Chromosome Rearrangements and Population Genomics

Alexander Mackintosh

THE UNIVERSITY
of EDINBURGH

Submitted for the degree of Doctor of Philosophy

School of Biological Sciences

2023

Abstract

Chromosome rearrangements result in changes to the physical linkage and order of sequences in the genome. Although we have known about these mutations for more than a century, we still lack a detailed understanding of how they become fixed and what their effect is on other evolutionary processes. Analysing genome sequences provides a way to address this knowledge gap. In this thesis I compare genome assemblies and use population genomic inference to gain a better understanding of the role that chromosome rearrangements play in evolution. I focus on butterflies in the genus *Brenthis*, where chromosome numbers are known to vary between species. In chapter 2, I present a genome assembly of *Brenthis ino* and show that its genome has been shaped by many chromosome rearrangements, including a Z-autosome fusion that is still segregating. In chapter 3, I investigate how synteny information in genome sequences can be used to infer ancestral linkage groups and inter-chromosomal rearrangements, implementing the methods in a command-line tool. In chapter 4, I test whether chromosome fissions and fusions have acted as barriers to gene flow between *B. ino* and its sister species *B. daphne*. I find that chromosomes involved in rearrangements have experienced less post-divergence gene flow than the rest of the genome, suggesting that rearrangements have promoted speciation. Finally, in chapter 5, I investigate how chromosome rearrangements have become fixed in *B. ino*, *B. daphne*, and a third species, *B. hecate*. I show that genetic drift is unlikely to be a strong enough force to have fixed very underdominant rearrangements, and that there is only weak evidence that chromosome fusions have become fixed through positive natural selection. In summary, this work provides methods for researching chromosome evolution as well as new results about how rearrangements evolve and impact the speciation process.

Lay summary

Closely related species sometimes have different numbers of chromosomes. This variation is typically due to mutations where chromosomes break or join together. These mutations, which are known as fission and fusion rearrangements, have been suggested to play an important role in evolution. In particular, they are thought to promote the formation of new species because individuals from populations with different numbers of chromosomes may be less likely to produce fertile offspring. However, we do not yet know how this process happens in nature, if at all. More fundamentally, it is not clear how new rearrangements are able to spread in a population. In this thesis I use genome sequence data to investigate the role of fission and fusion rearrangements in evolution. I focus on species of butterfly in the genus *Brenthis*, as they are known to vary considerably in chromosome number. I perform two main types of analysis: comparing the structure of complete genome sequences to identify past chromosome rearrangements and using mutations between genome sequences to infer the evolutionary history of populations. I find that fission and fusion rearrangements have led to a decrease in gene exchange between a pair of recently diverged species, therefore demonstrating that these rearrangements are involved in the formation of new species in nature. Additionally, I show that new rearrangements in these butterfly species likely spread through chance rather than because of natural selection. The methods that I use and develop provide a framework for future research on chromosome rearrangements in other groups of species.

Acknowledgements

During my PhD I have had the privilege of working alongside and spending time with many knowledgeable and kind people.

First of all, I would like to thank my supervisors, Konrad Lohse and Simon Martin. I have greatly benefited from the encouragement, expertise, support and enthusiasm that you have both provided.

I have also been fortunate enough to receive mentorship from Dominik Laetsch and Derek Setter. Thank you for teaching me so much.

I have enjoyed countless discussions with other researchers about genomics, science, and life more generally. In particular, I would like to thank Sam Ebdon, Rishi De-Kayne, Meng Lu, Charlotte Wright, Gertjan Bisschop, Pablo Manuel Gonzalez de la Rosa, Robert Baird and Rachel Blow. Thank you all for providing help, inspiration and fun.

My parents, Julia and Andy, and my close friend Benji, have provided support and encouragement whenever I needed it, for which I am very grateful.

Most of all, I would like to thank my wife and best friend, Malin. Spending time away from you has certainly been the hardest part of this PhD. Thank you for everything.

Publications

Chapters 2-5 of this thesis were prepared for publication. As a result, there is some repetition between them and the placement of the Methods section varies depending on the requirements of the journal. Chapters 2, 4 and 5 of this thesis have been published as:

- **Mackintosh, A.**, Laetsch, D. R., Baril, T., Foster, R. G., Dincă, V., Vila, R., Hayward, A., & Lohse, K. (2022). The genome sequence of the lesser marbled fritillary, *Brenthis ino*, and evidence for a segregating neo-Z chromosome. G3, 12(6), jkac069.
- **Mackintosh, A.**, Vila, R., Laetsch, D. R., Hayward, A., Martin, S. H., & Lohse, K. (2023). Chromosome fissions and fusions act as barriers to gene flow between *Brenthis* fritillary butterflies. Molecular Biology and Evolution, 40(3), msad043.
- **Mackintosh, A.**, Vila, R., Martin, S. H., Setter, D., & Lohse, K. (2023). Do chromosome rearrangements fix by genetic drift or natural selection? Insights from *Brenthis* butterflies. In Press at Molecular Ecology.

I have also contributed to the following publications during the preparation of this thesis:

- **Mackintosh, A.**, Laetsch, D. R., Baril, T., Ebdon, S., Jay, P., Vila, R., Hayward, A., & Lohse, K. (2022). The genome sequence of the scarce swallowtail, *Iphiclides podalirius*. G3, 12(9), jkac193.
- Lundberg, M., **Mackintosh, A.**, Petri, A., & Bensch, S. (2023). Inversions maintain differences between migratory phenotypes of a songbird. Nature Communications, 14(1), 452.
- Wright, C. J., Stevens, L., **Mackintosh, A.**, Lawniczak, M., & Blaxter, M. (2023). Chromosome evolution in Lepidoptera. In review at Nature Ecology and Evolution.

Declaration

I declare that this thesis has been composed by myself and that the work has not be submitted for any other degree or professional qualification. I confirm that the work submitted is my own, except where work which has formed part of jointly-authored publications has been included. My contribution and those of the other authors to this work have been explicitly indicated below. I confirm that appropriate credit has been given within this thesis where reference has been made to the work of others.

Chapter 1: I wrote this chapter with feedback from Konrad Lohse and Simon Martin.

Chapter 2: I wrote this chapter with input from Dominik Laetsch, Roger Vila and Konrad Lohse. Samples were collected by Konrad Lohse, Roger Vila, Raluca Vodă and Vlad Dincă. DNA extractions were performed by Andrés Garcia de la Filia and Katy McDonald. The transposable element (TE) annotation was performed by Tobias Baril and Alex Hayward. They generated Figure 2.4 and also wrote the Methods and Results sections describing the TE annotation. I performed all other bioinformatics analyses and generated the HiC sequence data together with Robert Foster.

Chapter 3: I wrote this chapter with feedback from Pablo Manuel Gonzalez de la Rosa, Simon Martin, Konrad Lohse and Dominik Laetsch. Pablo Manuel Gonzalez de la Rosa provided BUSCO gene information for each nematode genome and also shared results from a previous analysis of this data. I wrote the command-line tool with input from Dominik Laetsch and performed all simulations and analysis of data.

Chapter 4: I wrote this chapter with input from Dominik Laetsch, Roger Vila, Simon Martin and Konrad Lohse. Samples were collected by Roger Vila, Sabina Vila, Alex Hayward, Dominik Laetsch and Konrad Lohse. DNA extractions were performed by me and Katy McDonald. I performed all bioinformatics analyses with guidance from Dominik Laetsch and Konrad Lohse.

Chapter 5: I wrote this chapter with input from Derek Setter, Simon Martin and Konrad Lohse. Samples were collected by Roger Vila, Vlad Dincă, Raluca Vodă, Leonardo Dapporto, Alex Hayward, Dominik Laetsch and Konrad Lohse. DNA extractions were performed by me and Katy McDonald. Derek Setter and I jointly implemented the sweep inference method in Mathematica. I performed all other analyses.

Chapter 6: I wrote this chapter with feedback from Konrad Lohse and Simon Martin.

Alexander Mackintosh, 22nd September 2023

Contents

1 General introduction 1

1.1 The role of chromosome rearrangements in evolution . . . 1

1.2 Chromosomal speciation . . . 3

1.3 Identifying chromosome rearrangements . . . 5

1.4 Model-based inference from population genomic data . . . 6

1.5 Thesis aims and overview . . . 10

2 The genome sequence of the lesser marbled fritillary, *Brenthis ino*, and evidence for a segregating neo-Z chromosome 13

2.1 Abstract . . . 13

2.2 Introduction . . . 13

2.3 Materials and methods . . . 14

2.3.1 Sampling . . . 14

2.3.2 Sequencing . . . 14

2.3.3 Genome assembly . . . 15

2.3.4 Karyotype analysis . . . 16

2.3.5 Synteny comparison . . . 17

2.3.6 Genome annotation . . . 17

2.3.7 Estimating heterozygosity . . . 18

2.4 Results . . . 19

2.4.1 Genome assembly . . . 19

2.4.2 Evidence for a segregating neo-Z chromosome . . . 20

2.4.3 Synteny . . . 22

2.4.4 Genome annotation . . . 22

2.5 Discussion . . . 24

2.6 Data availability . . . 25

3 Inferring inter-chromosomal rearrangements and ancestral linkage groups from synteny 27

3.1 Abstract . . . 27

3.2 Introduction . . . 27

3.3 Results . . . 30

3.3.1 Inferring inter-chromosomal rearrangements between two genomes . . . 30

3.3.2 Inferring ALGs and inter-chromosomal rearrangements across a phylogeny 32
3.3.3 The effect of marker error 35
3.3.4 Evaluating evidence for translocations through simulation 37
3.3.5 Rhabditid nematodes and Nigon elements 38
3.4 Discussion 41
3.4.1 Rearrangement rates and marker error 41
3.4.2 Lessons from a reanalysis of nematode genomes 42
3.4.3 Gene tree discordance 44
3.4.4 Outlook 44
3.5 Methods 45
3.5.1 An overview of syngraph 45
3.5.2 ALG reconstruction under missingness and error 45
3.5.3 Simulations 46
3.5.4 Reanalysis of nematode genomes 47
3.6 Data availability 48

4 Chromosome fissions and fusions act as barriers to gene flow between *Brenthis* fritillary butterflies 49

4.1 Abstract 49
4.2 Introduction 49
4.2.1 Chromosomal speciation 49
4.2.2 Chromosome evolution in butterflies 52
4.2.3 Overview 53
4.3 Results 54
4.3.1 Diversity and divergence 54
4.3.2 Demographic history 55
4.3.3 Synteny 57
4.3.4 Variation in m_e across the genome 58
4.4 Discussion 61
4.4.1 The effect of fission and fusion rearrangements on gene flow 61
4.4.2 The underdominance paradox 62
4.4.3 Variation among rearrangements 63
4.4.4 Other determinants of m_e variation 64
4.4.5 Outlook 65
4.5 Materials and methods 65

4.5.1 Sampling . . . 65
4.5.2 Sequencing . . . 65
4.5.3 Genome assembly . . . 66
4.5.4 Synteny analysis . . . 66
4.5.5 Variant calling and filtering . . . 67
4.5.6 Summaries of diversity and divergence . . . 68
4.5.7 Demographic modelling with gIMble . . . 68
4.5.8 Statistical analysis . . . 69
4.6 Data availability . . . 70

5 Do chromosome rearrangements fix by genetic drift or natural selection? Insights from *Brenthis* butterflies 71

5.1 Abstract . . . 71
5.2 Introduction . . . 71
5.2.1 How do chromosome rearrangements fix? . . . 71
5.2.2 Inferring selective sweeps . . . 73
5.2.3 Overview . . . 74
5.3 Materials and methods . . . 75
5.3.1 Sampling and sequencing . . . 75
5.3.2 Genome assembly . . . 75
5.3.3 Synteny analysis . . . 75
5.3.4 Fitting a multi-species demographic history . . . 76
5.3.5 Identifying runs of homozygosity . . . 77
5.3.6 Inferring selective sweeps from blockwise mutation configurations . . . 78
5.3.7 Simulations . . . 79
5.3.8 Statistical analysis . . . 79
5.4 Results . . . 80
5.4.1 A genome assembly of *Brenthis hecate* . . . 80
5.4.2 Synteny between Argynnini butterfly species . . . 80
5.4.3 The demographic history of *Brenthis* butterflies . . . 81
5.4.4 Runs of homozygosity . . . 83
5.4.5 Parameter estimates and statistical support for simulated selective sweeps 83
5.4.6 Evidence for an enrichment of selective sweeps around chromosome fusions . . . 86
5.4.7 Evidence for individual selective sweeps . . . 88

5.5 Discussion 89
5.5.1 Patterns of chromosome evolution 89
5.5.2 Genetic drift and underdominance 90
5.5.3 The role of positive natural selection in the fixation of chromosome fusions 92
5.5.4 Outlook 94
5.6 Data availability 94

6 General discussion 96
6.1 Methods for investigating chromosome evolution 96
6.2 Methods for inferring the evolutionary history of populations 97
6.3 The role of fission and fusion rearrangements in *Brenthis* butterflies 99

A Supplementary Materials for Chapter 2 125
A.1 Supplementary figures 125
A.2 Supplementary tables 130

B Supplementary Materials for Chapter 3 132
B.1 Supplementary methods 132
B.2 Supplementary tables 135

C Supplementary Materials for Chapter 4 136
C.1 Supplementary note 1 136
C.2 Supplementary note 2 138
C.3 Supplementary figures 139
C.4 Supplementary tables 143

D Supplementary Materials for Chapter 5 145
D.1 Supplementary methods 145
D.1.1 Gene annotation 145
D.1.2 Fitting a multi-species demographic model with fastsimcoal2 145
D.1.3 Strategies for fitting sweep models to the bSFS 146
D.1.4 Fitting a finite-island model 147
D.2 Supplementary figures 148
D.3 Supplementary tables 150

E Appendix 153

General introduction

1.1 The role of chromosome rearrangements in evolution

Chromosomes are a fundamental unit of genetic inheritance (Sutton 1903; Boveri 1904). They determine which DNA sequences are inherited together and which are separated by random assortment. Chromosomes do, however, differ markedly between different organisms. In fact, differences in chromosome size, structure and number can often be observed between closely related species, or even individuals of the same species (White 1978b). The mutations that generate this kind of variation are collectively known as chromosome rearrangements and have been of interest to evolutionary biologists for more than a century (Robertson 1916; Sturtevant 1921). Recent advances in genome sequencing have shown that they are more common than once thought (Feuk et al. 2005; Jiao and Schneeberger 2020). Given their prevalence, an obvious question is what role, if any, chromosome rearrangements play in the evolutionary processes. Ideally, we would like to know how particular types of rearrangement interact with evolutionary forces such as recombination, genetic drift, and natural selection, as well as epi-phenomena like adaptation and speciation (Rieseberg 2001; Yeaman 2013; Feulner and De-Kayne 2017). While significant progress has been made in our understanding of the role that inversion rearrangements play in evolution (Wellenreuther and Bernatchez 2018), the same is not yet true for less common rearrangements such as fissions and fusions.

Part of a chromosome can change orientation through ectopic recombination or staggered breaks (Ranz et al. 2007), leading to an inversion. These rearrangements were first identified by comparing genetic maps of *Drosophila melanogaster* and *D. simulans* (Sturtevant 1921) and are now routinely found through population or comparative genomics methods (Rausch et al. 2012; Corbett-Detig and Hartl 2012; Jay et al. 2021; Lundberg et al. 2023). Importantly, recombination between ancestral and inverted sequences is suppressed because a single crossover generates unbalanced chromosomes. Although a low level of recombination can persist through double crossovers and gene conversion (Rozas and Aguade 1994), there is typically enough suppression for at least some divergence to accumulate between the two types of sequence.

