the represented fungal groups (Fig. 3-2). This is particularly the case for IPMDH, which recovers the Mucoromycota clade (Glomeromycotina, Mortierellomycotina, and Mucoromycotina) with moderately high support (88bs), as well as a well-supported clade containing species within Zoopagomycota (Zoopagomycotina, Kickxellomycotina, and Entomophthoromycotina) species, as well as the speculated sister relationship between these two large clades (64bs) (Fig. 3-3, left). In the IPMI gene tree (Fig. 3-3, right), we reconstruct a highly supported (100bs) clade containing Mucoromycota and Zoopagomycota and while species from each subphyla group with high support and both Mucoromycotina and Glomeromycotina (97bs) and Kickxellomycotina and Zoopagomycotina (77bs) respectively form sister clades, relationships between other subphyla are less stable, with both Entomopthoromycotina and Mortierellomycotina species constituting a clade, but grouping outside of their expect subphyla-level relationships. Given the species-like IPMDH grouping and the modest IPMI bootstrap support (77-85bs), we speculate that these portions of the tree also represent vertical inheritance but with phylogenetic reconstruction errors in the IPMI tree reconstruction.

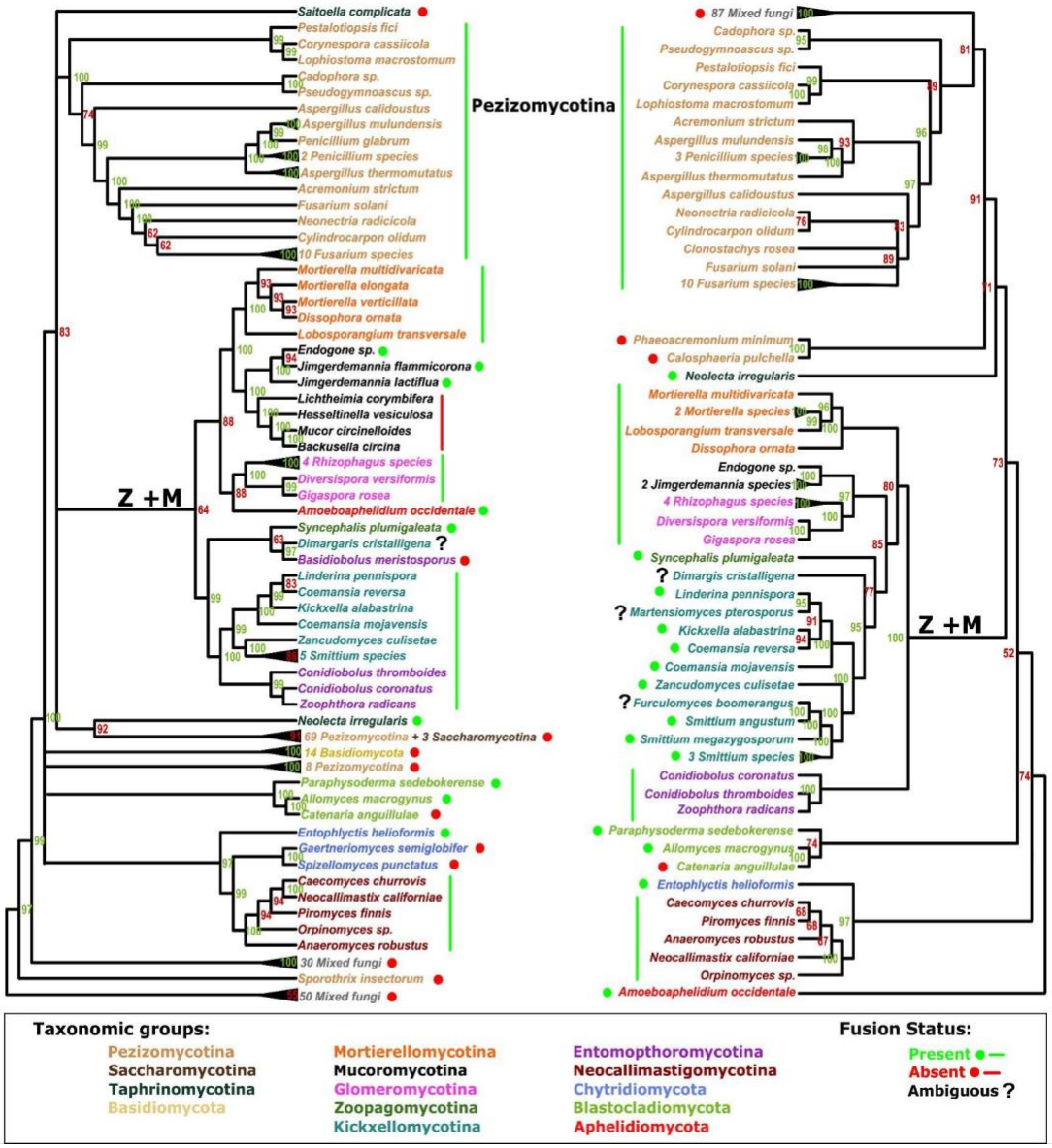


Figure 3-3. IPMDH (left) and IPMI (right) summary gene trees. Species with IPMDH-IPMI fusions are indicated with green circles and bars and those without with red circles and bars. Ambiguous cases (mostly likely missannotations, see text) are shown with question marks. Species are colored by subphyla, and congeners have been collapsed. Only nodes with > 50 ultrafast bootstrapping support (from 10,000 pseudoreplicates) are shown as resolved and BS values below 95 are shown in red (See Supplemental Figures 3-S1 and 3-S2 for full uncollapsed trees). The clade of fusions containing Zoopagomycota and Mucoromycota is marked as 'Z+M'.

The only clear exceptions to the Z+M broad species-like grouping are observed in the IPMDH gene tree. First, the aphelid *Amoeboaphelidium occidentale* falls within the Z+M clade as sister to a species-tree like group of Glomeromycotina. Second,

within the IPMDH Z+M clade, in contrast to the clearly species-like relationships between most species, there exists an additional three-gene clade including genes from three different Zoopagomycota groups.

This pattern, in which the gene tree in these clades largely follows the species tree, is as expected if the IPMDH-IPMI fusion was present in deep fungal ancestors and has been retained by vertical inheritance by a wide diversity of extant fungi. On the other hand, this pattern is difficult to explain otherwise. The only major possible exception to this species-like pattern is the lack of any representatives of the Dikarya (comprising Ascomycota + Basidiomycota). This could represent gene loss early in Dikarya or, depending on phylogeny (see above), gene gain after the Z+M lineage diverged from Dikarya.

Is it possible that this gene fusion evolved even deeper within the fungal tree? Notably, both gene trees include fused genes from the other two deeper branching major fungal groups [Blastocladiomycota and Chytridiomyceta (Neocallimastigomycota, Chytridiomycota, and Monoblepharidomycota), as well as the sister group to fungi (Aphelidiomycota, represented by the aphelid *Amoeboaphelidium occidentale*)], including most available Neocallimastigomycota and Blastocladiomycota species (Fig. 3-3). However, the case is far from simple. First, no fusions besides the single aphelid were found in other Opisthosporidians. Second, only one Chytridiomycota species contains the fusion, and no fusions were found in the closely related Monoblepharidomycota clade. Third, the Blastocladiomycota and Chytridiomyceta fusions do not group into a larger fusion clade along with the Z+M clade, as would be expected by this fungal ancestor hypothesis. Indeed, the Z+M clade groups closer to paralogs whose own broad representation and semi-species-like tree structure suggests they are ancient paralogs (and thus should not fall within another ancestral gene clade). In total, then, there is no clear preponderance of evidence for a fused gene in the fungal ancestor.

3.4.4 Evidence for recurrent lateral gene fusion in Pezizomycotina

A strikingly different pattern is seen in the second major clade of fused genes present in both trees, containing genes from within the large Ascomycota group Pezizomycotina. For this clade, the history implied is quite different at the gene and species level. At the level of the gene (i.e., gene fusion and fission), the clade's history may be straightforward: within these clades the two domains mostly show highly similar topologies for the two genes, suggesting a shared evolutionary history for an ancestrally fused gene.

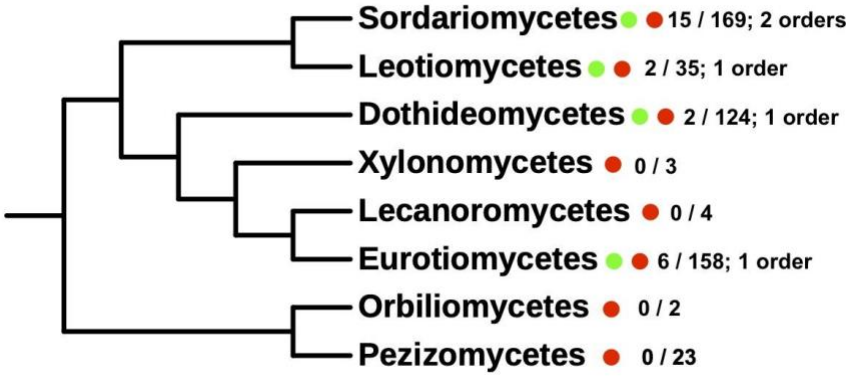


Figure 3-4. Presence and absence of IPMDH-IPMI fusions in Pezizomycotina classes. Species tree topology based on Li et al. 2021. Red circles indicate that species without the IPMDH-IPMI fusions were found in this class. Green circles represent fusion presence. For each class, numerator and denominator give the number of species containing fusions and the number of species in that class present in the JGI Mycocosm database at the time of searching (plus any fusions found on NCBI), respectively. For classes containing fusions, the number of orders represented with fusions is also given.

On the other hand, at the species level the tree is far less simple. First, the fused genes in this clade represent only a small fraction of studied Pezizomycotina species [21 out of 514 Pezizomycotina species searched from the downloaded JGI database (note that some fusions were found on NCBI and have been included in both numerator and denominator counts in Figure 3-4, e.g. 4 Pezizomycotina species with fusions were found only on NCBI, thus is counted as 25/518)]. However, the species that contain the fused gene are largely distantly related, being found in 4 different orders and 5 different classes (Fig. 3-3, Fig. 3-4). Moreover, the gene trees do not reflect species tree relationships (Supplemental figure 3-S3). Both of these patterns are not at all as expected by vertical inheritance (which would require massive parallel gene loss and widespread failure of phylogenetic reconstruction specific to this clade). Instead, both patterns are as expected by multiple HGT events among Pezizomycotina species (and, notably, only among Pezizomycotina species, as non-Pezizomycotina species are absent from this clade).

Taking a closer look at the Pezizomycotina species tree topology, we can infer at least 11-15 lateral transfer events (Fig. 3-4, Supplemental figure 3-S3). The range given is due to the possibility of either two transfers within Sordariomycetes (into the Hypocreales ancestor and the distantly related Xylariales species *Pestalotiopsis fici*, explainable with 7 losses based on the JGI Mycocosm species tree) or five transfers (transfers into each Sordariomycetes fusion species on the tips, except in Fusarium where most searched species have the fusion and thus likely was an ancestral gain explainable with 3 losses within Fusarium, discussed

below), or some intermediate. Notably the Sordariomycetes species do not group in species relationships outside of Fusarium (Fig. 3-3). In the other Pezizomycotina groups, while fusions occurred in single classes, the JGI Mycocosm species trees were also used to examine the positions of species with fusions to look for patterns of lateral transfer or gene loss. In all cases, species are sufficiently distantly related to parsimoniously conclude lateral transfer over gene loss.

3.4.5 Other possible HGT events are less well-supported

Other potential lateral gene transfer events are also observed in the larger trees, with clear and sometimes striking differences between the gene tree and species tree. These include the aphelid *Amoeboaphelidium occidentale* grouping with Glomeromycotina in the IPMDH tree and the presence of a fused gene in the single Taphrinomycotina species *Neolecta irregularis* (OLL24270.1) grouping with unrelated species (Fig. 3-3). However, these cases do not include the second line of evidence found in Pezizomycotina, namely spotty phylogenetic distribution of the fusion gene (their phylogenetic distribution being impossible to assess given relatively paltry genomic sequence of these lineages). Thus while these cases suggest HGT, given the real possibility of phylogenetic errors, we believe that misplacement alone is insufficient evidence to conclude an instance of lateral gene transfer. Possibly ongoing sequencing of these groups will allow for distinguishing the history of these genes in the future.

3.4.6 Evidence for gene fission and domain loss

Among the complex changes we see across the tree, we observe evidence for multiple instances of gene fission and domain loss, particularly loss of the IPMI domain. First, we observe a clade of four species within Mucoromycotina that lack a fused gene, but which contain unfused IPMDH but not IPMI genes (Fig. 3-3). The unfused IPMDH genes group in species-tree positions with the fused genes found within other Mucoromycotina species, exactly as expected if the ancestrally fused gene lost the IPMI gene in the ancestor of these four species. Indeed, the species with the IPMDH-IPMI fusions form a single clade (Endogonomycetes), the earliest branching Mucoromycotina class, suggesting a single event occurring after this early divergence involving fission of the ancestral fused gene and IPMI domain loss. Elsewhere in the three, the Entomopthoromycotina species *Basidiobolus meristoporus* also has an unfused IPMDH annotated without an IPMI domain (or any evidence for IPMI nearby in the genome). This unfused gene groups with related Zoopagomycota species in the IPMDH tree, consistent with secondary domain loss.

Four other potential cases of fission could be due to the possibility of annotation errors. Both the Kickxellomycotina species *Dimargaris cristalligena* (IPMDH: RKP40269.1, IPMI: RKP40268.1) and the Blastocladiomycota species *Catenaria anguillulae* (IPMDH: ORZ31678.1, IPMI: ORZ31677.1) IPMDH and IPMI copies

annotated as genomically adjacent but separate genes. While these cases could possibly indicate gene fission, it is hard to explain simple missannotation of a single fused gene as separate genes. Both species fall within the Z+M fusion clade and while *C. anguillulae* falls in its phyla-level species tree relationship in both gene trees and *D. cristalligena* in the IPMI tree, in the IPMDH gene tree, *D. cristalligena* groups in the 3-gene Zoopagomycota group shared with *B. meristoporus* discussed above (Fig. 3-3). The Kickxellomycotina species *Martensiomyces pterosporus* (strain CBS 209.56 v1.0, JGI protein ID 265168) and *Furculomyces boomerangus* (PVU91561.1) are present in the IPMI but not IPMDH Z+M fusion clade, however, manual inspection revealed that these exceptions are likely due to a truncated annotated gene sequence that we could not confirm as a true fission.

We also observe potential evidence for complete gene loss (e.g. as previously discussed in the dikaryon ancestor, and in the Chytrid phylum Monoblepharidomycota, depending on the species tree phylogeny and ancestral fusion timing). Another potential complete loss includes the Blastocladiomycota species *Blastocladiella britannica*, sister to *Catenaria anguillulae* which possessed unfused gene copies while grouping with the other Blastocladiomycota fusions in the IPMDH and IPMI gene trees, while *B. britannica* falls deep outside of the fusion clades. A similar pattern is seen for several Zoopagomycotina species, where one fusion groups with others in its phylum within the Z+M clade, while four searched species showed unfused genes and appear elsewhere or absent from the tree, indicating several losses. In Pezizomycotina, *Fusarium pseudograminearum, Fusarium verticillioides, and Fusarium graminearum* are our only sampled Fusarium species not present in our gene trees (See Fig. 3-S3), thus if the IPMDH-IPMI fusion was transferred into the Fusarium ancestor as the abundance of Fusarium species with the fusion and their species-tree like grouping suggests, these may also represent complete losses. However, given phylogenetic uncertainty and potential genome incompleteness it is more challenging to be confident about these events.

3.4.7 Homoplasies in complex gene evolution

The history documented here includes multiple cases of secondary domain loss and complete gene loss, as well as recurrent lateral gene transfer, underscoring the remarkably complex history of these core enzymatic functions. How many times has the gene fusion occurred? If the gene trees are to be believed, the parsimonious explanation could be to infer multiple cases of fusion leading separately to the Z+M, Blastocladiomycota and Chytridiomyceta clades. However, given the possibility of phylogenetic errors, we again believe caution is in order. The initial origins of the Pezizomycotina gene fusion is particularly curious. This clade groups deeply within the tree, suggesting an ancient origin for this clade, yet the fusion is found only within small recent groups (including some single species) (Fig. 3-3). This anomaly could be resolved if functional changes, changes in the cellular environment of Pezizomycotina, or other factors have led to an increase in

evolutionary rate of this clade, in which case long branch attraction could have led to misplacement of this clade. If so, this clade could have emerged from within the Z+M (or another) clade; alternatively, this clade could represent an independent fusion. Another interesting case involves the Chytridiomycota, in which subclades of fused and unfused genes group as sister, while Monoblepharidomycota is absent entirely (Fig. 3-3, Fig. 3-2). This pattern could be explained either by a novel fusion or by secondary fission and domain loss (depending on the ancestral state, which is difficult to infer given the weak bootstrap support for the deepest branches of the tree). In general, then, we do not believe that the current data provide clear evidence as to the number of gene fusion events responsible for the observed diversity.

3.4.8 *Reconstruction of the evolutionary history of IPMI and IPMDH genes in fungi*

In total, then, these results suggest a complex history, as depicted in Figure 3-5 below. This history involves dozens of unexpected events, including at least: (i) one or multiple early IPMDH-IPMI fusions; (ii) recurrent loss of the ancestral fused gene, either by complete loss or by domain loss; (iii) many events of HGT.