As a result, locally established inversions can reduce gene flow between populations and also facilitate local adaptation by reducing migration load (Kirkpatrick and Barton 2006). Individuals that are heterozygous for an inversion often suffer only minor or no reduction in fitness (Coyne et al. 1993), meaning that inversions can spread by genetic drift or natural selection if they trap co-adapted alleles. Analysis of genome sequence data has uncovered many example of inversions underlying ecologically important phenotypes (Twyford and Friedman 2015; Küpper et al. 2016; Mérot et al. 2021; Koch et al. 2021). Although one could argue that current population genomics studies are biased towards identifying such effects, the fact that inversions play a role in adaptation and speciation is unequivocal.

Unlike inversions, fission and fusion rearrangements result in increases and decreases in chromosome number, respectively. Fissions are the result of an unrepaired double-strand break (DSB) during meiosis, whereas fusions happen when 'sticky' ends of non-homologous chromosomes (generated through DSB or telomere shortening) are joined together (White 1973; Lysak 2022). These rearrangements tend to be less common than inversions. This idea is supported by the fact that synteny (i.e. the co-occurrence of loci on the same chromosome) can often be conserved across hundreds of millions of years of evolution, unchanged by fissions or fusions. By contrast, the order of sequences within chromosomes are usually mixed extensively by inversion over such long time scales (Simakov et al. 2022; Schultz et al. 2023). However, certain groups of closely related taxa display remarkable variation in chromosome number due to frequent fissions and fusions (Zima et al. 1996; Hipp et al. 2009; Lukhtanov et al. 2011; Talavera et al. 2013; Potter et al. 2017). In fact, recent analyses have shown that fissions and fusions are more common than large inversions in *Lysandra* butterflies (Pazhenkova and Lukhtanov 2023) and *Carex* sedges (Escudero et al. 2023). Both of these groups have holocentric chromosomes, without localised centromeres, which is thought to increase the probability that a new fission or fusion spreads in a population (Melters et al. 2012; Lucek et al. 2022). There are, however, plenty of examples of taxa with monocentric chromosomes displaying high rates of fission and fusion (White et al. 1964; Wahrman et al. 1969; Craddock 1970), suggesting that factors other than centromere structure explain why these rearrangements are more common in certain species (Ruckman et al. 2020).

There are two immediate evolutionary consequences of a new fission or fusion rearrangement: (i) a change in the number and distribution of crossovers, and (ii) reduced fitness in heterozygotes. Changes in recombination rate have been observed in species with either recent or polymorphic fissions and fusions (Davey et al. 2017; Capilla et al. 2014; Näsvall et al. 2023). Unlike inversions, these rearrangements do not generally lead to local recombination suppression in heterozygotes (Davisson and Akeson 1993). Nonetheless, broad reductions in the rate of recombination can promote speciation (Martin et al. 2019) and increases should lead to more efficient natural selection (Hill and Robertson 1966). The second effect mentioned above, reduced fitness in heterozygotes, is typically attributed to the multivalents that form at meiosis being prone to unbalanced segregation and thus leading to aneuploidy in subsequent generations (White 1973). Although there is evidence for this phenomenon across many taxa, the fitness reduction varies from undetectable to complete (Narain and Fredga 1997; Castiglia and Capanna 2000; Hora et al. 2019; Yoshida et al. 2023), and likely depends on the specific rearrangements and dynamics of meiosis in different species (Lukhtanov et al. 2018). In principle, the reduced fitness of heterozygotes could lead to speciation (see next section). Overall, there is good reason to suspect that fission and fusion rearrangements influence the evolution of populations.

1.2 Chromosomal speciation

Speciation research focuses on identifying the genetic and ecological factors that lead to reproductive isolation between populations and ultimately new species. It is often suggested that fission and fusion rearrangements promote speciation (White 1968; Bush 1975). While there are many different models of chromosomal speciation (see Rieseberg 2001), those that focus on fissions and fusions tend to rely on the following reasoning: Assuming a rearrangement rises to high frequency in one population but is absent from another, offspring of migrant individuals will tend to be heterozygous for the rearrangement. These heterozygotes will have reduced fitness due to unbalanced segregation during meiosis and so gene flow between these populations will be reduced by natural selection.

This simplistic model offers an appealing explanation for the observation that closely related species with differences in chromosome number often have strong post-zygotic isolation. However, a fission / fusion that confers a significant enough heterozygous disadvantage (underdominance) to prevent gene flow, is very unlikely to rise to high frequency in the first place (Hedrick 1981; Walsh 1982). Models of chromosomal speciation involving fissions and fusions (hereafter simply referred to as chromosomal speciation) must therefore include some kind of solution to this 'underdominance paradox'. For example, strong genetic drift and meiotic drive have both been suggested as mechanisms allowing underdominant rearrangements to become fixed (Wright 1941; White 1978b; Hedrick 1981). Models sometimes assume that a single rearrangement has only a weak fitness effect, and so can become fixed in one population, and that only the build-up of multiple rearrangements leads to underdominance in hybrids (Walsh 1982; Baker and Bickham 1986). Alternatively, models of chromosomal speciation may focus exclusively on the recombination modifying effects of fission / fusion rearrangements, which could promote speciation without underdomiance (Rieseberg 2001).

While there is no general consensus yet on how chromosomal speciation may proceed in nature, a simpler and more fundamental question is whether there is any evidence that it happens at all, and, if so, how frequent it is. There are several taxonomic groups where rates of chromosome number change and speciation are positively correlated (Leaché et al. 2016; de Vos et al. 2020), suggesting either that fissions and fusions promote speciation, or that both rates are associated with another variable such as effective population size (Bush et al. 1977). Sharp clines in rearrangement frequency within species are consistent with selection against heterokaryotypes (Barton and Hewitt 1981), perhaps representing the early stages of speciation, but could be a result of other genetic factors in linkage disequilibrium with the rearrangement. Increased genetic differentiation / divergence around fissions and fusions, suggesting reduced gene flow, has been found in *Sorex* shrews (Basset et al. 2006) but the opposite was found in rock-wallabies (Potter et al. 2022). All in all, there is certainly some evidence implicating fissions and fusions as drivers of speciation, but still no examples where we can be confident that these rearrangements led to the formation of new species.

During the preparation of this thesis, Yoshida et al. (2023) published evidence for the effect of chromosome fusions on reproductive isolation between a pair of *Pristionchus* nematode species. They show an absence of prezygotic isolation between *P. pacificus* and *P. exspectatus*, but reduced fertility in F1 individuals. Their comparison of genome assemblies shows that an ancestral chromosome has fused independently to a different chromosome in each species. These fused chromosomes have reduced rates of crossover (within species), contain a large effect quantitative trait locus for hybrid sterility, and often result in trisomy in the offspring of F1s. This elegant work is perhaps the strongest evidence so far for the idea that fusions promote speciation, and fits one model of chromosomal speciation particularly well (speciation by monobrachial fusions, Baker and Bickham 1986). One limitation is that these species are highly diverged from one another, with synonymous-site divergence at $\sim$ 0.3 which corresponds to a split time of around 20 million generations. So while there is strong evidence that these fusions cause reduced fertility in present-day crosses, we do not know at what stage of the speciation process they arose and what, if any, impact they had on the build up of barriers. Ideally, one would investigate a pair of sister species whose genomes differ due to fissions / fusions, but their speciation time is recent enough for genealogical histories to contain information about whether those rearrangements caused a reduction in gene flow. While most species pairs will not fulfill these criteria, those that do will provide much needed information about how (if at all) chromosomal speciation happens in nature.

1.3 Identifying chromosome rearrangements

The ability to reliably identify and characterise chromosome rearrangements is a prerequisite for investigating their role in evolution. Robertson (1916) identified centric fusions (i.e. Robertsonian translocations) by comparing the chromosomes of different species of Orthoptera through microscopy. The strongest evidence for these rearrangements came from bush crickets in the genus *Jamaicana*, where heterokaryotype individuals showed pairing of two rod-shaped (acrocentric) chromosomes with a single V-shaped (metacentric) chromosome (Woolsey 1915). Similarly, Muller (1940) used genetic maps of *Drosophila* to show that chromosome arms are conserved across species but that their arrangement varies due to fusions. Researchers today have the possibility of generating complete genome sequences rather than relying on microscopy or genetic maps. Yet, the core principle for identifying rearrangements –

comparing the arrangement of homologous sequences or markers across different genomes – remains the same.

Pairwise genome alignments are a common method used to identify chromosome rearrangements (Li 2018; Goel et al. 2019; Song et al. 2022). However, such comparisons do not provide information about the timing of rearrangements nor which arrangement represents the ancestral state. More detailed inference can be made by considering the genome evolution of many species along a phylogenetic tree. The aim of such analyses is to estimate characteristics of ancestral genomes at the internal nodes of the tree and place individual rearrangements on branches (Fertin et al. 2009). In principle, such ancestral state reconstructions provide information about the rates of different types of rearrangements and how these vary across the tree. A multispecies approach is an obvious choice for investigating the processes that drive and constrain chromosome evolution over deep evolutionary time (Simakov et al. 2022) and can equally be used to reconstruct the recent rearrangement history of a single genus (Ostevik et al. 2020).

There has been considerable theoretical work on the combinatorics of genome rearrangements (Fertin et al. 2009) and multiple software tools have been developed for reconstructing ancestral genomes and past rearrangements from sequence data (Bourque and Pevzner 2002; Tesler 2002; Ma 2010; Hu et al. 2013; Perrin et al. 2015; Feijão and Araujo 2016; Kim et al. 2017). Such methods would ideally fit probabilistic models that allow for variation in the rate of different rearrangements across a phylogeny (Moshe et al. 2022). However, there is often a vast number of plausible rearrangement histories that can explain the data, making this model fitting task challenging. Most methods therefore rely on parsimony to efficiently estimate rearrangement histories. While convenient, it is not clear how well parsimony-based methods can infer past chromosome fissions and fusions, especially in groups of organisms where these rearrangements are very common.

1.4 Model-based inference from population genomic data

One way to determine the role that fission and fusion rearrangements play in evolution is to infer the evolutionary history of populations with recently fixed rearrangements.

In particular, one could ask whether rearrangements have acted as barriers to gene flow between populations or if their evolution is associated with strong genetic drift. These questions can only be addressed by connecting models of evolution to the sequence variation that exists in present-day genomes. In the first part of the 20$^{\text{th}}$ century, population geneticists modelled evolution as changes in allele frequency through time (Fisher 1923; Kimura 1957; Moran 1958). Forwards-in-time modelling is certainly still useful and often unavoidable (Gutenkunst et al. 2009; Messer 2013), however a backwards-in-time approach – the coalescent (Kingman 1982; Tajima 1983; Hudson 1983a) – has become the cornerstone of modern population genomic inference.

Coalescent theory focuses on the genealogical history of a sample of present-day individuals, rather than the entire population. As an illustration, consider a sample of two sequences (or lineages) from a population that underwent a bottleneck t generations ago (Figure 1.1A). Backwards in time, the time for two lineages to coalesce is exponentially distributed with rate $\frac{1}{2N_e}$, where $2N_e$ is the (diploid) effective population size. The probability that two lineages reach the bottleneck t generations ago without coalescing is therefore $e^{-t/2N_e}$ and the probability that they then coalesce during the bottleneck is $1 - e^{-dB/2N_e}$, where d is the duration of the bottleneck and B the relative reduction in effective population size (Figure 1.1B).

This simple example illustrates how the coalescent can relate evolutionary processes to genealogical histories (Figure 1.1). Although this example includes only two lineages, the coalescent process can model k lineages as long as $k << 2N_e$. The coalescent can also be extended to include multiple populations (Tajima 1983), recombination (Hudson 1983b; McVean and Cardin 2005) and even approximations of natural selection (Hudson and Kaplan 1988; Barton 1998).

One approach to population genomic inference is to derive expectations for a null-model under the coalescent process and then test for deviations from it, e.g. Tajima's D and Patterson's D (Tajima 1989; Green et al. 2010). Ideally, however, explicit models of evolution should be fit to sequence data, as this allows insight into the individual evolutionary forces that have shaped variation in present-day genome sequences. Felsenstein (1988) formalised calculating the likelihood of a model (T) given sequence data (D) as

$$P(D \mid T) = \sum_{G} P(G \mid T)\, P(D \mid G)\,,$$

where G represents a genealogical history. For most practicable cases, G is unobservable and therefore unknown. This is because D is almost always an alignment of genome sequences rather than the actual genealogical history. As a result, the likelihood can only be obtained by summing over all possible genealogical histories that are consistent with the data.

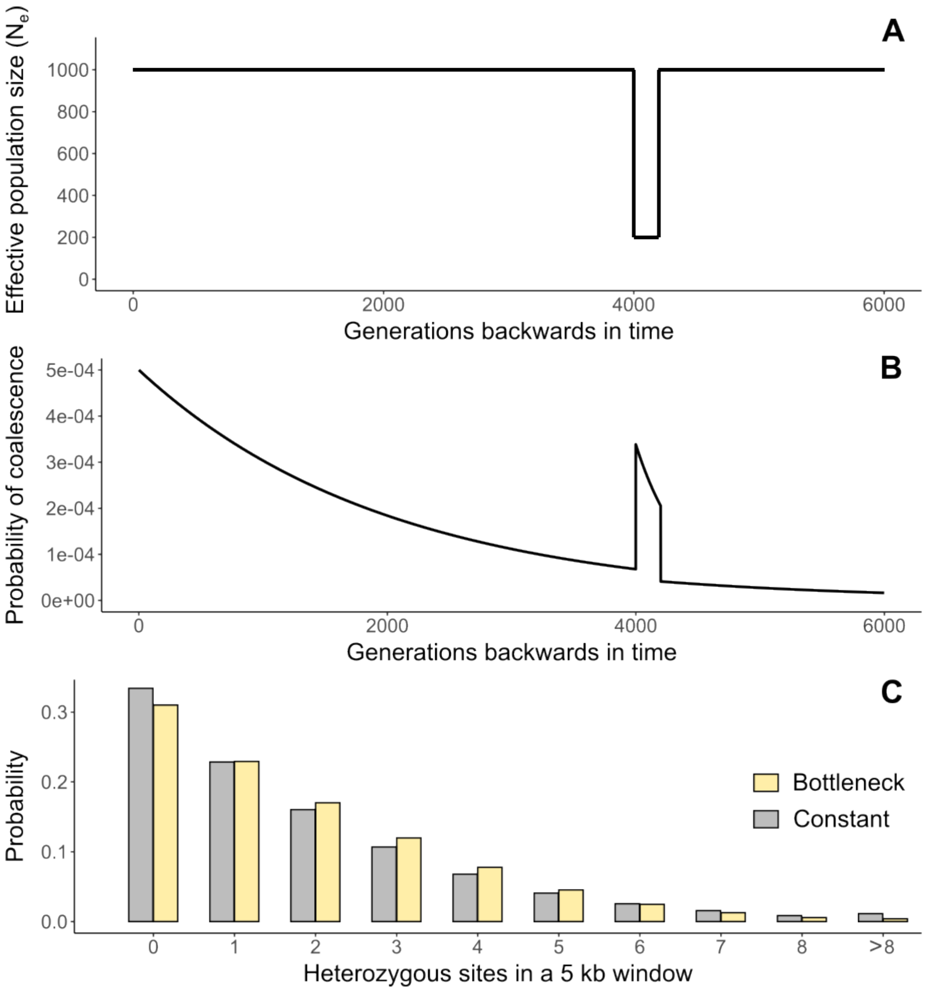

Figure 1.1: An example of a population bottleneck. **(A)** The effective size (N_e) of the population through time. The bottleneck temporarily reduces N_e by a factor of five and lasts for 200 generations. **(B)** The expected distribution of coalescence times for a pair of lineages under the bottleneck history in (A). **(C)** Expected counts of heterozygous sites in 5 kb windows under the bottleneck history and a constant population size history with equivalent overall diversity (N_e = 894). Counts were approximated by Monte-Carlo simulation (100,000 simulated windows with mutation and recombination rates of 1×10^{-7} per-base per-generation).