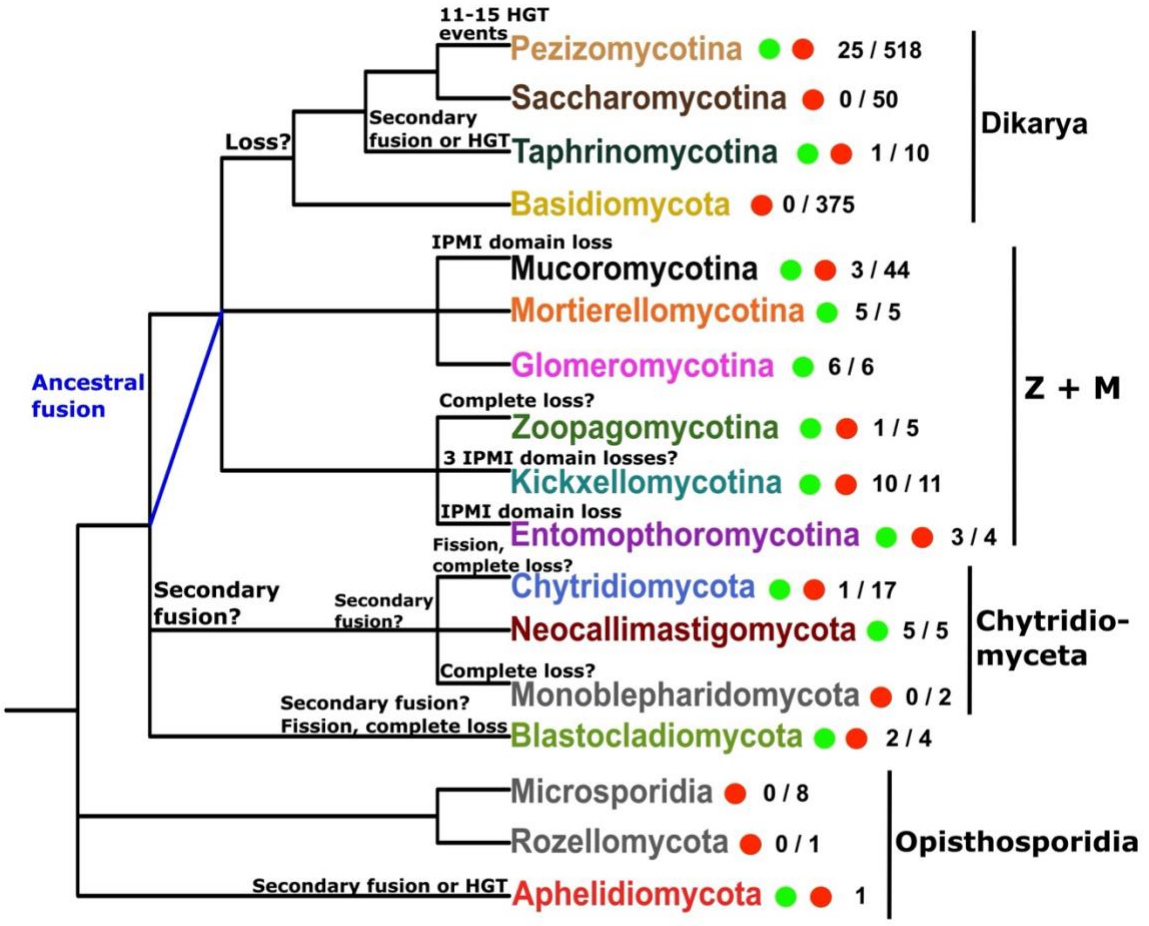


Figure 3-5. Summary of reconstructed evolutionary events involving the IPMDH-

IPMI fusion across fungi and outgroups. HGT indicates horizontal gene transfer. “IPMI domain loss” indicates presence of a IPMDH gene related to fusions but lacking a fused IPMI, indicating gene fission and loss of the IPMI domain. Question marks indicate events with incomplete evidence (e.g., possible gene misannotation), or where different potential histories are plausible. In particular, the pattern observed in the Chytridiomyceta clade could by secondary fusion events in Chytridiomycota and Neocallimastigomycota, or alternatively by a retention of the ancestral fused gene or a single secondary fusion event, in either case followed loss in Monoblepharidomycota. The blue line indicates ambiguity as to the timing of the ancestral fusion event. As in Figure 3-2, green and red circles indicate presence and absence of gene fusion within each group, and numbers for fusions out of the searched JGI database are included as in Figure 3-3.

3.4.9 An independent IPMDH-IPMI fusion event in green algae

The recurrent acquisition of IPMDH-IPMI gene fusions in fungi raises the question as to whether such fusions are common across the tree of life. A search for additional annotated cases of such fusions through BlastP searches of IPMI and IPMDH domains for all of GenBank yielded no results outside of fungi beyond the single aforementioned aphelid, suggesting that IPMDH-IPMI fusions are much more common in fungi than other groups. However, an additional search of the predicted proteomes of 670 diverse eukaryotes represented within the MMETSP database (Keeling et al. 2014) revealed a single additional case in green algae. The species *Chloroparvula japonica* (MMETSP1310/RC2339) shows a very similar fusion to those observed in fungi, with the IPMDH domain upstream of the IPMI domain with a much shorter linker region between IPMDH and IPMI of 4 amino acids, sequence below:

>CAMPEP_0119120364:9-1048
KIVVLPGDGIGPEIVEPAVSVLETVARARGHEFVFEEHLFGAAAIDAVGKPLPDTTLAAC
KSAEAALLGAVGHPKYDGKPVRPEQGLLGLRKELGLYANLRPVSILSSSLSERTSPIRPE
RLEGTDFIVVRELTGGIYFGQRQEANEKGEAWDRLEYSEKEIERIIRIAAQTARSRRKLV
TSVDKNNVLATSRLWRRTAERVMREEFPDVQLEHMLVDAAAMHMLREPRHFDVMVTENMF
GDILTDEASMLPGSLGLLPSASLSESGFGLYEPIHGSAPDIAGQGVANPMGTVLSAAMLL
RHSLGLEEEARAVEAACRDVLDAGIHTPDIAVAGGRAATTREVGNAVLIRLQDLLRESPA
TLYDKIFNEHVVRDCGDGTITMFVDRHLVHEVTSPQAFEGLRANGRAVRRRDCTLATVDH
NVPTSSRAKYGGLAKYITEPESLNQCSALEQNVREFGVSYYGLGDRRQGIVHVIGPEQGF
TLPGTTVVCGDSHTATHGAFGALAFGIGTSEVEHVLATQTLPQSRARNMRVTINGSLPSG
VTSKDLMLHIIGVVGTAGGTNHTIEFAGKAIEEMSMEARMSICNMAIEAGARAGLVAPDE
VTFEYLKDRPMAPTGKEWEAAVEHWKQLRTDEGAVFDKEVEINARDIAPTVTWGTSPEDV
LPITGRVPDPAEEKDATKRAAMERALQYMGLTAGVALTDVPVDKVFIGSCTNSRIEDLRA
VAAVAAGHKVADGVHAMIVPGSGVVKEQAEREGLVKIFEDAGFDWREPGCSMCLGMNPDQ
LKPQERCASTSNRNFEGRQGAGGRTHLMSPAMAAAAAVTGCLADVRQLERKEHIAPSAMP
DATAKQLSRPVGESGAGFVEAPPAAAVPRQAGGGAGGAGGVAKVDVLRGALAPLDRVNVD
TDMIIPKQFLKTVQRSGLGKSAFYELRYNADGTERDSFVLNQPQYRGAPVLVTGANFGCG

SSREHAPWALLDCGVRCIIAESFADIFYNNCFKNGILLVTLPREQVAKLMADAAAGCEVE
VNLGEQYVQTGDGSRYKFEV

Interestingly, in contrast to the fusions observed in fungi, in this instance both domains involved in the fusion show clear similarity to unfused copies in related species (data not shown), indicating a simpler fusion occurring within these species.

3.5 Discussion

Here we report a complex history of the leucine biosynthesis enzymes IPMDH and IPMI across fungi, including recurrent gene fusion, gene fission, and lateral gene transfer of fused genes (Fig. 3-5). In addition, the loss from the genome of ancestral enzyme-coding genes (since one of the domains/genes is generally lost in the cases of fission) suggests functional replacement of ancestral genes by paralogs or by lateral transfers, particularly given the core functions encoded by the studied genes. This history represents, to our knowledge, among the most complex history of a single pair of genes to be reported, the most comparable being a similar story of gene fission, fusion and loss in ATP citrate lyase enzymes across eukaryotes (Gawryluk et al. 2015).

Several observations from literature on enzyme evolution, gene fusion, fission and lateral gene transfer collectively render aspects of the history reported here at once more and less surprising than it may initially seem. First, while lateral gene transfer was initially thought to be rare in eukaryotes, many studies have now shown that lateral gene transfers are not uncommon events in some eukaryotic lineages (Van Etten and Bhattacharya 2020; Gabaldón 2020). In particular, the phylogenetic concentration of HGT events observed in this study within Pezizomycotina is consistent with previous results indicating elevated rates in this group (Marcet-Houben and Gabaldón 2010). It has been demonstrated that gene fusions are transferred laterally more frequently than they occur independently or than they are inherited vertically with subsequent fission (Yanai et al. 2002). Moreover, metabolic enzymes have been shown to be frequently laterally transferred, consistent with the current study (Andersson 2005; Schönknecht et al. 2014; Van Etten and Bhattacharya 2020). On the other hand, work on HGT of metabolic enzymes has tended to concentrate on nonessential enzymes, in particular in cases of concerted transfer of entire metabolic pathways responsible for synthesizing or degrading particular compounds that often produce secondary metabolites (Wisecaver and Rokas 2015; Rokas et al. 2020). By contrast, here we describe recurrent transfer of genes involved in the essential process of leucine biosynthesis, in clear contrast to the paradigm of HGT of auxiliary enzymes. The essentiality of the reactions catalyzed also makes the finding of recurrent loss of ancestral enzymes more surprising. Presumably these species still synthesize leucine, however, what enzymes perform the function of those lost remains obscure. However, some previous work suggests that the specific reactions

studied here may be more prone to functional replacement, as in other fungi it has been found that homologous enzymes can take up this function, suggesting a route to gene loss in these species (Larson and Idnurm 2010; Aguirre-López et al. 2020).

The pattern of recurrent fusion and fission is also challenging to interpret. Fusion between adjacent enzymes in a metabolic pathway immediately suggests the possibility of optimization of the pathway due to more efficient intermediate transfers, perhaps by substrate channeling, or by coordination of gene expression (Tsoka and Ouzounis 2000; Henry et al. 2016; Hagel and Facchini 2017). Indeed, a specific fusion of IPMDH and IPMI is supported by previous work. Experimental evidence suggests that physical complexing between IPMDH and IPMI *in vivo* may act to regulate substrate channeling (Chen et al. 2021). This association may also be analogous to the metabolon complex of the TCA cycle (see Bulutoglu et al. 2016) where aconitase associates with citrate synthase and malate dehydrogenase and exhibits substrate channeling. Thus, fusion of the two genes encoding for the partner enzymes can lead to a structure with improved functions.

However, here there are several reasons for caution. The first is that we observe both fusion and “un-fusion”, whether by domain loss or by entire gene loss and presumably replacement by non-fused alternatives: such recurrent reversal through evolution is not expected for a generally adaptive change. Secondly, substrate channeling may be a more major force in prokaryotes over eukaryotes, though may also simply be better studied in prokaryotes. In the absence of direct biochemical evidence, this complexity generally urges caution in interpreting the evolutionary history, caution which certainly applies to interpretations of the gene fusions.

3.5.1 Robustness of core findings to methodological limitations

The results presented here rely on genome annotation and phylogenetic tree building, both of which are notoriously challenging and prone to error, thus caution is in order. However, while the specifics of the reconstruction of the evolutionary history of these genes could be altered by methodological issues, we believe that our most important broad conclusions are robust to these concerns. The conclusion of an IPMDH-IPMI fusion in early-branching fungi depends only on the concordance between the gene tree and species tree, which is not expected based on errors in gene annotation (since failure to annotate fused genes in some species would tend to cause fusions to appear more recent, not more ancient) or in phylogenetic reconstruction (which should tend to destroy rather than create concordance of gene trees with species trees). While gene annotation errors could indeed lead to a conclusion of loss of the gene fusion, the fact that the gene fusion is not annotated in large groups of species is not expected by stochastic gene annotation errors. Indeed, gene annotation errors would be most likely to manifest as absence of the fusion from a single species, which we see in 3 species discussed above with additional evidence of potential misannotation, and one

singular loss in *B. meristoporus* that appears to be a genuine fission and loss. Recurrent lateral transfer of a fused gene between Pezizomycotina species is supported by the highly punctate phylogenetic pattern of presence of the fusion; by the non-species relationship reconstructed for the IPMI domain; and by a highly concordant non-species relationship reconstructed for the IPMDH domain (Fig. 3-3, Fig. 3-4). That secondary fusions have occurred is independently supported by the grouping of both fused domains outside of pan-fungal fusion clade for the same group of species (e.g., Pezizomycotina).

3.5.2 Concluding remarks

These results, along with cumulative evidence from other studies, highlights the limitations of the paradigm of genes encoding core functions in eukaryotes evolving conservatively and by strictly vertical inheritance. Genes with fused and non-fused versions may be particularly useful in illuminating complex histories, since it can help to distinguish true HGTs from errors of phylogenetic reconstruction. It will be interesting to learn whether these results truly represent an extreme example or rather a typical example of an overlooked subset of eukaryotic genes with complex histories.

3.6 Supplemental figure legends

Supplemental Figure 3-S1. IPMDH gene tree in full. As opposed to main text Figure 3-3, clades are not collapsed here by either congeners or by low bootstrap values. NCBI accession numbers and JGI protein IDs with taxa strain names are marked on the tips. Taxonomic groups are color coded and the Z+M clades (Zoopagomycota and Mucoromycota) and Pezizomycotina HGT fusion clade are labelled. The presence of IPMDH-IPMI fusions in a clade is marked by a green bar, but note that here this does not indicate presence of the fusion in all members within, see main text Figure 3-3 for more detail.

Supplemental Figure 3-S2. IPMI gene tree in full. As opposed to main text Figure 3-3, clades are not collapsed here by either congeners or by low bootstrap values. NCBI accession numbers and JGI protein IDs with taxa strain names are marked on the tips. Taxonomic groups are color coded and the Z+M clades (Zoopagomycota and Mucoromycota) and Pezizomycotina HGT fusion clade are labelled. The presence of IPMDH-IPMI fusions in a clade is marked by a green bar, but note that here this does not indicate presence of the fusion in all members within, see main text Figure 3-3 for more detail.

Supplemental Figure 3-S3. Pezizomycotina species tree phylogeny of all fungal species within the searched JGI database at time of download (December 2018), and 7 fusion species only found in NCBI have been added. Green circles on tips represent presence of the IPMDH-IPMI fusion in or near the species marked. Tree was generated by TimeTree (http://www.timetree.org/) and not all species names

were located, thus some taxa of interest were replaced by congeners or alternative names: *Pseudogymnoascus sp. => Pseudogymnoascus destructans*; *Neonectria radicicola* => Neonectria dissimita; Cylindrocarpon olidum => lyonectria destructans; Acremonium strictum => *Saroctadium strictum*; *Aspergillus thermomutatus => Aspergillus clavatus*; *Penicillium glabrum => Penicillium decumbens*; In the case of *Lophiostoma macrostomum*, its closest relatives in the tree are the *Corynespora cassiicola* clade in the TimeTree, but *L. macrostomum* is not within that clade and would require at least 7 losses according to the JGI's Mycocosm species trees, were the fusions in these two relatives ancestrally shared. In cases where taxa names could not be found at all by TimeTree, green circles are indicated near their closest relatives. Notably not all relationships here mimic most recent phylogenies, and in some cases species with fusions appear closer related than they are based on more updated evidence. Even so, from this tree we can estimate at least 8 horizontal transfer events, while comparing gene trees more complete from JGI's Mycocosm allows us to estimate 11-15 HGT events within Pezizomycotina.

References

Abbott JK, Nordén AK, Hansson B. 2017. Sex chromosome evolution: historical insights and future perspectives. *Proceedings of the Royal Society B: Biological Sciences* [Internet] 284:20162806. Available from: http://rspb.royalsocietypublishing.org/lookup/doi/10.1098/rspb.2016.2806

Aggarwal R, Benatti TR, Gill N, Zhao C, Chen M-S, Fellers JP, Schemerhorn BJ, Stuart JJ. 2009. A BAC-based physical map of the Hessian fly genome anchored to polytene chromosomes. *BMC Genomics* 10:293.

Aguirre-López B, Escalera-Fanjul X, Hersch-González J, Rojas-Ortega E, El-Hafidi M, Lezama M, González J, Bianchi MM, López G, Márquez D, et al. 2020. In Kluyveromyces lactis a Pair of Paralogous Isozymes Catalyze the First Committed Step of Leucine Biosynthesis in Either the Mitochondria or the Cytosol. *Front Microbiol* 11:1843.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* [Internet] 215:403–410. Available from: https://pubmed.ncbi.nlm.nih.gov/2231712/

Andersson JO. 2005. Lateral gene transfer in eukaryotes. *Cellular and Molecular Life Sciences* [Internet] 62:1182–1197. Available from: https://link.springer.com/article/10.1007/s00018-005-4539-z

Asar Ö, Ilk O, Dag O. 2017. Estimating Box-Cox power transformation parameter via goodness-of-fit tests. *Commun Stat Simul Comput* 46:91–105.