Considering all possible genealogical histories may seem like daunting task, but there are several ways to approach or circumvent it. One solution is to focus on summaries of genealogies that can be derived under the coalescent while integrating over G. For example, the average length of branches in a genealogical tree with i descendants (approximated by the site frequency spectrum or SFS) is one such summary that has been used extensively for demographic and selective inference (Fu 1995; Liu and Fu 2020; Kamm et al. 2020; Nielsen et al. 2005; Setter et al. 2020). While certainly useful, there are limits to the complexity of models for which the expected SFS can be derived analytically and similar models can sometimes result in identical spectra (Lapierre et al. 2017).

An alternative strategy is to approximate the likelihood of a model by Monte-Carlo coalescent simulation (Hudson 2002; Baumdicker et al. 2021, see Figure 1.1C for an example). This requires the simulation of many thousands of genealogies, and so is often slow, but has the benefit of allowing likelihoods to be approximated for arbitrarily complex models (Excoffier et al. 2021). Methods employing this approach often still only consider certain aspects of genealogical trees (e.g. the SFS), but can in principle use any information that is accessible in both simulated genealogies and genome sequences (Beeravolu et al. 2018).

A third approach is to estimate the genealogical history of a sample of genomes, and then perform the likelihood calculation assuming that history. The ancestral recombination graph (ARG) describes the history of a sample through the coalescent process with recombination (Griffiths and Marjoram 1997). Methods that estimate the ARG from a sample of genome sequences have recently become more efficient (Kelleher et al. 2019; Speidel et al. 2019) and, as a result, are growing in popularity. One attraction of these methods is that the ARG contains genealogies that are informed by all linked mutations and so are information-rich. The downside is that this approach often ignores the uncertainty in the ARG given the data, which will be substantial whenever mutation and recombination happen at similar rates.

It would certainly be possible to gain useful information about the evolutionary history of populations (with recent chromosome rearrangements) using many of the methods discussed above. However, in this thesis I instead use a class of inference method that derives the joint distribution of branch-specific mutation counts for genealogies with small sample size (Lohse et al. 2011). This has multiple benefits: Firstly, the joint distribution of branch lengths contains more information than the expected length of each branch (the SFS), *a priori*. Secondly, likelihoods can be calculated analytically for models that are complex enough to capture demographic and selective process of interest (Bunnefeld et al. 2015; Bisschop et al. 2021; Laetsch et al. 2022). Finally, the method integrates over all possible (non-recombinant) genealogies that could underlie the data in a short block of sequence, rather than assuming a single genealogy and ignoring the uncertainty associated with it. Although haplotype / ARG-based analyses may be preferable for datasets containing hundreds of phased genomes, the fact that the method of Lohse et al. (2011) can be applied to unphased mutations from a small sample of genomes makes is a natural choice for investigating the population history of non-model organisms.

1.5 Thesis aims and overview

It should be clear from the above sections that fission and fusion rearrangements have the potential to influence evolution, but their exact role and importance is unresolved. The primary aim of this thesis is therefore to use comparative and population genomic methods to gain information about the role of these rearrangements in evolution, with a particular focus on speciation.

In this thesis I focus on chromosome rearrangements in a non-model system: butterflies in the genus *Brenthis* (Hübner, 1819, Nymphalidae). The genus consists of four species: *B. daphne* (Denis and Schiffermüller, 1775), *B. ino* (Rottemburg, 1775), *B. hecate* (Denis and Schiffermüller, 1775) and *B. mofidii* (Wyatt, 1969). They have Palearctic distributions with *B. daphne* and *B. ino* being particularly widespread. Species within this genus vary in chromosome number (Pazhenkova and Lukhtanov 2019), suggesting that fission and fusion rearrangements are common. Interestingly, chromosome numbers of other species within the tribe Argynnini display much less variation (Robinson 1971), consistent with a change in chromosome evolution dynamics in a recent common ancestor of *Brenthis* butterflies. Three chapters of this thesis focus on the genomics of *Brenthis* butterflies, whereas one explores the more general problem of identifying past chromosome rearrangements. Here I give a brief description of each chapter.

In chapter 2 I present the first genome assembly for any *Brenthis* species. I compare the *B. ino* genome assembly with that of another nymphalid butterfly, revealing broad patterns of chromosome evolution. Additionally, I use the phase information in HiC sequence data to partition reads into haplotye-specific datasets. I demonstrate that this approach can be used to detect rearrangements in a heterozygous state.

In chapter 3 I explore how synteny information in genome sequences can be used to infer past rearrangements and chromosome content. I implement previously described algorithms in a command-line tool and evaluate their performance through simulation. I also reanalyse a set of 14 nematode genomes and compare the results to those of an alternative method.

In chapter 4 I investigate the speciation history of *B. ino* and *B. daphne*. I compare genome assemblies of the two species and find that they differ as a result of multiple fissions and fusions. I then fit demographic models to patterns of mutation in short sequence blocks, estimating effective rates of drift (N_e) and migration (m_e) along the genome. Given estimates of post-divergence m_e, I ask whether fissions and fusions have acted as barriers to gene flow between these species and therefore whether they have promoted speciation.

In chapter 5 I ask whether fissions and fusions in the genus *Brenthis* have fixed through genetic drift or positive natural selection. I used the method from chapter 3 to infer past rearrangements, revealing a complex history of fissions and fusions. I then fit a three-population demographic model to obtain estimates of N_e and consider the likelihood of underdominant rearrangement becoming fixed through drift. I also fit models of hard selective sweeps to explore whether chromosome fusions became fixed through natural selection or meiotic drive.

Finally, in chapter 6, I briefly discuss the results of chapters 2 - 5 in relation to each other and the broader literature. I suggest alternative approaches for making inference from genome sequence data, and I also discuss how my research has contributed to our understanding of chromosome rearrangements in evolution.

The genome sequence of the lesser marbled fritillary, *Brenthis ino*, and evidence for a segregating neo-Z chromosome

2.1 Abstract

The lesser marbled fritillary, *Brenthis ino* (Rottemburg, 1775), is a species of Palearctic butterfly. Male *B. ino* individuals have been reported to have between 12 and 14 pairs of chromosomes, a much reduced chromosome number than is typical in butterflies. Here we present a chromosome-level genome assembly for *B. ino*, as well as gene and transposable element annotations. The assembly is 411.8 Mb in length with a contig N50 of 9.6 Mb and a scaffold N50 of 29.5 Mb. We also show evidence that the male individual from which we generated HiC data was heterozygous for a neo-Z chromosome, consistent with inheriting 14 chromosomes from one parent and 13 from the other. This genome assembly will be a valuable resource for studying chromosome evolution in Lepidoptera, as well as for comparative and population genomics more generally.

2.2 Introduction

The lesser marbled fritillary, *Brenthis ino* (Rottemburg, 1775), is a species of butterfly in the family Nymphalidae. It has a Palearctic distribution, is widespread in Europe with variance in local abundance, and can be found as far East as Japan and Siberia. It is monovoltine and feeds on plants in the family Rosaceae, including some species in the genera *Filipendula*, *Aruncus*, *Sanguisorba*, and *Rubus*. While most butterflies in the family Nymphalidae, and Lepidoptera more widely, have 31 (or close to 31) pairs of chromosomes (de Vos et al. 2020), *B. ino*, along with its sister species *B. daphne* (Denis and Schiffermüller, 1775), has an unusually low chromosome count. Federley (1938) reported male haploid chromosome numbers of 12 and 13 for individuals collected in Finland, consistent with segregating chromosomal fissions or fusions in the population. However, other males sampled in Finland and Sweden consistently displayed 13 chromosome pairs (Saitoh 1987, 1991). In Japan, where the subspecies

B. ino mashuensis (Kono, 1931) and *B. ino tigroides* (Fruhstorfer, 1907) are found, a male chromosome number of 14 has been consistently observed (Maeki and Makino 1953; Saitoh et al. 1989).

Currently, there are no genome assemblies for species in the genus *Brenthis* and information about chromosome evolution in the genus is confined to cytological data. Here we present a chromosome-level genome assembly of *B. ino* as well as gene and transposable element (TE) annotations. We also show that one of the individuals we sampled was heterozygous for a neo-Z chromosome, consistent with there being karyotypic variation within the Spanish population from which we sampled.

2.3 Materials and methods

2.3.1 Sampling

Three individuals were collected by hand netting in Somiedo, Braña de Mumian, Asturias, Spain (SO_BI_364, SO_BI_375, SO_BI_376) and one in Larche, Alpes-de-Haute-Provence, France (FR_BI_1497, RVcoll12O846) (Table A.1). Spanish individuals were flash frozen in a liquid nitrogen dry shipper. The French specimen was dried and, after some days, stored in ethanol at -20°C.

2.3.2 Sequencing

High molecular weight (HMW) DNA was extracted from the thorax of a flash frozen individual (SO_BI_364) using a salting out extraction protocol. In brief, tissue was homogenised in cell lysis buffer using a micro-pestle and then incubated with Proteinase K overnight at 56°C, followed by a further one hour incubation at 37°C with RNase A, before precipitating and discarding proteins. Finally, DNA was precipitated using isopropanol and the resulting pellet was washed with ethanol.

Edinburgh Genomics (EG) generated a SMRTbell sequencing library from the HMW DNA which was sequenced on three SMRT cells on a Sequel I instrument to generate 28.4 Gb of Pacbio continuous long read (CLR) sequence data. From the same HMW DNA extraction, EG also generated a TruSeq library (350 bp insert) and 33.5 Gb of Illumina whole genome (WGS) paired-end reads on a Novaseq 6000. Pacbio and

Illumina protocols were followed for library preparation, QC and sequencing.

A second individual (SO_BI_375) was used for chromatin conformation capture (HiC) sequencing. The HiC reaction was done using an Arima-HiC kit, following the manufacturer's instructions for flash frozen animal tissue. The NEBNext Ultra II library was sequenced on an Illumina MiSeq at EG, generating 4.8 Gb of paired-end reads.

Illumina WGS paired-end reads were also generated for the same individual used for HiC sequencing (SO_BI_375) as well as the French female individual (FR_BI_1497) that did not contribute to the assembly.

Paired-end RNA-seq data (for individual SO_BI_376) was previously generated and analysed by Ebdon et al. (2021) (ENA experiment accession ERX5086186).

2.3.3 Genome assembly

Illumina WGS, RNA-seq, and HiC reads were adapter and quality trimmed with fastp v0.2.1 (Chen et al. 2018).

The Pacbio reads were assembled with Nextdenovo v2.4.0 (Hu et al. 2023) using default parameters. Contigs were polished twice by aligning Illumina WGS reads and correcting consensus errors with HAPO-G v1.1 (Aury and Istace 2021). Contigs belonging to non-target organisms were identified using blobtools v1.1.1 (Laetsch and Blaxter 2017) and subsequently removed. Duplicated regions (haplotigs and overlaps) were identified and removed with purge_dups v1.2.5 (Guan et al. 2020). Mapping of Pacbio reads and Illumina WGS reads for the above steps were performed with minimap2 v2.17 and bwa-mem v0.7.17, respectively (Li 2018, 2013).

The trimmed HiC reads were aligned to the contig-level assembly with Juicer v1.6 (Durand et al. 2016). Scaffolding was performed with 3d-dna v180922 (Dudchenko et al. 2017). The initial scaffolding generated by 3d-dna was manually partitioned into chromosomes and misassembly corrected with Juicebox v1.11.08 (Robinson et al. 2018).

A k-mer spectrum, with $k = 21$ and a maximum counter value of 10^7, was generated using KMC v3.1.1 (Kokot et al. 2017) and genome size was estimated from the spectrum using Genomescope v2.0 (Ranallo-Benavidez et al. 2020).

Gene completeness was evaluated using BUSCO v5.2.2 with the insecta_odb10 dataset (n=1367) (Manni et al. 2021). Kmer QV was calculated using Merqury v1.3 (Rhie et al. 2020).

The mitochondrial genome was assembled and annotated using the Mitofinder pipeline v1.4 (Allio et al. 2020). Illumina WGS reads from SO_BI_364 were assembled with metaSPAdes v3.14.1 (Nurk et al. 2017) and tRNAs were annotated with MiTFi (Jühling et al. 2011).

2.3.4 Karyotype analysis

After scaffolding, chromosomes 11 and 13 displayed an intermediate HiC contact map pattern, suggesting a potential fusion of the chromosomes in one of the haplotypes.

In order to investigate this further we generated haplotype-specific HiC maps for chromosomes 11 and 13. First, we created a version of the assembly where chromosomes 11 and 13 were scaffolded together. WGS and HiC reads (from SO_BI_375) were mapped to this assembly with bwa-mem v0.7.17. Alignments were deduplicated with sambamba v0.6.6 (Tarasov et al. 2015). Heterozygous variants were called from the WGS alignments with freebayes v1.3.2-dirty (Garrison and Marth 2012). Variants were then normalised with bcftools v1.8 (Danecek et al. 2021) and decomposed with vcfallelicprimitives (Garrison et al. 2021). Normalisation involves left-aligning variants and ensuring that they are represented parsimoniously. Decomposition is the splitting up of MNPs and complex variants into multiple SNPs and/or indels. Variants were filtered for coverage ($>$ 7 and $<$ 56 reads) with bcftools. The remaining SNPs were phased using HAPCUT2 v1.3.3 with both the WGS and HiC alignments as input (Edge et al. 2017).

We developed a tool (chomper.py, see Data availability) which uses the phased SNPs from HAPCUT2 to partition aligned HiC reads by haplotype. For any read pair whose alignment encompasses at least one phased SNP, we can ask whether the alleles in the read are associated with haplotype 1 or 2. If a read pair contains alleles exclusively associated with one haplotype, then it is assigned to that haplotype-specific read set. If it instead contains alleles associated with both haplotypes, then it is discarded. Haplotype-specific HiC read sets were then aligned back to the original assembly with Juicer and visualised with HiC_view.py (parameters `-b 250 -s 10`, see Data availability).

To identify the Z chromosome, one male (SO_BI_364) and one female (FR_BI_1497) individual were mapped to the assembly with bwa-mem v0.7.17 and median, window-wise coverage was calculated using mosdepth v0.3.2 (Pedersen and Quinlan 2017).

2.3.5 Synteny comparison

Synteny in the *B. ino* genome was compared to synteny in another nymphalid genome, *Melitaea cinxia* (GCA_905220565.1 Vila et al. 2021). A total of 5178 lepidoptera_obd10 BUSCO genes were identified in both assemblies using BUSCO v5.2.2. The positions of these genes in both assemblies were visualised using busco2synteny.py (see Data availability).

2.3.6 Genome annotation

The Illumina RNA-seq reads were mapped to the assembly with HISAT2 v2.1.0 (Kim et al. 2019). The softmasked assembly and RNA-seq alignments were used for gene prediction with braker2.1.5 (Hoff et al. 2015, 2019; Li et al. 2009; Barnett et al. 2011; Lomsadze et al. 2014; Buchfink et al. 2015; Stanke et al. 2006, 2008). Gene annotation statistics were calculated with GenomeTools v1.6.1 (Gremme et al. 2013).

Transposable elements (TEs) were annotated using the Earl Grey TE annotation pipeline (`https://github.com/TobyBaril/EarlGrey`, Baril et al. 2022). Briefly, known repeats were masked with RepeatMasker v4.1.2 (Smit et al. 2015) using the Lepidoptera library from RepBase v23.08 and Dfam release 3.3 (Jurka et al. 2005; Hubley et al. 2015). Following this, a *de novo* repeat library was constructed using RepeatModeler2 v2.0.2 (Flynn et al. 2020) with RECON v1.08 and RepeatScout v1.0.6. Subsequently, Earl Grey generated maximum-length consensus sequences for the *de novo* sequences identified by RepeatModeler2 using an automated version of the ‘BLAST, Extract, Extend’ process, as previously described (Platt et al. 2016). The resulting *de* *novo* repeat library was combined with the RepBase and Dfam libraries used in the initial masking step to annotate repetitive elements using RepeatMasker. Full-length LTR elements were identified using LTR_Finder v1.07 with the LTR_Finder parallel wrapper (Xu and Wang 2007; Ou and Jiang 2019). Final TE annotations were defragmented and refined using a loose merge in RepeatCraft (`-loose`), followed by maintaining the longest of any overlapping annotations with MGkit v0.4.1 (`filter-gff -c length -a length`) (Wong and Simakov 2018; Rubino and Creevey 2014). Finally, all repeats $<$ 100bp in length were removed before final TE quantification to decrease spurious hits.

Following gene annotation, gene flanks were defined as regions that were $>=$ 20*kb* upstream and downstream of genes. We expect these regions to be enriched for regulatory sequences, including both proximal promoters and distal elements. We define regions as intergenic if they are neither genic (start/stop codons, exons, and introns) nor gene flanks. Bedtools intersect v2.27.1 (Quinlan and Hall 2010) was used to determine overlap (`-wao`) between TEs and genomic features. Following this, quantification and plotting was performed in R, using the tidyverse package (Wickham et al. 2019; R Core Team 2021; RStudio Team 2020).

2.3.7 Estimating heterozygosity

To estimate heterozygosity, WGS reads were mapped to the assembly with bwa-mem v0.7.17 and variants were called with freebayes v1.3.2-dirty. Variant calls were normalised with bcftools v1.8 and decomposed using vcfallelicprimitives (for an explanation of these terms see Methods, Karyotype analysis). Callable sites, where coverage was $>$ 7 and $<$ twice the sample mean, were identified using mosdepth v0.3.2. Fourfold-degenerate sites, where all possible nucleotide substitutions have no effect on the amino acid sequence, were identified using partition_cds.py (see Data availability). Biallelic SNPs within callable fourfold-degenerate sites were counted using bedtools v2.30.0. To calculate heterozygosity, SNP counts were divided by the total number of callable fourfold-degenerate sites for each individual.

2.4 Results

2.4.1 Genome assembly

We sequenced and assembled the genome of a male *Brenthis ino* individual collected in Asturias, Spain (SO_BI_364, Figures 2.1A and 2.1B). We generated 69.0-fold and 81.2-fold coverage of Pacbio CLR and Illumina WGS reads, respectively. The initial assembly consisted of 119 contigs and had a total length of 411.8 Mb, which is consistent with the kmer-based estimate of haploid genome size of 414.0 Mb (Figure A.1). HiC reads (11.7-fold coverage) from a male specimen collected at the same locality (SO_BI_375, Figures 2.1C and 2.1D) were used to scaffold the contigs into 14 chromosome-level sequences. These scaffolds range in size from 21.9 to 43.0 Mb and encompass 99.7% of the assembly. The contig and scaffold N50 of the assembly is 9.6 and 29.5 Mb, respectively.

The BUSCO score of the assembly is 99.0% (S:98.6%, D:0.4%, F:0.3%, M:0.7%), suggesting the assembly is missing very few single-copy insect orthologues and has little duplication. The estimated mean Phred quality score of the consensus sequence is 39.85.

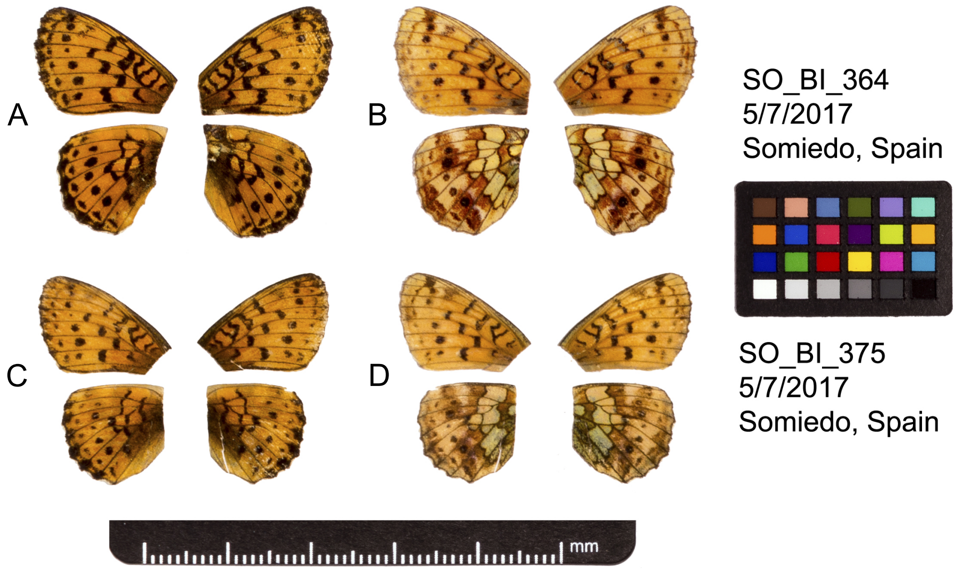

Figure 2.1: Fore and hind wings of the two *B. ino* individuals used to generate the genome sequence. **(A)** Dorsal and **(B)** ventral surface view of wings of specimen SO_BI_364, used to generate Pacbio and Illumina WGS reads. **(C)** Dorsal and **(D)** ventral surface view of wings of specimen SO_BI_375, used to generate HiC reads.

We assembled and annotated a circular mitochondrial genome of 15,180 bases with 13 protein coding genes, 22 tRNAs, and two rRNAs. The cytochrome oxidase subunit 1 (COI) nucleotide sequence has 99.85% identity (657/658 b) with a previously published COI sequence from a *B. ino* individual collected in Castilla y León, Spain (GenBank accession MN144802, Dapporto et al. 2019).

2.4.2 Evidence for a segregating neo-Z chromosome

While the HiC data support the scaffolding of 14 chromosome-level sequences (hereafter simply referred to as chromosomes), there is an excess of HiC contacts between chromosomes 11 and 13 (Figure 2.2A). This excess is not distributed evenly over the two chromosomes and is instead concentrated at one of the four possible junctions (Figure 2.2B), supporting the scaffolding of these two chromosomes in a specific orientation. However, while the number of HiC contacts between chromosomes 11 and 13 exceeds what we see between any other pair of chromosomes, it is below what we typically observe within chromosomes in this dataset (Figure A.2), making it unclear whether chromosomes 11 and 13 are fused and should be scaffolded together.

We tested whether the HiC contacts between chromosomes 11 and 13 are haplotype-specific, as this would result in half the number of contacts, and so could explain the reduced frequency (Figure A.2). Haplotype-specific HiC maps (see Methods) confirm that HiC contacts between chromosomes 11 and 13 are almost entirely limited to one haplotype (Figures 2.2C and 2.2D) and the proportions of haplotype-specific reads (49.6% and 50.4% of partitioned reads support haplotypes 1 and 2, respectively) are consistent with these chromosomes being fused in one haplotype but not the other.

We identified chromosome 11 as the Z-chromosome in *B. ino*: the female individual (Figure A.3) has half coverage for this chromosome, whereas the male used for assembly has full coverage (Figure A.4). By contrast, chromosome 13 has full coverage in both males and females (Figure A.4), consistent with the expectation for autosomal chromosomes (although see discussion). As one of these chromosomes is Z-linked, while the other has autosomal patterns of sex-specific coverage, we conclude that the individual from which we generated the HiC library must be heterozygous for a

Z-autosome fusion, i.e. a neo-Z chromosome.

The Pacbio reads, which were generated from SO_BI_364 rather than SO_BI_375, do not span the gap between chromosomes 11 and 13. However, it is still possible that SO_BI_364 does possess a copy of the neo-Z chromosome, if the fusion point is within a region of the genome that is too repetitive to be assembled and the gap is too large for successful chimeric alignment. It is therefore uncertain whether only SO_BI_375 possesses a copy of the neo-Z or if SO_BI_364 does as well.

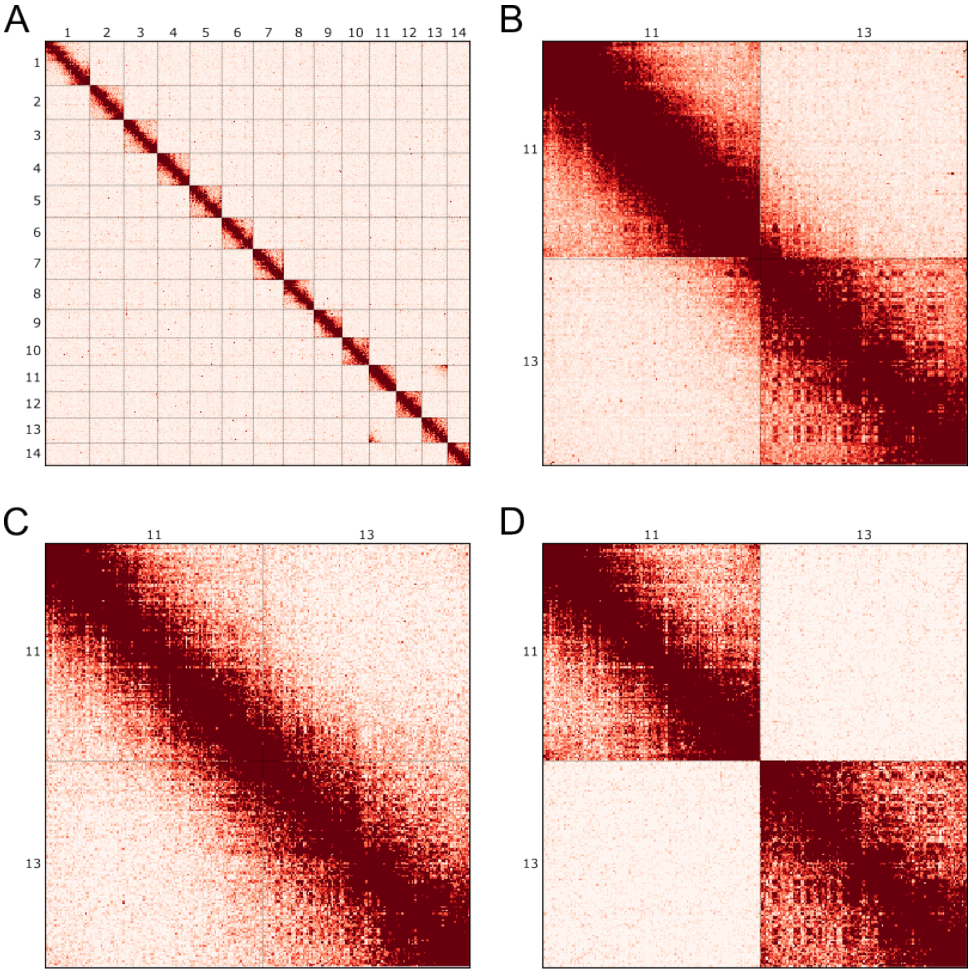

Figure 2.2: HiC contact heatmaps for the assembly of *B. ino*. **(A)** HiC contacts across all 14 chromosomes (HiC_view params: `-b 2500 -s 25`). **(B)** Contacts across chromosomes 11 and 13, with both chromosomes in the reverse orientation. (HiC_view params: `-b 250 -s 30`) **(C)** The same as in (B) but restricted to HiC reads containing alleles exclusively associated with haplotype 1 (HiC_view params: `-b 250 -s 10`). **(D)** The same as in (C) but associated with haplotype 2 rather than 1 (HiC_view params: `-b 250 -s 10`)

2.4.3 Synteny

We expect that the *B. ino* genome has been shaped by many chromosome fusions because it has a much lower chromosome number than other nymphalid butterflies. A pairwise comparison of synteny between *B. ino* and *Melitaea cinxia* shows that all *B. ino* chromosomes contain genes from multiple *M. cinxia* chromosomes (Figure 2.3). Additionally, nine *M. cinxia* chromosomes have genes distributed over multiple *B. ino* chromosomes (Figure 2.3). Because *M. cinxia* possesses the ancestral karyotype of nymphalid butterflies (Ahola et al. 2014), the differences in synteny observed in Figure 2.3 are all the result of rearrangements on the lineage leading to *B. ino*. These patterns of synteny therefore show that chromosome fusions, alongside fissions and/or reciprocal translocations, have shaped the *B. ino* genome.

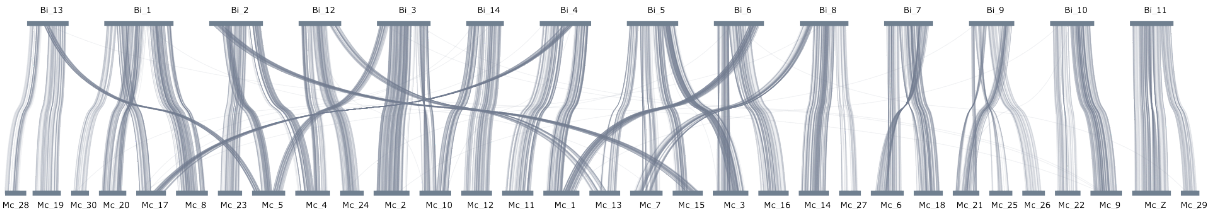

Figure 2.3: A synteny comparison between *B. ino* (top) and *M. cinxia* (bottom). Each line connects the same BUSCO gene in either genome assembly. Chromosomes are ordered to minimise the number of lines that cross one another. The correspondence between *M. cinxia* and *B. ino* chromosomes can only be explained by chromosome fusions alongside fissions and/or reciprocal translocations.

2.4.4 Genome annotation

We annotated 16,844 protein coding genes. Given this annotation, we estimate that 33.5% of the genome assembly is intronic and 5.6% exonic. Chromosomes display some variation in gene density; chromosome 14, the shortest and most gene poor, is 32.8% genic whereas chromosome 11 (the Z) is 47.7% genic. Across the annotation, the median length of genes, introns, and exons is 4,084, 616 and 148 b, respectively (Figure A.5).