Ashman TL, Bachtrog D, Blackmon H, Goldberg EE, Hahn MW, Kirkpatrick M, Kitano J, Mank JE, Mayrose I, Ming R, et al. 2014. Tree of Sex: A database of sexual systems. *Sci Data* 1:140015.

Bachtrog D. 2020. The Y Chromosome as a Battleground for Intragenomic Conflict. *Trends in Genetics* 36:510–522.

Baines JF, Harr B. 2007. Reduced X-Linked Diversity in Derived Populations of House Mice. *Genetics* [Internet] 175:1911. Available from: /pmc/articles/PMC1855135/

Baines JF, Sawyer SA, Hartl DL, Parsch J. 2008. Effects of X-linkage and sex-biased gene expression on the rate of adaptive protein evolution in Drosophila. *Mol Biol Evol* [Internet] 25:1639–1650. Available from: https://pubmed.ncbi.nlm.nih.gov/18477586/

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19:455–477.

Bertone MA, Brian M. Wiegmann. 2009. True flies (Diptera). In: Hedges SB, Kumar S, editors. The Timetree of Life. Oxford University Press. p. 270–277.

Blackmon H, Hardy NB, Ross L. 2015. The evolutionary dynamics of haplodiploidy: Genome architecture and haploid viability. *Evolution (N Y)* 69:2971–2978.

Brisson JA, Nuzhdin S V., Stern DL. 2009. Similar patterns of linkage disequilibrium and nucleotide diversity in native and introduced populations of the pea aphid, Acyrthosiphon pisum. *BMC Genet* [Internet] 10:1–10. Available from: https://bmcgenomdata.biomedcentral.com/articles/10.1186/1471-2156-10-22

Brown SW. 1964. Automatic Frequency Response in the Evolution of Male Haploidy and Other Coccid Chromosome Systems. *Genetics* 49:797–817.

Bull JJ. 1979. An advantage for the evolution of male haploidy and systems with similar genetic transmission. *Heredity (Edinb)* 43:361–381.

Bulutoglu B, Garcia KE, Wu F, Minteer SD, Banta S. 2016. Direct Evidence for Metabolon Formation and Substrate Channeling in Recombinant TCA Cycle Enzymes. *ACS Chem Biol* [Internet] 11:2847–2853. Available from: https://pubs.acs.org/doi/full/10.1021/acschembio.6b00523

Burt A, Trivers R. 2006. Genes in Conflict: The Biology of Selfish Genetic Elements.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* [Internet] 10:1–9. Available from: http://dx.doi.org/10.1186/1471-2105-10-421

Castellana M, Wilson MZ, Xu Y, Joshi P, Cristea IM, Rabinowitz JD, Gitai Z, Wingreen NS. 2014. Enzyme clustering accelerates processing of intermediates through metabolic channeling. *Nature Biotechnology 2014 32:10* [Internet] 32:1011–1018. Available from: https://www.nature.com/articles/nbt.3018

Charlesworth B, Campos JL, Jackson BC. 2018. Faster-X evolution: Theory and evidence from Drosophila. *Mol Ecol* [Internet] 27:3753–3771. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/mec.14534

Charlesworth B, Coyne JA, Barton NH. 1987. The Relative Rates of Evolution of Sex Chromosomes and Autosomes. *Am Nat* 130:113–146.

Charlesworth D. 2021. The timing of genetic degeneration of sex chromosomes. *Philosophical Transactions of the Royal Society B: Biological Sciences* [Internet] 376. Available from: /pmc/articles/PMC8273506/

Chen LQ, Chhajed S, Zhang T, Collins JM, Pang Q, Song W, He Y, Chen S. 2021. Protein complex formation in methionine chain-elongation and leucine biosynthesis. *Scientific Reports 2021 11:1* [Internet] 11:1–8. Available from: https://www.nature.com/articles/s41598-021-82790-4

Connallon T, Hall MD. 2018. Genetic constraints on adaptation: a theoretical primer for the genomics era. *Ann N Y Acad Sci* [Internet] 1422:65–87. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/nyas.13536

Connallon T, Knowles LL. 2005. Intergenomic conflict revealed by patterns of sex-biased gene expression. *Trends in Genetics* [Internet] 21:495–499. Available from: http://www.cell.com/article/S0168952505001964/fulltext

Cote-L'Heureux A, Maurer-Alcalá XX, Katz LA. 2022. Old genes in new places: A taxon-rich analysis of interdomain lateral gene transfer events. *PLoS Genet* [Internet] 18:e1010239. Available from:

https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1010239

Dallai R, Fanciulli PP, Frati F. 2000. Aberrant spermatogenesis and the peculiar mechanism of sex determination in symphypleonan Collembola. *Journal of Heredity* 91:351–358.

Darolti I, Fong LJM, Sandkam BA, Metzger DCH, Mank JE. 2023. Sex chromosome heteromorphism and the Fast-X effect in poeciliids. *Mol Ecol* [Internet]. Available from: https://pubmed.ncbi.nlm.nih.gov/37309716/

Davis GK. 2012. Cyclical Parthenogenesis and Viviparity in Aphids as Evolutionary Novelties. *J Exp Zool B Mol Dev Evol* [Internet] 318:448–459. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/jez.b.22441

Dean R, Harrison PW, Wright AE, Zimmer F, Mank JE. 2015. Positive Selection Underlies Faster-Z Evolution of Gene Expression in Birds. *Mol Biol Evol* [Internet] 32:2646–2656. Available from: https://dx.doi.org/10.1093/molbev/msv138

Elcock AH, Andrew McCammon J. 1996. Evidence for Electrostatic Channeling in a Fusion Protein of Malate Dehydrogenase and Citrate Synthase†. *Biochemistry* [Internet] 35:12652–12658. Available from: https://pubs.acs.org/doi/full/10.1021/bi9614747

Ellegren H. 2007. Characteristics, causes and evolutionary consequences of male-biased mutation. *Proc Biol Sci* [Internet] 274:1–10. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17134994

Ellegren H. 2009. The different levels of genetic diversity in sex chromosomes and autosomes. *Trends Genet* [Internet] 25:278–284. Available from: https://pubmed.ncbi.nlm.nih.gov/19481288/

Ellegren H. 2011. Sex-chromosome evolution: recent progress and the influence of male and female heterogamety. *Genetics* [Internet] 12:157–166. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21301475

Van Etten J, Bhattacharya D. 2020. Horizontal Gene Transfer in Eukaryotes: Not if, but How Much? *Trends in Genetics* 36:915–925.

Felsenstein J. 2005. Using the quantitative genetic threshold model for inferences between and within species. *Philos Trans R Soc Lond B Biol Sci* 360:1427–1434.

Fitzpatrick DA. 2012. Horizontal gene transfer in fungi. *FEMS Microbiol Lett* [Internet] 329:1–8. Available from: http://femsle.oxfordjournals.org/content/329/1/1.abstract

Gabaldón T. 2020. Patterns and impacts of nonvertical evolution in eukaryotes: a paradigm shift. *Ann N Y Acad Sci* [Internet] 1476:78–92. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/nyas.14471

Gallun RL, Hatchett JH. 1969. Genetic Evidence of Elimination of Chromosomes in the Hessian Fly. *Ann Entomol Soc Am* 62:1095–1101.

Galtier N, Roux C, Rousselle M, Romiguier J, Figuet E, Glemin S, Bierne N, Duret L. 2018. Codon Usage Bias in Animals: Disentangling the Effects of Natural Selection, Effective Population Size, and GC-Biased Gene