Transposable elements (TEs) compromise 37.9% of the genome (Table A.2, Figure 2.4A). Most TE activity appears to be relatively recent, as a large proportion of repeats exhibit a low genetic distance from their respective consensus sequences (Figure 2.4B). The genome contains all major TE types (Table A.2). Rolling circle elements,

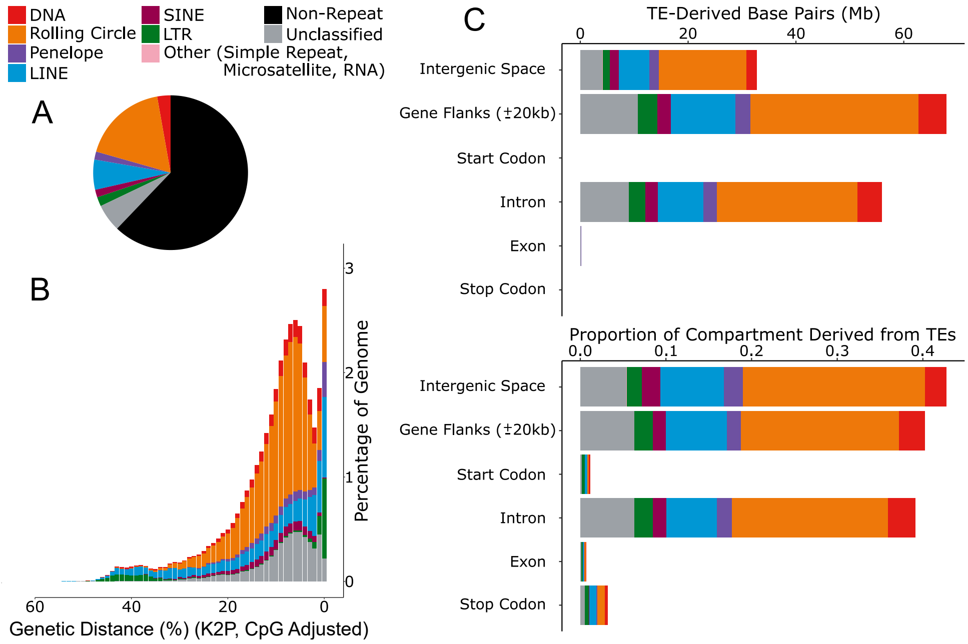

Figure 2.4: TEs within the genome assembly of *B. ino*. **(A)** The proportion of the assembly comprised of the main TE classifications, as represented by the colours in the key. **(B)** A repeat landscape plot illustrating the proportion of repeats in the genome at different genetic distances (%) to their respective RepeatModeler consensus sequence. Greater similarity to consensus (i.e. lower genetic distance) is suggestive of recent activity. **(C)** The abundance of TEs in different partitions of the genome, shown in bases and as a proportion of the partition.

also known as helitrons, appear to have been the most successful progenitors within the genome, accounting for 17.8% of total genome length, and $\sim$ 47% of total TE content (Table A.2). There is also evidence of very recent activity in LINEs and LTR elements, with a sharp increase in the number of identified elements with very low genetic distance to their consensus sequences (Figure 2.4B). The reasons for the bursts in LINEs and LTRs are unknown, although the likely recent age of these insertions is consistent with recent host colonisation, potentially via horizontal transposon transfer (HTT) from another host genome (Ivancevic et al. 2018; Gilbert et al. 2010; Wallau et al. 2012).

Considering all TE classifications, most TEs are found outside of genes (Figure 2.4C). Gene flanks and introns have a similar density of TEs, whereas intergenic space has a slightly higher density (Figure 2.4C). Exons are largely devoid of TE sequence, with only 0.7% of exonic sequences consisting of TEs. This is to be expected given the likely detrimental effects of TE insertions in host exons (Bourque et al. 2018; Sultana et al. 2017). The most abundant TEs in the genome, rolling circle elements, comprise

21.3% of intergenic space, $\sim$ 18% of gene flanks and intronic regions, and just 0.1% of exonic regions (Figure 2.4C).

Satellite repeats are found immediately adjacent to the putative neo-Z fusion point. Chromosome 11 starts with a 5.8 kb array of repeats (RND-5_FAMILY-919) and chromosome 13 ends in a 10.9 kb array (RND-6_FAMILY-6270). The array on chromosome 11 consists of repeat units of $\sim$ 110 bases, whereas the array on chromosome 13 has larger repeat units of $\sim$ 325 bases. We conclude that, due to a lack of similarity, these repeats are unlikely to have facilitated a non-homologous recombination event that led to the neo-Z fusion.

2.5 Discussion

We have resolved the sequences of 14 *Brenthis ino* chromosomes: 13 autosomes and the Z sex-chromosome. The number of chromosomes in the assembly is higher than previously reported for *B. ino* in Europe (Federley 1938; Saitoh 1987, 1991), but equal to counts reported for this species in Japan (Maeki and Makino 1953; Saitoh et al. 1989). We note that previous karyotype data from Europe were all from Scandinavian samples, whereas the individuals contributing to the assembly were collected in Spain. Scandinavian populations of *B. ino* may therefore have a high frequency of the neo-Z fusion that we report or other chromosome fusions that are not identifiable in our data.

We have interpreted the excess of HiC contacts between chromosomes 11 and 13, as well as the stark contrast in haplotype-specific HiC maps, as strong evidence for a segregating neo-Z chromosome. Lab contamination from a closely related - but karyotypically divergent - species is not a plausible alternative explanation given that the haplotype partitioned HiC reads are approximately equal in frequency (see Results). We can also rule out the possibility that we sampled an admixed individual, e.g. an F1 between *B. ino* and its sister species *B. daphne*, and that the neo-Z is fixed in one species but absent in the other. Both species are present in Northern Spain, so sampling an F1 is possible, at least in principle. However, if SO_BI_375 were a recent hybrid, we would expect its heterozygosity to be considerably elevated compared to other *B. ino* individuals which is not the case: heterozygosity at autosomal fourfold degenerate sites for SO_BI_375, SO_BI_364, and FR_BI_1497, is 0.0108, 0.0106, and 0.0100, respectively, and in all cases is far lower than we would expect for an F1 between *B. ino* and *B. daphne* ($\sim$ 0.025, Ebdon et al. 2021).

Because we have only observed evidence for the neo-Z in one individual, we do not know its frequency in the wider *B. ino* population. This rearrangement could be restricted to certain populations, or it may have evolved so recently that it is only found in a small number of closely related individuals. One way to estimate the frequency of the neo-Z would be to test whether any females have half the normalised coverage over both chromosomes 11 and 13; which would be consistent with a single copy of the neo-Z (chromosomes 11 and 13 fused together), a W chromosome, but no additional copy of chromosome 13. However, if chromosome 13 is yet to evolve a dosage compensation mechanism, females carrying the neo-Z may only be viable with two copies of the autosomal sequence. Under this scenario, the female coverage seen in Figure A.4 is consistent with both presence or absence of the neo-Z chromosome. Population level cytological or HiC data would be required to estimate the frequency of the neo-Z and understand its evolutionary history.

While we have mainly focused on karyotypic variation within a single individual, we have also shown that the *B. ino* genome has a complex rearrangement history that includes many fusions as well as fissions and/or reciprocal translocations (Figure 2.3). The assembly therefore provides an opportunity to test the causes and consequences of chromosome rearrangements more widely. Additonally, the assembly will enable population genomic studies in the genus *Brenthis*, expanding on previous reference-free analyses (Pazhenkova and Lukhtanov 2019; Ebdon et al. 2021). More generally, it adds to a growing number of high quality resources for comparative genomics in the Lepidoptera.

2.6 Data availability

Table A.1 contains the metadata for the four individuals used for this project. The genome assembly, gene annotation, and raw sequence data can be found at the European Nucleotide Archive under project accession PRJEB49202. The scripts used for analysing HiC data (chomper.py and HiC_view.py), the script used for calculating site degeneracy (partition_cds.py), and the script used for visualising synteny (busco2synteny.py) can be found at the following github repository: https://github.com/A-J-F-Mackintosh/Mackintosh_et_al_2022_Bino. The mitochondrial genome sequence and the TE annotation can be found at the same repository.

Inferring inter-chromosomal rearrangements and ancestral linkage groups from synteny

3.1 Abstract

Chromosome rearrangements shape the structure of the genome and influence evolutionary processes. Inferring ancestral chromosomes and rearrangements across a phylogenetic tree is therefore an important analysis within evolutionary genetics. One approach to this inference problem is to focus on synteny information, i.e. the co-occurrence of loci on the same chromosome. Although algorithms for inferring ancestral linkage groups (ALGs) and inter-chromosomal rearrangements from synteny have been previously described, they have seldom been applied to modern genome data. Here we implement these algorithms in a command-line tool, `syngraph`, and evaluate their performance using simulations that include a mix of different rearrangements and types of error. We show that ALGs and rearrangements can be recovered when the rearrangement frequency per-branch is well below the number of chromosomes. We demonstrate that competing models of rearrangement can be inferred by comparing observed results to simulations. Finally, we reanalyse genome assemblies of rhabditid nematodes and find that independent fusions of the same ALGs pose a challenge that is difficult to overcome without gene-order information. Our simulations and analysis of real data demonstrate both the promise and limitations of using synteny information to infer patterns of genome evolution.

3.2 Introduction

The fact that the genomes of different organisms vary in chromosome number and structure has long been appreciated (Robertson 1916; Sturtevant 1921). Changes in ploidy explain at least some of this variation, especially in plants (Otto and Whitton 2000). Chromosome rearrangements are another mechanism by which genomes undergo large-scale changes, and they are common across eukaryotes (Zhao and Schranz 2019; Li et al. 2022; Muffato et al. 2023). Intra-chromosomal rearrangements involve a single chromosome (e.g. inversions) while inter-chromosomal involve two chromosomes (e.g. fissions, fusions, and translocations). Chromosome rearrangements can influence fundamental evolutionary processes, such as recombination (Bidau et al. 2001; Näsvall et al. 2023), as well as broader processes like speciation (Yoshida et al. 2023; Mackintosh et al. 2023), so there is considerable interest in reconstructing how genomes have rearranged through time. Ancestral karyotypes and rearrangements have been estimated for a number of different taxa, including *Drosophila*, ruminants, and birds (Muller 1940; Farré et al. 2019; Damas et al. 2018), as well as large taxonomic groups such as rosid plants and animals (Murat et al. 2015; Simakov et al. 2022). Typically, such analyses rely on sequence or linkage map data for tens of genomes (or less), but it is becoming more common to analyse hundreds of chromosome-level genome assemblies (Muffato et al. 2023; Wright et al. 2023), highlighting the need for efficient inference methods.

Estimating ancestral genomes and rearrangements given a set of present-day genomes and a phylogenetic tree is a challenging combinatorics problem (Sankoff 2003; Fertin et al. 2009). Methods for inferring ancestral genomes are typically either event-based or adjacency-based (Feng et al. 2017). Event-based methods use an explicit model of rearrangement and aim to construct maximally parsimonious ancestral genomes at internal nodes by minimising the number of rearrangements across the tree (e.g. Bourque and Pevzner 2002; Zheng and Sankoff 2011). By contrast, adjacency-based methods reconstruct ancestral genomes by assuming that genome structure that is conserved in present-day genomes was also present in their most recent common ancestor, without explicitly modelling individual rearrangements (e.g. Kim et al. 2017; Muffato et al. 2023). The event-based approach has the advantage of co-estimating ancestral genomes and rearrangements, and so gives direct insight into the evolutionary process, but comes at a computational cost. These approaches can therefore be viewed as complementary, with adjacency-based analyses being the most practical way to summarise patterns of genome evolution from hundreds of genomes (Muffato et al. 2023) and event-based approaches being a better choice for detailed rearrangement inference from a handful of genomes (Ostevik et al. 2020).

Genome evolution through time can be reconstructed at different resolutions; the most detailed being the full reconstruction of ancestral sequences. However, co-estimating base and indel substitutions and chromosome rearrangements requires accurate whole genome alignments and therefore considerable computational resources (Armstrong et al. 2020). An alternative is to instead infer the order of sequences in ancestral genomes without considering substitutions. This task is straightforward when gene-order is well conserved, but becomes more challenging when the intra-chromosomal rearrangement rate is high or if taxa are very distantly related (Farré et al. 2019; Muffato et al. 2023). Genome evolution can also be reconstructed at the level of synteny (Fertin et al. 2009). While the term synteny is sometimes used to denote co-linearity between chromosomes, here we use it to refer to the co-occurrence of loci on the same chromosome, regardless of their order (Renwick 1971; Passarge et al. 1999). Focusing exclusively on synteny means that only unordered sets of markers (i.e. linkage groups) and inter-chromosomal rearrangements can be reconstructed. Despite this limitation, methods that focus on synteny are likely to be applicable across a wide range of datasets, as synteny decays more slowly than gene-order across evolutionary time (Simakov et al. 2022). An efficient synteny-based method for reconstructing genome evolution would therefore allow for the processes that constrain and promote fission, fusion and translocation rearrangements to be investigated across many groups of species.

Although algorithms exist for reconstructing genome evolution from synteny (Ferretti et al. 1996; DasGupta et al. 1997; Liben-Nowell 2001), they have rarely been applied to data. Additionally, it is not clear how well these methods perform under different rearrangement scenarios or how well they can accommodate the types of error that exist in genome assemblies and annotations. Here we address these issues by implementing previously described algorithms in a command-line tool – `syngraph` – and performing analyses on both simulated and real data. The manuscript is structured as follows: First, we briefly recapitulate previously described synteny-based algorithms for inferring inter-chromosomal rearrangements between genomes and across phylogenies. Next, we evaluate the performance of these methods on data simulated under a range of rearrangement and error parameters. Finally, we reanalyse a set of nematode genomes from Gonzalez de la Rosa et al. (2021) and compare our results to theirs.

3.3 Results

3.3.1 Inferring inter-chromosomal rearrangements between two genomes

Ferretti et al. (1996) outlined the problem of calculating a syntenic edit distance between two genomes (hereafter simply referred to as syntenic distance). This is the minimum number of fissions, fusions and translocations required to transform one genome into another (Figure 3.1). The entire genome sequences are not required, instead markers present in both genomes only need to be assigned to chromosomes in each and positional information (i.e. marker order) is ignored. A marker can be any single-copy sequence feature that is identifiable across both genomes, e.g. ultra-conserved elements, genes or nucleotide alignments. Given this information, Ferretti et al. (1996) showed that the problem of transforming genome G_A to genome G_B by rearrangement can be reduced as follows: write each chromosome of G_A as a set populated by the labels of chromosomes in G_B with which it shares markers. Then, G_A must be rearranged through the smallest number of set operations such that the final sets are all unique and of length one (therefore each representing a single chromosome from G_B). In this scheme, fusions are unions of two sets and fissions replace one set with two disjoint sets (Ferretti et al. 1996). Additionally, a translocation can be modelled as an exchange of subsets. In principle, these set operations can be combined to find a maximally parsimonious series of rearrangements. Calculating the syntenic distance between two genomes therefore gives (i) a measure of how rearranged they are from one another and (ii) a putative history of the rearrangements between them (Figure 3.1).

We implemented two heuristic algorithms for calculating a syntenic distance between two genomes: the algorithm from DasGupta et al. (1997) where only fission and fusion rearrangements are permitted (FF) and a modified version of the algorithm from Ferretti et al. (1996) which allows fission, fusion and translocation (FFT) (Figure 3.1). While Ferretti et al. (1996) include the possibility of non-reciprocal translocations in their algorithm, we only consider reciprocal translocations where portions of both chromosomes are exchanged. The FF algorithm is straightforward: chromosomes of G_A that share markers syntenic in G_B are fused recursively, then fissions are implemented

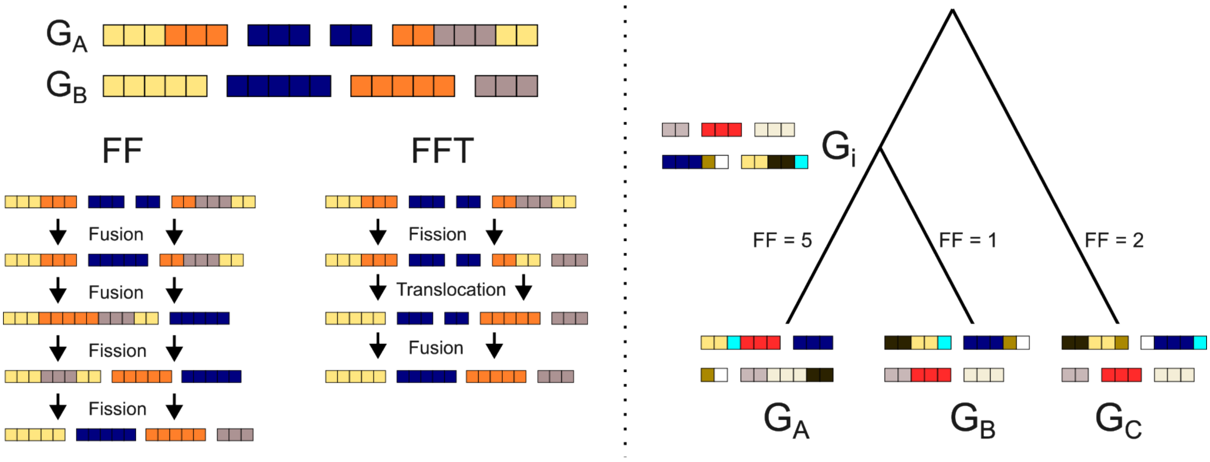

Figure 3.1: Inferring rearrangements between two genomes and across a phylogenetic tree. The left panel shows two genomes, G_A and G_B, where chromosomes (rectangles) consist of markers (squares), and the order of markers within a chromosome is arbitrary. Each marker in G_A is coloured by the chromosome that it is found on in G_B. Given this representation of G_A, the FF and FFT algorithms can be used to transform it to G_B, thus calculating a syntenic distance between G_A and G_B. The right panel shows three genomes, G_A, G_B and G_C, that are related by a phylogenetic tree. Here markers are coloured by the synteny-set that they are a part of, where a synteny-set is defined as a group of markers that are syntenic in genomes G_A, G_B and G_C. A genome at the internal node of the tree, G_i, can be constructed by combining synteny-sets and evaluated by summing the syntenic distances between G_i and the other three genomes. Branches of the phylogeny are labelled with the syntenic distance under the FF algorithm given the genome at G_i.

to recover the individual chromosomes of G_B (Figure 3.1). The FFT algorithm uses similar criteria for implementing fissions and fusions, but will also implement a translocation when two chromosomes share multiple sets of markers syntenic in G_B (Figure 3.1). Given that this algorithm is more complex, we do not describe the details here and instead provide a description in the Supplementary Methods.

We used simulations to test whether the syntenic distances estimated by these algorithms correspond to the true number of total rearrangements. More specifically, we simulated an initial genome containing 1,000 markers uniformly distributed among k chromosomes and then generated a second genome by rearranging the first r times. Note that here we only simulated the types of rearrangement considered by the respective algorithms. We performed simulations for three values of k (10, 20 and 30) and varied r between 1 and 50. We find that both algorithms accurately estimate the total number of rearrangements when $r <= k$ (Figure 3.2). Above this, both algorithms infer syntenic distances that are underestimates of the total number of rearrangements (Fig-

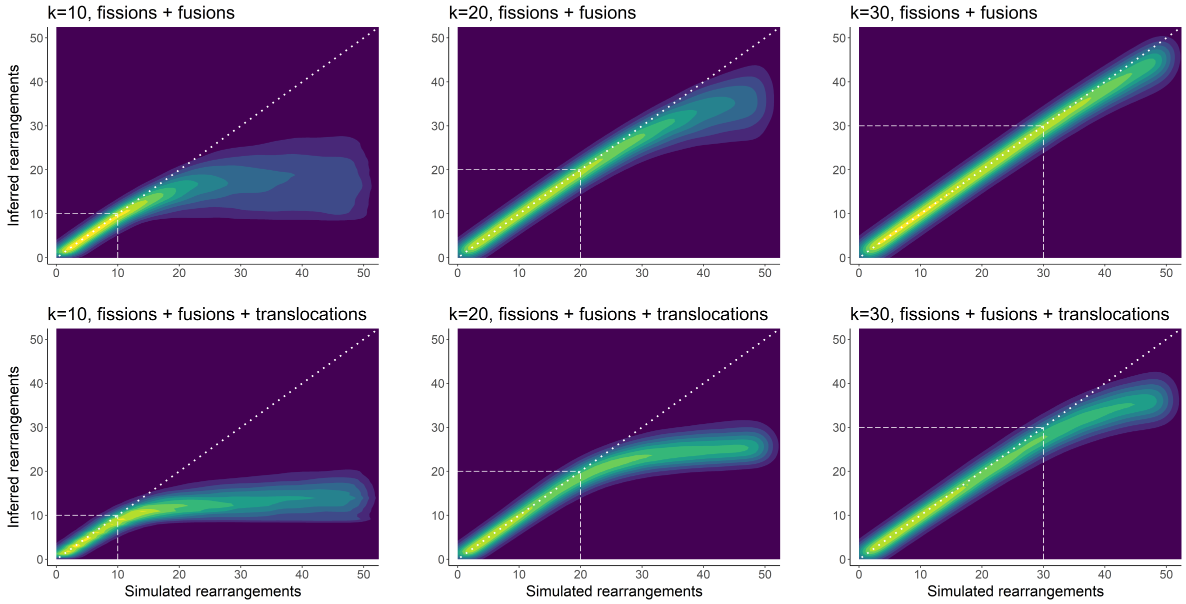

Figure 3.2: Estimating the total number of rearrangements between two genomes using syntenic distance. Each plot summarises results from 50,000 simulations, with brighter colour corresponding to a greater density of simulated data points. The total number of simulated rearrangements is plotted on the x-axis with the inferred number on the y-axis. The dotted white line along the diagonal ($x = y$) corresponds to inferred and simulated counts being equal, i.e. correct inference. Plots show simulations with $k = 10$ (left), $k = 20$ (middle), or $k = 30$ chromosomes (right). Additionally, plots either show simulations containing only fissions and fusions (top) or fissions, fusions and translocations (bottom). Dashed lines in each plot show where the number of rearrangements is equal to k.

ure 3.2). This bias is much more pronounced when simulations and inference include translocation, although this is expected given that a series of fissions and fusions can sometimes be explained by a single translocation. Put differently, histories that contain all three types of rearrangements are more challenging to estimate. Nonetheless, these results suggest that the syntenic distance between two genomes is a useful approximation of the true number of total rearrangements, as long as the total number of rearrangements is less than the number of chromosomes.

3.3.2 Inferring ALGs and inter-chromosomal rearrangements across a phylogeny

The syntenic distance between genomes can be used to reconstruct ancestral linkage groups (ALGs) and inter-chromosomal rearrangements across a phylogenetic tree (Ferretti et al. 1996; DasGupta et al. 1997). Consider a triplet of related genomes, G_A, G_B and G_C, where G_A consists of chromosomes A_1, A_2, ..., A_n with arbitrary labels. Any marker present in all three genomes can be written as the chromosomes it is found on, e.g. [A_4, B_{22}, C_9]. Markers that are syntenic in all three genomes will have the same notation as each other and can be considered as part of the same synteny-set (Figure 3.1). These synteny-sets can be used as building blocks to reconstruct the ALGs of G_i (the ancestral genome at the internal node of the tree relating G_A, G_B and G_C) (Figure 3.1). An adjacency-based approach to ALG reconstruction is to initiate a LG with a single synteny-set and add all synteny-sets that are syntenic with it in two of the triplet genomes (i.e. those that share two of the three elements in their notation), and repeat until all synteny-sets are part of a LG. An event-based approach is to build many different versions of ALGs from synteny-sets under more relaxed rules (e.g. synteny-sets in the same LG are allowed to share just one element in their notation) and then identify the most parsimonious set of ALGs using the sum of syntenic distances between G_i and G_A, G_B, G_C. Once a set of ALGs is obtained for G_i (by either method), rearrangements can be recorded between G_i and its descendants (G_A and G_B) using the FF or FFT algorithms.

To assess these approaches to ALG estimation, we performed simulations over a phylogenetic tree with $n = 3$ leaves (see Methods for details). The genome at the root of the tree always had 20 chromosomes and 1000 markers (these parameters are used for all subsequent simulations), and the number of rearrangements simulated across the tree varied between 1 and 50. Given the synteny of markers in the three sampled genomes, we estimated ALGs and rearrangements with the approaches described above and calculated two performance metrics:

- **ALG accuracy**: The proportion of markers within correctly estimated ALGs, averaged across all internal nodes of the phylogeny (unless stated otherwise).
- **Rearrangement accuracy**: The proportion of branches across the tree with the correct number of estimated rearrangements by type, e.g. 1 fission, 2 fusions, 1 translocation.

We find that both ALG and rearrangement accuracy are greater when simulations and inference only contain fissions and fusions (Figure 3.3). Both metrics decline with the number of simulated rearrangements, but rearrangement accuracy does so faster (Figure 3.3). This is expected given that poor rearrangement estimation will

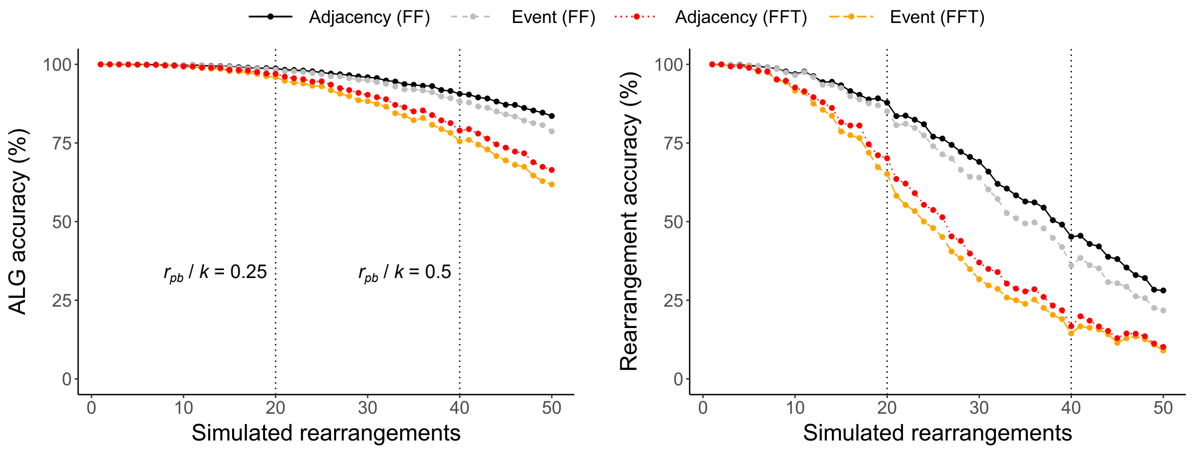

Figure 3.3: The accuracy of inferred ALGs and rearrangements across a tree with n = 3 leaves. The left plot show the accuracy of ALGs inferred from simulations with between 1 and 50 rearrangements. The right plots show the accuracy of inferred rearrangements. Points always represent averages from 1000 simulations. ALG estimation was either performed using an adjacency or event-based approach. Simulations, and inference, either include fissions and fusions (FF) or fissions, fusions, and translocations (FFT). Dotted vertical lines show the ratio between the number of rearrangements per-branch and the expected number of chromosomes at values of 0.25 and 0.5.

only typically involve a subset of ALGs, meaning that some ALGs (e.g. those that are invariant across the tree) can still be well estimated when rearrangements are not. These simulations also show that rearrangements across the tree are only well estimated when their frequency per-branch is well below the number of chromosomes (e.g. $r_{per-branch}/k \approx 0.25$). Surprisingly, we find that the event-based approach often gives worse results than the simpler and quicker adjacency-based approach (Figure 3.3). We therefore use adjacency-based ancestral genomes for all subsequent analyses.

The approach outlined for a phylogeny with n = 3 leaves can be extended to arbitrarily large trees. To do this, nodes are visited through a post-order traversal (from leaves to root) and ancestral genomes are reconstructed using a triplet of genomes (either observed genomes at the leaves or already inferred ancestral genomes). This process continues until the ancestral genomes at all internal nodes have been estimated. Given a large phylogeny (n = 100 leaves), we tested how well ALGs are estimated at a deep node in the tree (the child-node of the root with the most descendants). When the total number of fission and fusions rearrangements simulated was 198, 594 and 990, corresponding to an average of 1, 3 and 5 rearrangements on each of the $2n-2$

branches, ALG accuracy (under the FF algorithm) was 98.53%, 67.48%, and 13.94%, respectively. These performance estimates are far worse than analogous ones for a tree with only $n = 3$ leaves (99.98%, 99.65% and 98.72% for 1, 3 and 5 rearrangements per-branch, respectively). This shows that errors in ALG estimation accumulate upwards through the tree, especially when the rearrangement rate is high.

3.3.3 The effect of marker error

We have so far only evaluated the performance of ALG and rearrangement inference using simulations that produce perfect data. Real genomes sequences and annotations, however, are likely to contain errors. For example, some markers will be missing from certain genomes and some markers may be assigned incorrect orthology. We therefore added these sources of error to simulations.

We implemented a method to assign markers that are missing in a minority of genomes to ALGs (see Methods). We then tested this method by modifying our simulations so that genomes at the leaves of the tree have a small number of missing markers. We simulated fissions and fusions over a tree with $n = 10$ leaves and inferred them back under the FF algorithm. We find that, even when the amount of missingness is small (e.g. 5% of markers per-genome), ALG accuracy is significantly reduced (Figure 3.4). By contrast, missingness has a negligible effect on whether rearrangement histories are estimated accurately (Figure 3.4). This disparity can be explained by the fact that missingness removes information about individual markers from the data, which effects our ability to estimate all of the markers within an ALG correctly but not our ability to identify whether, for example, a single fusion rearrangement has happened on a particular branch. Given the sensitivity of our ALG accuracy measure to missingness, we considered an alternative metric:

- **Fuzzy ALG accuracy**: The proportion of markers within fuzzily estimated ALGs, averaged across all internal nodes of the phylogeny (unless stated otherwise). A fuzzy ALG contains at least 90% of markers from a single true ALG and no more than 5% of markers from any other.