Conversion. *Mol Biol Evol* [Internet] 35:1092–1103. Available from: https://dx.doi.org/10.1093/molbev/msy015
Gardner A, Ross L. 2014. Mating ecology explains patterns of genome elimination. *Ecol Lett* 17:1602–1612.
Gawryluk RMR, Eme L, Roger AJ. 2015. Gene fusion, fission, lateral transfer, and loss: Not-so-rare events in the evolution of eukaryotic ATP citrate lyase. *Mol Phylogenet Evol* [Internet] 91:12–16. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1055790315001475
Geer LY, Domrachev M, Lipman DJ, Bryant SH. 2002. CDART: Protein Homology by Domain Architecture. *Genome Res* [Internet] 12:1619–1623. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12368255
Goday C, Rosario Esteban M. 2001. Chromosome elimination in sciarid flies. *BioEssays* 23:242–250.
Goldstein DB. 1994. Deleterious mutations and the evolution of male haploidy. *American Society of Naturalists* 144:176–183.
Grafen A. 1989. The phylogenetic regression. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 326:119–157.
Grath S, Parsch J. 2012. Rate of amino acid substitution is influenced by the degree and conservation of male-biased transcription over 50 myr of Drosophila evolution. *Genome Biol Evol* [Internet] 4:346–359. Available from: https://pubmed.ncbi.nlm.nih.gov/22321769/
Grath S, Parsch J. 2016. Sex-Biased Gene Expression. *Annu Rev Genet* [Internet] 50:29–44. Available from: http://www.annualreviews.org/doi/10.1146/annurev-genet-120215-035429
Grigoriev I V., Cullen D, Goodwin SB, Hibbett D, Jeffries TW, Kubicek CP, Kuske C, Magnuson JK, Martin F, Spatafora JW, et al. 2011. Fueling the future with fungal genomics. *Mycology* [Internet] 2:192–209. Available from: https://www.tandfonline.com/action/journalInformation?journalCode=tmyc20
Grigoriev I V., Nikitin R, Haridas S, Kuo A, Ohm R, Otillar R, Riley R, Salamov A, Zhao X, Korzeniewski F, et al. 2014. MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res* [Internet] 42:D699–D704. Available from: https://academic.oup.com/nar/article/42/D1/D699/1058210
Haddrill PR, Loewe L, Charlesworth B. 2010. Estimating the Parameters of Selection on Nonsynonymous Mutations in Drosophila pseudoobscura and D. miranda. *Genetics* [Internet] 185:1381–1396. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20516497
Hadfield JD. 2015. Increasing the efficiency of MCMC for hierarchical phylogenetic models of categorical traits using reduced mixed models. *Methods Ecol Evol* 6:706–714.
Hadfield JD, Nakagawa S. 2010. General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *J Evol Biol* 23:494–508.
Hagel JM, Facchini PJ. 2017. Tying the knot: occurrence and possible significance of gene fusions in plant metabolism and beyond. *J Exp Bot*

[Internet] 68:4029–4043. Available from: https://academic.oup.com/jxb/article/68/15/4029/3832250

Haig D. 1993a. The evolution of unusual chromosomal systems in coccoids: extraordinary sex ratios revisited. *J Evol Biol* [Internet] 6:69–77. Available from: http://doi.wiley.com/10.1046/j.1420-9101.1993.6010069.x

Haig D. 1993b. The evolution of unusual chromosomal systems in sciarid flies: intragenomic conflict and the sex ratio. *J Evol Biol* 6:249–261.

Haldane JBS. 1924. A mathematical theory of natural and artificial selection-I. *Bull Math Biol* [Internet] 52:209–240. Available from: http://link.springer.com/10.1007/BF02459574

Hamilton WD. 1967. Extraordinary sex ratios. *Science (1979)* 156:477–488.

Hartl DL, Brown SW. 1970. The origin of male haploid genetic systems and their expected sex ratio. *Theor Popul Biol* 1:165–190.

Hayes K, Barton HJ, Zeng K, Slotte T. 2020. A Study of Faster-Z Evolution in the Great Tit (Parus major). *Genome Biol Evol* [Internet] 12:210–222. Available from: https://dx.doi.org/10.1093/gbe/evaa044

Hebert PDN, Ratnasingham S, Zakharov E V., Telfer AC, Levesque-Beaudin V, Milton MA, Pedersen S, Jannetta P, Dewaard JR. 2016. Counting animal species with DNA barcodes: Canadian insects. *Philosophical Transactions of the Royal Society B: Biological Sciences* 371.

Henry CS, Lerma-Ortiz C, Gerdes SY, Mullen JD, Colasanti R, Zhukov A, Frelin O, Thiaville JJ, Zallot R, Niehaus TD, et al. 2016. Systematic identification and analysis of frequent gene fusion events in metabolic pathways. *BMC Genomics* [Internet] 17:1–17. Available from: https://link.springer.com/articles/10.1186/s12864-016-2782-3

Hershberg R, Petrov DA. 2008. Selection on Codon Bias. *Annu Rev Genet* [Internet] 42:287–299. Available from: http://www.annualreviews.org/doi/10.1146/annurev.genet.42.110807.091442

Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. 2019. Whole-Genome Annotation with BRAKER. *Methods Mol Biol* 1962:65–95.

Hvilsom C, Qian Y, Bataillon T, Li Y, Mailund T, Sallé B, Carlsen F, Li R, Zheng H, Jiang T, et al. 2012. Extensive X-linked adaptive evolution in central chimpanzees. *Proc Natl Acad Sci U S A* [Internet] 109:2054–2059. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22308321

Ives AR, Garland Jr. T. 2010. Phylogenetic Logistic Regression for Binary Dependent Variables. *Syst Biol* 59:9–26.

Ives AR, Garland Jr. T. 2014. Phylogenetic Regression for Binary Dependent Variables. In: Garamszegi LZ, editor. Modern Phylogenetic Comparative Methods and their Application in Evolutionary Biology. Berlin Heidelber. p. 231–261.

Jaquiéry J, Peccoud J, Ouisse T, Legeai F, Prunier-Leterme N, Gouin A, Nouhaud P, Brisson JA, Bickel R, Purandare S, et al. 2018. Disentangling the Causes for Faster-X Evolution in Aphids. *Genome Biol Evol* [Internet] 10:507–520. Available from: https://dx.doi.org/10.1093/gbe/evy015

Jaquiéry J, Stoeckel S, Rispe C, Mieuzet L, Legeai F, Simon JC. 2012. Accelerated evolution of sex chromosomes in aphids, an X0 system. *Mol Biol Evol* 29:837–847.

Jaron KS, Hodson CN, Ellers J, Baird SJE, Ross L. 2022. Genomic evidence of paternal genome elimination in the globular springtail Allacma fusca. *Genetics* [Internet] 222. Available from: https://dx.doi.org/10.1093/genetics/iyac117

Jung S-J, Seo Y, Lee K-C, Lee D, Roe J-H. 2015. Essential function of Aco2, a fusion protein of aconitase and mitochondrial ribosomal protein bL21, in mitochondrial translation in fission yeast. *FEBS Lett* [Internet] 589:822–828. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0014579315001076

Katoh K, Misawa K, Kuma KI, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* [Internet] 30:3059. Available from: /pmc/articles/PMC135756/

Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, Armbrust EV, Archibald JM, Bharti AK, Bell CJ, et al. 2014. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biol* [Internet] 12:e1001889. Available from: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001889

Keller Valsecchi CI, Marois E, Basilicata MF, Georgiev P, Akhtar A. 2021. Distinct mechanisms mediate X chromosome dosage compensation in Anopheles and Drosophila. *Life Sci Alliance* 4:e202000996.

Kessler MD, Dean MD. 2014. Effective population size does not predict codon usage bias in mammals. *Ecol Evol* 4:3887–3900.

Kirkpatrick M, Hall DW. 2004. MALE-BIASED MUTATION, SEX LINKAGE, AND THE RATE OF ADAPTIVE EVOLUTION. *Evolution (N Y)* [Internet] 58:437–440. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/j.0014-3820.2004.tb01659.x

Kourtz L, Dillon K, Daughtry S, Madison LL, Peoples O, Snell KD. 2005. A novel thiolase-reductase gene fusion promotes the production of polyhydroxybutyrate in Arabidopsis. *Plant Biotechnol J* 3:435–447.

Kousathanas A, Halligan DL, Keightley PD. 2014. Faster-X adaptive protein evolution in house mice. *Genetics* [Internet] 196:1131–1143. Available from: /pmc/articles/PMC3982706/

Laetsch DR, Blaxter ML. 2017. BlobTools: Interrogation of genome assemblies [version 1; peer review: 2 approved with reservations]. *F1000Res* 6.

Langley CH, Stevens K, Cardeno C, Lee YCG, Schrider DR, Pool JE, Langley SA, Suarez C, Corbett-Detig RB, Kolaczkowski B, et al. 2012. Genomic Variation in Natural Populations of Drosophila melanogaster. *Genetics* [Internet] 192:533–598. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22673804

Larson EM, Idnurm A. 2010. Two Origins for the Gene Encoding α-Isopropylmalate Synthase in Fungi. *PLoS One* [Internet] 5:e11605. Available

from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0011605

Leffler EM, Bullaughey K, Matute DR, Meyer WK, Ségurel L, Venkat A, Andolfatto P, Przeworski M. 2012. Revisiting an Old Riddle: What Determines Genetic Diversity Levels within Species? *PLoS Biol* 10.

Leonard G, Richards TA. 2012. Genome-scale comparative analysis of gene fusions, gene fissions, and the fungal tree of life. *Proc Natl Acad Sci U S A* 109:21402–21407.

Lesnoff M, Lancelot R. 2010. aod: Analysis of Overdispersed Data.

Li D, Luo R, Liu C-M, Leung C-M, Ting H-F, Sadakane K, Yamashita H, Lam T-W. 2016. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* 102:3–11.

Li WH, Yi S, Makova K. 2002. Male-driven evolution. *Curr Opin Genet Dev* 12:650–656.

Li Y, Steenwyk JL, Chang Y, Wang Y, James TY, Stajich JE, Spatafora JW, Groenewald M, Dunn CW, Hittinger CT, et al. 2021. A genome-scale phylogeny of the kingdom Fungi. *Current Biology* 31:1653-1665.e5.

Lindbladh C, Rault M, Hagglund C, Small WC, Mosbach K, Buelow L, Evans C, Srere PA. 1994. Preparation and kinetic characterization of a fusion protein of yeast mitochondrial citrate synthase and malate dehydrogenase. *Biochemistry* [Internet] 33:11692–11698. Available from: http://pubs.acs.org/doi/abs/10.1021/bi00205a004

M. ZE, Christian von M, Ivica L, David T, Mikita S, R. CR, K. CG, Dana T, A. HR, Mani SG, et al. 2002. Comparative Genome and Proteome Analysis of Anopheles gambiae and Drosophila melanogaster. *Science (1979)* 298:149–159.

Mank JE. 2017. Population genetics of sexual conflict in the genomic era. *Nature Reviews Genetics 2017 18:12* [Internet] 18:721–730. Available from: https://www.nature.com/articles/nrg.2017.83

Mank JE, Axelsson E, Ellegren H. 2007. Fast-X on the Z: Rapid evolution of sex-linked genes in birds. *Genome Res* [Internet] 17:618–624. Available from: https://genome.cshlp.org/content/17/5/618.full

Mank JE, Hosken DJ, Wedell N. 2014. Conflict on the sex chromosomes: cause, effect, and complexity. *Cold Spring Harb Perspect Biol* 6:a017715.

Mank JE, Nam K, Ellegren H. 2010. Faster-Z Evolution Is Predominantly Due to Genetic Drift. *Mol Biol Evol* 27:661–670.

Marcet-Houben M, Gabaldón T. 2010. Acquisition of prokaryotic genes by fungal genomes. *Trends in Genetics* 26:5–8.

Márcia Barbosa A, Real R, Muñoz A-R, Brown JA. 2013. New measures for assessing model equilibrium and prediction mismatch in species distribution models. *Divers Distrib* 19:1333–1338.

Marygold SJ, Crosby MA, Goodman JL, Consortium F. 2016. Using FlyBase, a Database of Drosophila Genes and Genomes. *Methods Mol Biol* 1478:1–31.

Meisel RP, Connallon T. 2013. The faster-X effect: Integrating theory and data. *Trends in Genetics* [Internet] 29:537–544. Available from: http://dx.doi.org/10.1016/j.tig.2013.05.009

Meisel RP, Delclos PJ, Wexler JR. 2019. The X chromosome of the German cockroach, Blattella germanica, is homologous to a fly X chromosome despite 400 million years divergence. *BMC Biol* 17:100.

Metz CW. 1938. Chromosome Behavior, Inheritance and Sex Determination in Sciara. *Am Nat* 72:485–520.

Meynial Salles I, Forchhammer N, Croux C, Girbal L, Soucaille P. 2007. Evolution of a Saccharomyces cerevisiae metabolic pathway in Escherichia coli. *Metab Eng* 9:152–159.

Minh BQ, Nguyen MAT, Von Haeseler A. 2013. Ultrafast Approximation for Phylogenetic Bootstrap. *Mol Biol Evol* [Internet] 30:1188–1195. Available from: https://academic.oup.com/mbe/article/30/5/1188/997508

Mongue AJ, Hansen ME, Walters JR. 2022. Support for faster and more adaptive Z chromosome evolution in two divergent lepidopteran lineages. *Evolution (N Y)* [Internet] 76:332–345. Available from: https://dx.doi.org/10.1111/evo.14341

Muller H. 1940. Bearings of the “Drosophila” work on systematics. In: Huxley J, editor. The New Systematics. Oxford: Clarendon Press. p. 185–268.

Müller L, Grath S, von Heckel K, Parsch J. 2012. Inter- and intraspecific variation in Drosophila genes with sex-biased expression. *Int J Evol Biol* [Internet] 2012:1–10. Available from: https://pubmed.ncbi.nlm.nih.gov/22315698/

Naranjo-Ortiz MA, Gabaldón T. 2019. Fungal evolution: major ecological adaptations and evolutionary transitions. *Biological Reviews* 94:1443–1476.

Naranjo-Ortiz MA, Gabaldón T. 2020. Fungal evolution: cellular, genomic and metabolic complexity. *Biological Reviews* [Internet] 95:1198–1232. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/brv.12605

Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol* [Internet] 32:268. Available from: /pmc/articles/PMC4271533/

Normark BB. 2004. Haplodiploidy as an outcome of coevolution between male-killing cytoplasmic elements and their hosts. *Evolution (N Y)* 58:790–798.

Normark BB. 2006. Maternal Kin Groups and the Origins of Asymmetric Genetic Systems: Genomic Imprinting, Haplodiploidy, and Parthenogenesis. *Evolution (N Y)* 60:631–642.

Normark BB, Ross L. 2014. Genetic conflict, kin and the origins of novel genetic systems. *Philosophical Transactions of the Royal Society B: Biological Sciences* 369.

Nützmann H-W, Scazzocchio C, Osbourn A. 2018. Metabolic Gene Clusters in Eukaryotes. *Annu Rev Genet* 52:159–183.

Orr HA. 2010. The population genetics of beneficial mutations. *Philosophical Transactions of the Royal Society B: Biological Sciences* [Internet]

365:1195–1201. Available from: http://rstb.royalsocietypublishing.org/cgi/doi/10.1098/rstb.2009.0282

Orr HA, Betancourt AJ. 2001. Haldane's sieve and adaptation from the standing genetic variation. *Genetics* 157:875–884.

Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35:526–528.

Pettersson H, Olsson P, Bülow L, Pettersson G. 2000. Kinetics of the coupled reaction catalysed by a fusion protein of yeast mitochondrial malate dehydrogenase and citrate synthase. *Eur J Biochem* [Internet] 267:5041–5046. Available from: https://onlinelibrary.wiley.com/doi/full/10.1046/j.1432-1327.2000.01558.x

Pinharanda A, Rousselle M, Martin SH, Hanly JJ, Davey JW, Kumar S, Galtier N, Jiggins CD. 2019. Sexually dimorphic gene expression and transcriptome evolution provide mixed evidence for a fast-Z effect in Heliconius. *J Evol Biol* [Internet] 32:194–204. Available from: https://pubmed.ncbi.nlm.nih.gov/30523653/

Plummer M, Best N, Cowles K, Vines K. 2005. CODA: Convergence diagnosis and output analysis for MCMC. *R News* 6.

Purandare SR, Bickel RD, Jaquiery J, Rispe C, Brisson JA. 2014. Accelerated Evolution of Morph-Biased Genes in Pea Aphids. *Mol Biol Evol* [Internet] 31:2073–2083. Available from: https://dx.doi.org/10.1093/molbev/msu149

Qiu S, Bergero R, Zeng K, Charlesworth D. 2011. Patterns of Codon Usage Bias in Silene latifolia. *Mol Biol Evol* [Internet] 28:771–780. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20855431

Radhakrishnan S, Valenzuela N. 2017. Chromosomal Context Affects the Molecular Evolution of Sex-linked Genes and Their Autosomal Counterparts in Turtles and Other Vertebrates. *Journal of Heredity* [Internet] 108:720–730. Available from: http://academic.oup.com/jhered/article/108/7/720/4259380

Rao Y, Wu G, Wang Z, Chai X, Nie Q, Zhang X. 2011. Mutation bias is the driving force of codon usage in the Gallus gallus genome. *DNA Res* [Internet] 18:499–512. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22039174

Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol Evol* 3:217–223.

Rice WWR. 1984. Sex chromosomes and the evolution of sexual dimorphism. *Evolution (N Y)* [Internet] 38:1416–1424. Available from: http://apps.webofknowledge.com/full_record.do?product=WOS&search_mode=GeneralSearch&qid=115&SID=N2JBCmJk4Ik9okHfkNb&page=1&doc=7

Richards S, Stuart JJ. BCM-HGSC Hessian Fly Genome Project. *Baylor College of Medicine Human Genome Sequencing Center* [Internet]. Available from: https://www.hgsc.bcm.edu/arthropods/hessian-fly-genome-project

Richards T a., Leonard G, Soanes DM, Talbot NJ. 2011. Gene transfer into the fungi. *Fungal Biol Rev* [Internet] 25:98–110. Available from: http://dx.doi.org/10.1016/j.fbr.2011.04.003

Richards TA. 2011. Genome Evolution: Horizontal Movements in the Fungi. *Current Biology* 21:R166–R168.

Richards TA, Talbot NJ. 2013. Horizontal gene transfer in osmotrophs: Playing with public goods. *Nat Rev Microbiol* 11:720–727.

Rokas A, Mead ME, Steenwyk JL, Raja HA, Oberlies NH. 2020. Biosynthetic gene clusters and the evolution of fungal chemodiversity. *Nat Prod Rep* 37:868–878.