We find that fuzzy ALG accuracy is generally high at all levels of missingness (Fig-

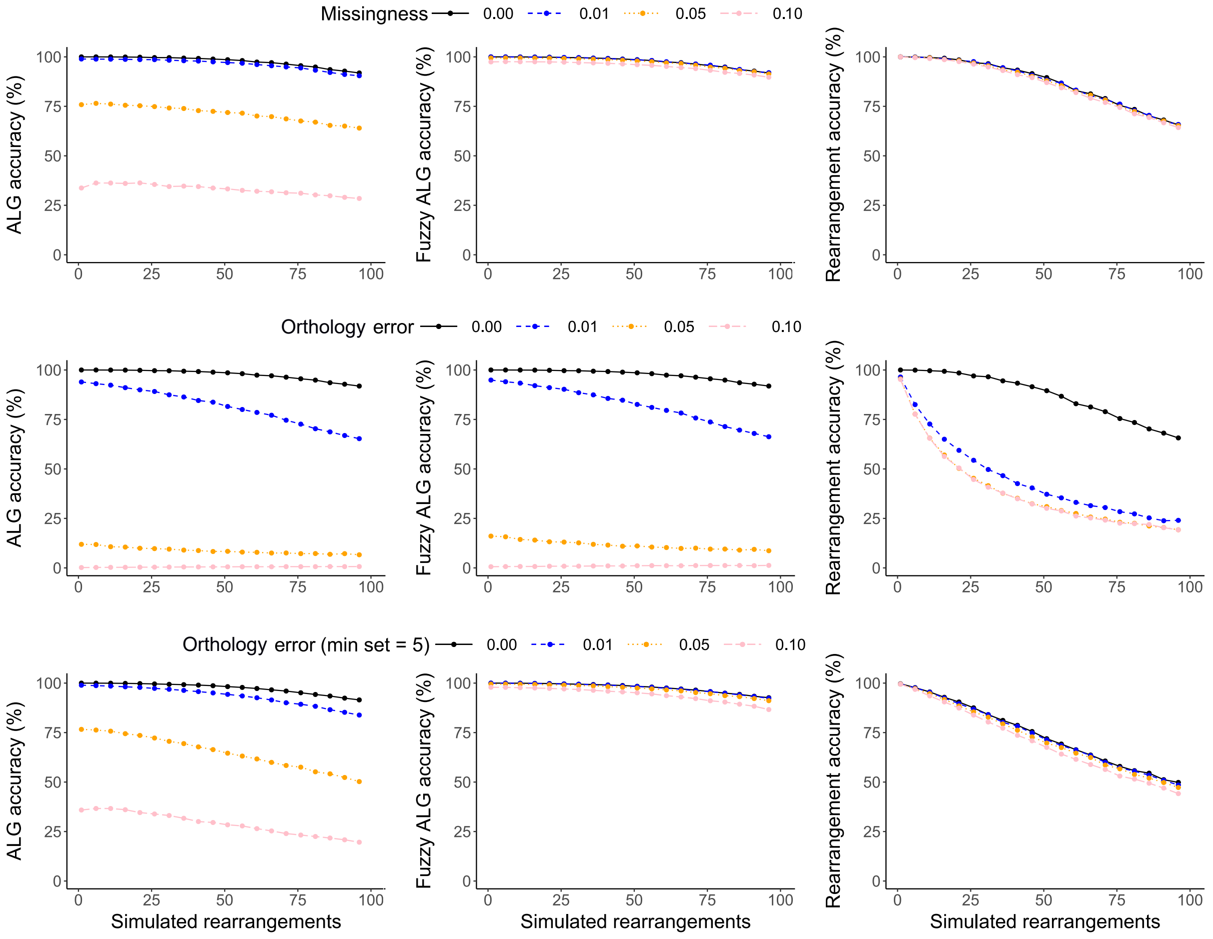

Figure 3.4: The effect of marker error and rearrangement frequency on inference accuracy. Each plot shows how the number of rearrangements (x-axis) affects a performance metric (y-axis). These metrics include ALG accuracy (left column), fuzzy ALG accuracy (middle column), and rearrangement accuracy (right column). The top row of plots include simulations with varying levels of marker missingness. The middle row of plots include different levels of marker orthology error, whereas the bottom row include the same levels of orthology error but inference was performed with a minimum synteny-set size of 5. Points always represent averages from 1000 simulations.

ure 3.4). Rearrangements and good approximations of ALGs can therefore still be estimated despite missingness.

We next added orthology error to simulations by randomising the chromosome assignment of a small proportion of markers. We again simulated rearrangements across a tree with n = 10 leaves and find that, perhaps unsurprisingly, orthology error has a large effect on all performance metrics (Figure 3.4). In particular, it becomes difficult to estimate rearrangement histories as erroneous rearrangements are introduced to account for the movement of markers (Figure 3.4). These inferred rearrangements will involve a much smaller number of markers than fission, fusion, or translocation events which typically involve large portions of chromosomes. We therefore set a minimum synteny-set size for inferring ALGs and rearrangements in an attempt to improve performance (see Methods). Setting the minimum synteny-set size to five markers resulted in large improvements in fuzzy ALG and rearrangement accuracy (Figure 3.4). However, even without error, the addition of a minimum synteny-set size reduces the accuracy of estimated rearrangement histories (Figure 3.4). This is presumably due to an inability to identify fissions that involve a small number of markers, and a similar difficulty in reconstructing complex rearrangement sequences that result in small sets of syntenic markers. These simulations show that a small amount of orthology error can be overcome by introducing a minimum set size, although this itself does have a performance cost.

3.3.4 Evaluating evidence for translocations through simulation

Only fission and fusions rearrangements are inferred when using the FF algorithm, even if the true rearrangement history contains translocations. The FFT algorithm, by contrast, allows inference of all three types of rearrangement, but it is not clear how often it erroneously infers a series of fissions and fusions as a translocation, or vice versa. We therefore investigated whether inference under the FFT algorithm recovers the correct ratio of fission, fusion and translocation events. We again simulated rearrangements over a tree with $n = 10$ leaves. We varied the number of rearrangements as well as the ratio of fission, fusion and translocation events (1:1:0, 1:1:1 or 1:1:2), then performed inference under the FFT algorithm. When only fission and fusion rearrangements are simulated (ratio 1:1:0), a small number of translocations are still inferred (Figure 3.5). For example, when 50 fissions / fusions are simulated across the tree, an average of 4.1% of inferred rearrangements are translocations, with 95% confidence intervals (95% CIs) of 0.0 and 12.2%. This shows that although the false-positive rate is generally low, rearrangement histories that include only fissions and fusions can result in inferred histories where $\sim$ 10% of rearrangements are translocations. When translocations are included in simulations with ratio 1:1:1 or 1:1:2, they are inferred at the expected frequencies of 33.3% and 50.0%, albeit with a downwards bias that increases with the number of simulated rearrangements (Figure 3.5). Although inferred rearrangements across many simulations do tend to reflect the un-

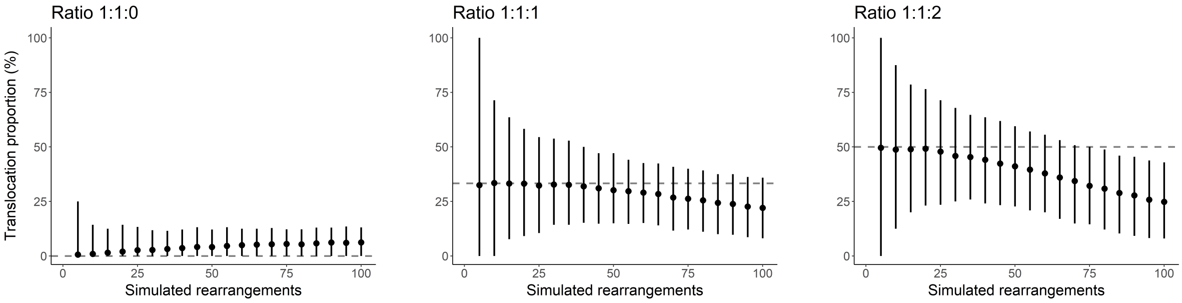

Figure 3.5: The proportion of rearrangements inferred as translocations for simulations across a tree with $n = 10$ leaves. Each plot shows the mean translocation proportion and 95% CIs (y-axis) for simulations involving different numbers of rearrangement (x-axis). The fission:fusion:translocation ratios of simulated rearrangements are 1:1:0 (left), 1:1:1 (middle), and 1:1:2 (right). Dashed horizontal lines correspond to the expected translocation proportion under each rearrangement ratio.

derlying rearrangement ratio, the wide variation among simulation replicates (Figure 3.5) shows that a single analysis only contains limited information about the relative rates of fission, fusion and translocation.

3.3.5 Rhabditid nematodes and Nigon elements

We implemented the methods described above for inferring ALGs and rearrangements in a command line tool, `syngraph`. Here we apply `syngraph` to genomes of nematodes in the order Rhabditida. Species in this order typically possess a small number of chromosomes ($\sim$ 6), albeit with some exceptions (Table B.1). Pairwise comparisons between genomes have shown that gene-order is highly variable across species (Lee et al. 2003; Hillier et al. 2007; Stevens et al. 2020). By contrast, synteny is more conserved between genomes and has therefore been used to identify seven ancestral linkage groups (ALGs), often referred to as Nigon elements, as well as inter-chromosomal rearrangements (Tandonnet et al. 2019; Gonzalez de la Rosa et al. 2021). We chose to reanalyse 14 rhabditid nematode genomes from Gonzalez de la Rosa et al. (2021). They used a clustering algorithm to identify ALGs and then manually inferred fissions and fusions across the tree (Gonzalez de la Rosa et al. 2021). By contrast, the synteny-based method we focus on aims to co-infer ALGs and rearrangements across a phylogeny in a single automated analysis.

We used BUSCO genes as orthologous markers for measuring synteny. After excluding non-chromosome-level sequences, as well as Y chromosomes, the number

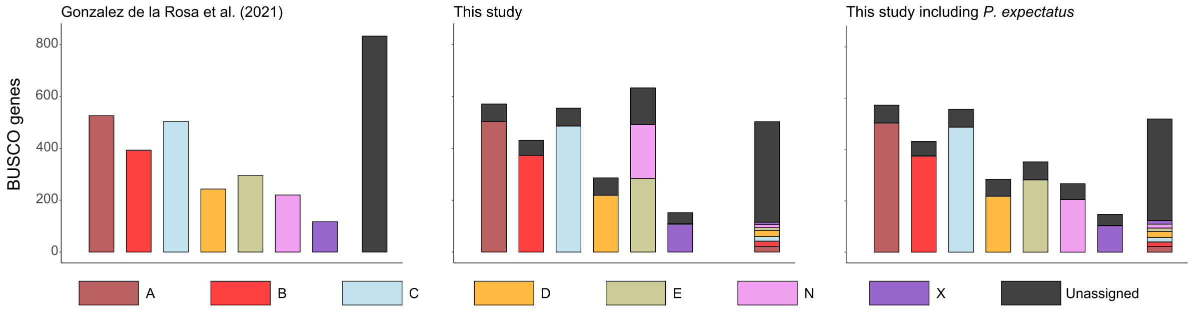

Figure 3.6: Ancestral linkage groups at the most recent common ancestor of Rhabditina and Tylenchina nematodes. Each plot shows ALGs as stacks of BUSCO genes, with the furthest right stack consisting of genes that were not assigned to any ALG. Stacks are coloured by the ALGs (i.e. Nigon elements) defined in Gonzalez de la Rosa et al. (2021) (left and legend below). Reconstructed ALGs in this study (middle) are similar to those of Gonzalez de la Rosa et al. (2021) but include a fusion of Nigons E + N. This fusion is not present, however, when including the genome of *Pristionchus exspectatus* in the analysis (right). Our method assigns more markers to ALGs than the clustering method of Gonzalez de la Rosa et al. (2021).

of BUSCO genes per-assembly ranged from 1820 in *Strongyloides ratti* to 3103 in *Caenorhabditis briggsae*, with only 961 being shared by all 14 genomes. We included all BUSCOs genes in the analysis, regardless of missingness.

We next inferred rearrangements and ALGs across the phylogeny using the FF algorithm. The number of ALGs at each internal node varied between four and seven and we inferred a total of 16 fusions and 30 fissions. In contrast to Gonzalez de la Rosa et al. (2021), our analysis suggests the existence of six ALGs at the deepest node in the tree (the most recent common ancestor of Rhabditina and Tylenchina). These ALGs correspond to Nigon elements A, B, C, D and X with E + N fused (Figure 3.6). Another key difference between our results and those of Gonzalez de la Rosa et al. (2021) is that we infer a single fusion event of Nigons N + X in the ancestor of *Caenorhabditis sp.* and *Haemonchus contortus* (Figure 3.7), whereas they suggest that these ALGs fused independently in each of these lineages.

The differences mentioned above can be explained by two limitations of our method. The first is that ancestral genomes are reconstructed locally, using information from only three other genomes at a time rather than the entire dataset. The reconstruction of Nigons E + N as one ALG deep in the tree, for example, can be explained by the fact that these elements co-occur on chromosomes of *Pristionchus pacificus* and *S.*

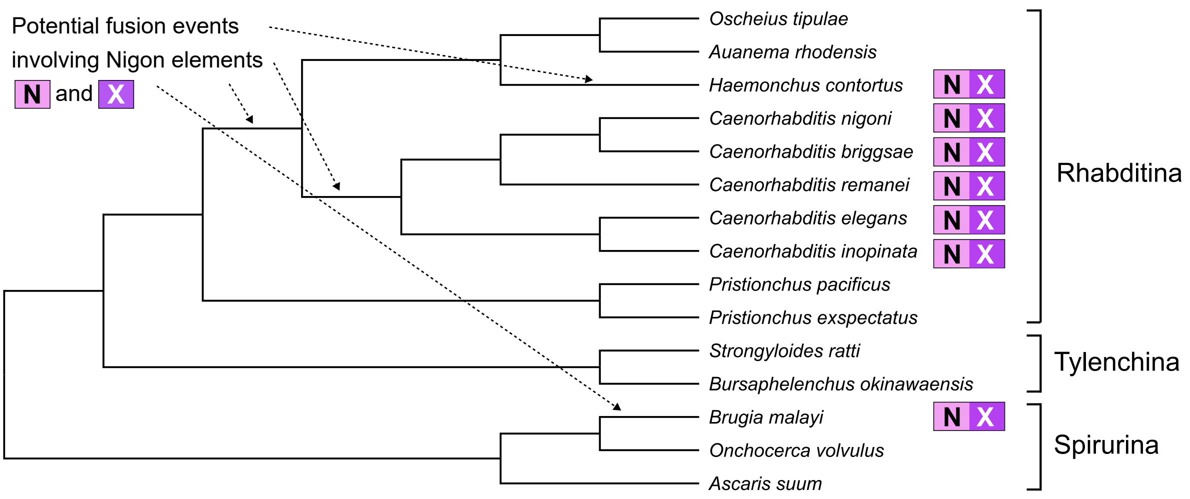

Figure 3.7: A tree showing the phylogenetic relationships between the 14 rhabditid nematode species analysed by Gonzalez de la Rosa et al. (2021), as well as *P. exspectatus*. The topology is from Gonzalez de la Rosa et al. (2021) and the branch lengths are arbitrary. Species with genomes where Nigon elements N + X are fused together are marked with a symbol to the right of the tree. Arrows point to branches where Nigon elements N + X could have fused.

ratti. Local reconstructions that rely on these genomes will therefore place Nigons E + N together, even though this necessitates multiple fission events on other branches of the phylogeny. The second limitation is that gene-order information is not used. The single fusion of Nigons N + X is supported by synteny but an examination of gene-order (as was done in Gonzalez de la Rosa et al. 2021) shows that these elements are well ’mixed’ in *Caenorhabditis sp.* but much less so in *H. contortus* (see Figure 5 of Gonzalez de la Rosa et al. 2021). This is consistent with two independent fusion events.

Recently, Yoshida et al. (2023) investigated the evolution and consequences of chromosome fusions in *P. pacificus* and another closely related species in the genus, *P. exspectatus*. A comparison of these genomes shows that the E + N fusion in *P. pacificus* must be recent and not an ancestral state in rhabditid nematodes. Indeed, when we include the *P. exspectatus* genome assembly in our analysis we recover seven Nigon elements (A, B, C, D, E, N and X, Figure 3.6). So while limitations in our method can lead to incorrect inference, a denser sampling of present-day genomes does, unsurprisingly, improves performance.