Rokas A, Wisecaver JH, Lind AL. 2018. The birth, evolution and death of metabolic gene clusters in fungi. *Nature Reviews Microbiology 2018 16:12* [Internet] 16:731–744. Available from: https://www.nature.com/articles/s41579-018-0075-3

Ross L, Pen I, Shuker DM. 2010. Genomic conflict in scale insects: The causes and consequences of Bizarre genetic systems. *Biological Reviews* 85:807–828.

Rousselle M, Faivre N, Ballenghien M, Galtier N, Nabholz B. 2016. Hemizygosity Enhances Purifying Selection: Lack of Fast-Z Evolution in Two Satyrine Butterflies. *Genome Biol Evol* [Internet] 8:3108–3119. Available from: https://dx.doi.org/10.1093/gbe/evw214

Sackton TB, Corbett-Detig RB, Nagaraju J, Vaishna L, Arunkumar KP, Hartl DL. 2014. Positive selection drives faster-Z evolution in silkmoths. *Evolution (N Y)* 68:2331–2342.

Sánchez L. 2014. Sex-determining mechanisms in insects based on imprinting and elimination of chromosomes. *Sexual Development* 8:83–103.

Schaeffer SW, Bhutkar A, McAllister BF, Matsuda M, Matzkin LM, O'Grady PM, Rohde C, Valente VLS, Aguadé M, Anderson WW, et al. 2008. Polytene Chromosomal Maps of 11 Drosophila Species: The Order of Genomic Scaffolds Inferred From Genetic and Physical Maps. *Genetics* 179:1601–1655.

Schaffner SF. 2004. The X chromosome in population genetics. *Nat Rev Genet* [Internet] 5:43–51. Available from: http://dx.doi.org/10.1038/nrg1247

Schönknecht G, Weber APM, Lercher MJ. 2014. Horizontal gene acquisitions by eukaryotes as drivers of adaptive evolution. *BioEssays* [Internet] 36:9–20. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/bies.201300095

Ševčík J, Hippa H, Burdíková N. 2022. Just a fragment of undescribed diversity: Twenty new oriental and palearctic species of sciaroidea (diptera), including dna sequence data and two new fossil genera. *Insects* 13.

Ševčík J, Kaspřák D, Mantič M, Fitzgerald S, Ševčíková T, Tóthová A, Jaschhof M. 2016. Molecular phylogeny of the megadiverse insect infraorder Bibionomorpha sensu lato (Diptera). *PeerJ* [Internet] 4:e2563. Available from: https://peerj.com/articles/2563

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* .

Singh ND, Davis JC, Petrov DA. 2005. X-linked genes evolve higher codon bias in Drosophila and Caenorhabditis. *Genetics* [Internet] 171:145–155. Available from: http://www.genetics.org/cgi/doi/10.1534/genetics.105.043497

Slot JC. 2017. Fungal Gene Cluster Diversity and Evolution. *Adv Genet* [Internet] 100:141–178. Available from: https://pubmed.ncbi.nlm.nih.gov/29153399/

Soucy SM, Huang J, Gogarten JP. 2015. Horizontal gene transfer: building the web of life. *Nature Reviews Genetics 2015 16:8* [Internet] 16:472–482. Available from: https://www.nature.com/articles/nrg3962

de Souza Amorim D, Brown B V., Boscolo D, Ale-Rocha R, Alvarez-Garcia DM, Balbi MIPA, de Marco Barbosa A, Capellari RS, de Carvalho CJB, Couri MS, et al. 2022. Vertical stratification of insect abundance and species richness in an Amazonian tropical forest. *Sci Rep* 12:1–10.

Stuart JJ, Hatchett JH. 1988a. Cytogenetics of the hessian fly: I. Mitotic karyotype analysis and polytene chromosome correlations. *Journal of Heredity* 79:184–189.

Stuart JJ, Hatchett JH. 1988b. Cytogenetics of the Hessian Fly: II. Inheritance and Behavior of Somatic and Germ-Line-Limited Chromosomes. *Journal of Heredity* 79:190–199.

Sundström H, Webster MT, Ellegren H. 2004. Reduced variation on the chicken Z chromosome. *Genetics* [Internet] 167:377. Available from: /pmc/articles/PMC1470867/?report=abstract

Sved JA, Chen Y, Shearman D, Frommer M, Gilchrist AS, Sherwin WB. 2016. Extraordinary conservation of entire chromosomes in insects over long evolutionary periods. *Evolution (N Y)* 70:229–234.

Sweetlove LJ, Fernie AR. 2018. The role of dynamic enzyme assemblies and substrate channelling in metabolic regulation. *Nature Communications 2018 9:1* [Internet] 9:1–12. Available from: https://www.nature.com/articles/s41467-018-04543-8

Torgerson DG, Singh RS. 2005. Enhanced adaptive evolution of sperm-expressed genes on the mammalian X chromosome. *Heredity 2006 96:1* [Internet] 96:39–44. Available from: https://www.nature.com/articles/6800749

Tsoka S, Ouzounis CA. 2000. Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion. *Nature Genetics 2000 26:2* [Internet] 26:141–142. Available from: https://www.nature.com/articles/ng1000_141

Tung Ho L si, Ané C. 2014. A Linear-Time Algorithm for Gaussian and Non-Gaussian Trait Evolution Models. *Syst Biol* 63:397–408.

Vicoso B, Bachtrog D. 2013. Reversal of an ancient sex chromosome to an autosome in Drosophila. *Nature* 18:1199–1216.

Vicoso B, Bachtrog D. 2015. Numerous Transitions of Sex Chromosomes in Diptera. *PLoS Biol* 13.

Vicoso B, Charlesworth B. 2006. Evolution on the X chromosome: unusual patterns and processes. *Nat Rev Genet* 7:645–653.

Vicoso B, Charlesworth B. 2009. Effective population size and the faster-X effect: An extended model. *Evolution (N Y)* 63:2413–2426.

Watkinson SC. 2016. Physiology and Adaptation. *The Fungi: Third Edition*:141–187.

Werren JH, Beukeboom LW. 1998. Sex determination, sex ratios, and genetic conflict. *Ann.Rev.Ecol.Syst.* 29:233–261.

White MJD (Michael JD. 1950. Cytological studies on gall midges (Cecidomyidae). Austin: The University

Wiley JF. 2013. postMCMCglmm: Average marginal predicted probabilities from Bayesian ordered probit models. R package version 0.1-2.

Wilson ACC, Sunnucks P, Hales DF. 1997. SHORT PAPER Random loss of X chromosome at male determination in an aphid, Sitobion near fragariae, detected using an X-linked polymorphic microsatellite marker. *Genet Res (Camb)* [Internet] 69:233–236. Available from: https://www.cambridge.org/core/journals/genetics-research/article/short-paper-random-loss-of-x-chromosome-at-male-determination-in-an-aphid-sitobion-near-fragariae-detected-using-an-xlinked-polymorphic-microsatellite-marker/EA36BDC08EAFF466A0C0A6D76B4AC536

Wilson Sayres MA. 2018. Genetic diversity on the sex chromosomes. *Mol Biol Evol* [Internet]. Available from: https://academic.oup.com/gbe/advance-article/doi/10.1093/gbe/evy039/4895090

Wisecaver JH, Rokas A. 2015. Fungal metabolic gene clusters-caravans traveling across genomes and environments. *Front Microbiol* [Internet] 6:1–11. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25784900

Wisecaver JH, Slot JC, Rokas A. 2014. The Evolution of Fungal Metabolic Pathways. *PLoS Genet* 10.

Wright AE, Fumagalli M, Cooney CR, Bloch NI, Vieira FG, Buechel SD, Kolm N, Mank JE. 2018. Male-biased gene expression resolves sexual conflict through the evolution of sex-specific genetic architecture. *Evol Lett* [Internet] 2:52–61. Available from: https://pubmed.ncbi.nlm.nih.gov/30283664/

Wright AE, Harrison PW, Zimmer F, Montgomery SH, Pointer MA, Mank JE. 2015. Variation in promiscuity and sexual selection drives avian rate of Faster-Z evolution. *Mol Ecol* [Internet] 24:1218–1235. Available from: http://doi.wiley.com/10.1111/mec.13113

Wright S. 1934. An Analysis of Variability in Number of Digits in an Inbred Strain of Guinea Pigs. *Genetics* 19:506–536.

Yanai I, Wolf YI, Koonin E V. 2002. Evolution of gene fusions: horizontal transfer versus independent events. *Genome Biology 2002 3:5* [Internet] 3:1–13. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/gb-2002-3-5-research0024

Zhaxybayeva O, Doolittle WF. 2011. Lateral gene transfer. *Current Biology* 21:R242–R246.

Zhou Q, Bachtrog D. 2012. Sex-specific adaptation drives early sex chromosome evolution in Drosophila. *Science* [Internet] 337:341. Available from: /pmc/articles/PMC4107656/