Previous analyses of chromosome evolution in nematodes did not test for the presence of translocation rearrangements (Tandonnet et al. 2019; Gonzalez de la Rosa et al. 2021). To investigate this possibility, we performed inference allowing for translocation rearrangements while again including *P. exspectatus*. We inferred 42 rearrangements using the FFT algorithm (13 fusions, 24 fissions and 5 translocations), which is fewer than the 47 inferred with the FF algorithm (18 fusions and 29 fissions). We tested whether the five translocations inferred, representing 11.9% of the total rearrangements, could be a result of incorrect inference of a rearrangement history that only includes fissions and fusions. We simulated 18 fusions and 29 fissions over the phylogenetic tree (Figure 3.7) and inferred rearrangements using the FFT algorithm (see Methods for details). We find that only 2.7% of rearrangements are (incorrectly) inferred as translocations, with 95% CIs of 0.0 - 10.3%. This provides some support for the idea that these nematode genomes have rearranged through translocation as well as fission and fusion (but see Discussion).

3.4 Discussion

3.4.1 Rearrangement rates and marker error

The need for efficient and accurate methods that infer past genome evolution will only increase as larger genome sequence datasets become available. Here we have implemented and evaluated one such class of method which focuses exclusively on synteny information. Our simulations show that it is straightforward to infer ALGs and rearrangements when the frequency of rearrangements per-branch is low relative to chromosome number. Although this will not always be the case, it is at least true in some groups of organisms. Most lepidoptera, for example, have a large number of chromosomes ($\sim$ 30) and slow rates of rearrangement, making inference straightforward (Wright et al. 2023). Higher rearrangement rates, however, impose a limit on the accuracy of reconstructed ALGs and rearrangements (Figures 3.2 and 3.3). The loss of synteny information with increasing rearrangement rates is hard to overcome within a parsimony-based framework, meaning that a different approach to modelling genome evolution may be required; e.g. estimating rearrangement parameters through simulation (Moshe et al. 2022) or performing a Bayesian sampling of rearrangement histories (Miklós and Tannier 2010). Interestingly, Markov models of chromosome number evolution have recently been developed (Yoshida and Kitano 2021; Setter 2023) and could be adapted to estimate model parameters and sample likely rearrangement histories given a set of genomes. Alternative methods aside, our results show that high rearrangement rates hinder accurate inference and so we encourage researchers to consider the inherent uncertainty and limits to reconstructing rearrangement histories when genomes rearrange frequently.

We also explored the effect of marker errors on performance, finding that our method is robust to missing markers but not orthology error (Figure 3.4). Implementing a minimum set size did alleviate some of this effect, but this must be balanced against the risk of masking real rearrangements. We simulated orthology errors by moving markers between chromosomes, which should emulate incorrect orthology assignment as well as small scaffolding mistakes. It is unclear exactly how frequent these errors are in present-day genome assemblies and annotations. Anecdotally, initial assemblies of large genomes do often require extensive manual curation (e.g. Streicher et al. 2021), suggesting the possibility for small and overlooked scaffolding errors. Similarly, approximate methods for identifying orthology relationships, such a reciprocal best hits, can lead to false single copy orthologues (Emms and Kelly 2019). Limiting analyses to curated genome assemblies with high confidence single copy markers is therefore a sensible precaution when inferring ALGs and rearrangements.

3.4.2 Lessons from a reanalysis of nematode genomes

Our analysis of nematode genomes highlighted some important limitations of our synteny-based inference method, `syngraph`. Firstly, `syngraph` only analyses a single triplet of genomes at a time, thereby ignoring useful synteny information contained in other genomes. The fact that we infer Nigons E + N as syntenic deep in the phylogeny (Figure 3.6) is a result of this limitation, as consideration of more genomes would generate a more parsimonious history involving two independent fusions of E + N on the lineages leading to *P. pacificus* and *S. ratti*. This problem could be alleviated by considering four or five closely related genomes at a time (at the expense of computation time) or by using a weighted graph as in Kim et al. (2017) and Muffato et al. (2023). An even simpler approach would be to perform iterative traversals of the tree until there is no further improvement in parsimony, although this would not guarantee a globally optimal solution (Adam and Sankoff 2008).

The second limitation highlighted by our analysis of nematode genomes is that we ignore useful information in the form of how mixed ALGs are in present-day genomes. Following a fusion of two ALGs, markers belonging to either ALG can become mixed through intra-chromosomal rearrangements. This mixing makes a fusion non-reversible as a subsequent fission is no longer likely to recover the original ALGs (Simakov et al. 2022; Schultz et al. 2023). Additionally, under the assumption that intra-chromosomal rearrangement rates are similar across a given clade, the amount of mixing between fused ALGs provides information about the timing of the fusion (Gonzalez de la Rosa et al. 2021). Without considering ALG mixing, it is difficult to infer whether Nigons N + X fused once in Rhabditina (and later fissioned apart in the lineage leading to *Oscheius tipulae* and *Auanema rhodensis*) or if they instead underwent two independent fusions (Figure 3.7). While our analysis with `syngraph` suggested the former, we agree with the interpretation of Gonzalez de la Rosa et al. (2021) that this fusion likely happened twice given patterns of ALG mixing in *Caenorhabditis sp.* and *H. contortus* and the improbability of a recent fission recovering two separate ALGs. This example suggests that there are limitations to considering synteny information alone and that including at least some gene-order information is likely to improve the accuracy of results.

Finally, we applied a simulation-based test for the presence of translocation rearrangements in the nematode dataset. The observed number of inferred translocations was greater than in 95% of simulations. While it is tempting to view this as strong evidence that the rearrangement history of nematode chromosomes involved translocations, it is important to remember that this result could be generated by other differences between the simulations and the rearrangement process underlying the real data. In particular, a non-uniform rate of rearrangement across the phylogeny (not captured by our simulations) could generate such a result, with (incorrectly) inferred translocations concentrated on branches of the tree with the most fissions and fusions. We therefore interpret this result as weak evidence for translocations and acknowledge the limitations of using simulations under highly simplified models that may only partially capture the complexities of real rearrangement histories.

3.4.3 Gene tree discordance

We have so far assumed that the genealogies underlying rearrangements always follow the species tree. However, this assumption may not hold for groups of closely related populations / species due to appreciable levels of incomplete lineage sorting (ILS) and gene flow. Given that multiple chromosome-level assemblies are now being routinely generated for single species or genera (Kim et al. 2022; Liao et al. 2023; Shi et al. 2023), it is worth considering how gene tree discordance might affect our ability to accurately infer rearrangement histories. A rearrangement that has a history incongruent with the species tree will result in non-sister species carrying the derived chromosome arrangement (Jay et al. 2018). Assuming the species tree, this can be interpreted as two independent rearrangements or a single rearrangement with a subsequent reversion / loss. The challenge is therefore to discern between those scenarios as well the possibility of introgression / ILS. Useful evidence includes the probability of gene tree discordance estimated from polymorphism data (Dutheil et al. 2009), as well as whether rearrangement break points are consistent with multiple origins (Lundberg et al. 2023). A probabilistic method for rearrangement inference that includes gene tree discordance seems like a distant goal at present, but will be useful in any analysis where rearrangements happen on the same time scale as lineage sorting.

3.4.4 Outlook

We have investigated how synteny information can be used to estimate past genome evolution, co-estimating both ALGs and rearrangements. Genome rearrangement problems have garnered considerable interest within the field of mathematics (Fertin et al. 2009), but here we have focused on a relatively simple version of the problem as well as the practicalities of analysing real data. We have been motivated by the fact that accurate inference of past rearrangements has the ability to improve our understanding of how genomes evolve. For example, we still do not know the relative fitness effects of different types of rearrangements (e.g. sex-autosome fusions vs. autosome-autosome fusions) or whether the majority of new fission and fusions are weakly deleterious (Pennell et al. 2015). Additionally, while the importance of fissions and fusions in speciation is becoming clearer (Yoshida et al. 2023; Mackintosh et al. 2023), we still have an incomplete understanding of exactly how such rearrangements prevent gene flow. Identifying inter-chromosomal rearrangements across species and populations will be the first step in answering these biological questions, and it is encouraging that our synteny-based method has already been used to generate new results about genome evolution (Mackintosh et al. 2023; Wright et al. 2023). We anticipate that a variety of methods, focusing on different rearrangements types and relying on different inference procedures, will be required to investigate genome evolution across the tree of life and improve our understanding of the role of chromosome rearrangements in evolution.

3.5 Methods

3.5.1 An overview of syngraph

We implemented methods for investigating past genome evolution from synteny in a modular python tool, `syngraph` (https://github.com/A-J-F-Mackintosh/syngraph). The suggested workflow is to first generate an adjacency graph from orthology data using the `syngraph build` module. A file containing markers and their chromosome assignments must be provided for each genome, with matching marker IDs denoting orthology across genomes. Given the adjacency graph and a phylogenetic tree, ALGs and rearrangements can then be estimated with `syngraph infer`. This generates descriptions of rearrangements across the phylogeny as well as a new graph which includes reconstructed ALGs. This graph can be summarised with `syngraph tabulate`, which produces a table with the assignment of each marker to an ALG.

3.5.2 ALG reconstruction under missingness and error

For a triplet of genomes, G_A, G_B, G_C, the ALGs at G_i (the genome at the internal node of the tree connecting them) are estimated by considering synteny-sets. An adjacency-based approach builds ALGs by combining synteny-sets that are syntenic in > 2 genomes, whereas an event-based finds the set of ALGs that minimises the syntenic distance between G_i and G_A, G_B, G_C. However, ALG reconstruction must be modified when some markers are missing. If a marker is missing from G_C then the synteny-set for such a marker would only have a partial notation, e.g. [A_4, B_{22}] rather than [A_4, B_{22}, C_9]. One solution is to limit the analysis to markers shared across all genomes, with full notations, but this quickly becomes restrictive as the number of leaves in the phylogeny grows. We instead attempt to assign markers with partial missingness to ALGs. Specifically, a marker with notation [A_4, B_{22}] can be assigned if there is a single (already reconstructed) ALG with notation [A_4, B_{22}, C_*], with $*$ representing any sequence in G_C. Markers that are missing from two genomes, or those that are not consistent with a single ALG, are not assigned. This procedure is performed by default within `syngraph infer`, but can be disable by restricting the analysis to markers present in all genomes when reading in data with `syngraph build`.

We also considered the effect of orthology error and how ALG estimation can be made more robust to it. Synteny-sets will vary in size depending on: (i) the total number of markers in the analysis, (ii) the size of chromosomes, (iii) the number of rearrangements between genomes, and (iv) the frequency of orthology error. As an example, consider a chromosome that is conserved in a triplet of genomes, so that hundreds of markers have the same notation, e.g. [A_4, B_{22}, C_9], forming a large synteny-set. If a gene on sequence C_{15} is falsely annotated as having orthology with genes on sequences A_4 and B_{22} we now obtain a new synteny-set with notation [A_4, B_{22}, C_{15}]. Importantly, this synteny set is small, consisting of only a single marker. We therefore implemented an option in `syngraph infer` to enforce a minimum size of synteny-sets (`--minimum`), with the aim of minimising the effect of orthology / scaffolding error.

3.5.3 Simulations

To investigate the performance of these methods we simulated rearrangement histories and attempted to infer them with `syngraph`. The general simulation procedure is as follows: For each simulation a phylogenetic tree is generated under a birth-death process using Dendropy (Sukumaran and Holder 2010). The birth rate is set to one 1.0 and the death rate is 0.5. The tree is sampled once there are $n+1$ leaves. As a result, two leaves will have external branch lengths of zero and so one of them is removed to recover a tree with n leaves. This means that the time of sampling is effectively a random event determined by the time it takes for the number of leaves to increase from n to $n+1$. A genome with k chromosomes is initiated at the root of the tree, with g markers uniformly distributed among them. A total of r rearrangements are placed onto branches of the tree, with probability proportional to branch lengths and different rearrangement types being sampled under a specific fission:fusion:translocation ratio. The initial genome at the root is then simulated forwards in time across the tree and rearranged through set operations. The resulting markers at the leaves of the tree and the phylogeny are parsed to `syngraph infer`. Unless stated otherwise, simulations were parameterised with k = 20 initial chromosomes and g = 1000 markers. Simulations involving only fission and fusions were simulated with rearrangement ratio 1:1:0, whereas those involving fissions, fusions, and translocations were simulated with ratio 1:1:1 (unless otherwise stated). Performance metrics were always estimated using 1,000 simulations for a given parameter combination.

Marker missingness was introduced by randomly removing m% of markers per-genome before parsing the markers to `syngraph infer`. Similarly, marker orthology error was introduced by randomly selecting e% of markers per-genome, and then placing each selected marker on a chromosome with uniform probability. This allowed for the possibility of a selected marker being placed on the chromosome from which it was sampled.

3.5.4 Reanalysis of nematode genomes

We reanalysed 14 rhabditid nematode genomes (Table B.1) and also performed analyses including the genome sequence of *Pristionchus exspectatus*. For each genome assembly we identified sequences corresponding to nematoda_odb10 single-copy genes using BUSCO v5.2.2 (Manni et al. 2021). BUSCO genes annotated on non-chromosome-level sequences or Y chromosomes were removed from the analysis. We then performed the `syngraph` workflow described above using a phylogenetic tree with topology from Gonzalez de la Rosa et al. (2021) and branch lengths corresponding to at least one time unit between speciation events. Missingness was allowed when reading in markers and the minimum synteny-set size for inference was set to 20.

We also performed a simulation test for translocation rearrangements. Each simulation (1,000 in total) had an initial genome with 7 chromosomes, 1,000 markers, and a total of 47 rearrangements with fission:fusion:translocation ratio 29:18:0. The simulations were conditioned on the phylogenetic tree presented in Figure 3.7, and the minimum synteny-set size was again set to 20. The proportion of inferred rearrangements that were translocations were recorded for each simulation. The mean was calculated across simulations and the 95% CIs were estimated using the 25th and 975th percentiles.

3.6 Data availability

The command-line tool, `syngraph`, is available at `https://github.com/A-J-F-Mackintosh/syngraph`. Scripts for simulating rearrangement histories are available at the same Github directory. The NCBI accessions for genome assemblies analysed in this work are given in Table B.1.

Chromosome fissions and fusions act as barriers to gene flow between *Brenthis* fritillary butterflies

4.1 Abstract

Chromosome rearrangements are thought to promote reproductive isolation between incipient species. However, it is unclear how often, and under what conditions, fission and fusion rearrangements act as barriers to gene flow. Here we investigate speciation between two largely sympatric fritillary butterflies, *Brenthis daphne* and *B. ino*. We use a composite likelihood approach to infer the demographic history of these species from whole genome sequence data. We then compare chromosome-level genome assemblies of individuals from each species and identify a total of nine chromosome fissions and fusions. Finally, we fit a demographic model where effective population sizes and effective migration rate vary across the genome, allowing us to quantify the effects of chromosome rearrangements on reproductive isolation. We show that chromosomes involved in rearrangements experienced less effective migration since the onset of species divergence and that genomic regions near rearrangement points have a further reduction in effective migration rate. Our results suggest that the evolution of multiple rearrangements in the *B. daphne* and *B. ino* populations, including alternative fusions of the same chromosomes, have resulted in a reduction in gene flow. While fission and fusion of chromosomes are unlikely to be the only processes that have led to speciation between these butterflies, this study shows that these rearrangements can directly promote reproductive isolation and may be involved in speciation when karyotypes evolve quickly.

4.2 Introduction

4.2.1 Chromosomal speciation

The process of speciation, where groups of individuals become reproductively isolated from one another, is driven by evolutionary forces that prevent gene flow. Many closely related species show differences in karyotype and there has been much discussion about the role of chromosome rearrangements (e.g. inversions, translocations, fissions, and fusions) in preventing gene flow and promoting speciation. Early work on *Drosophila* demonstrated that inversions suppress recombination (Sturtevant 1921; Dobzhansky and Epling 1948). More recently, both theoretical models (Noor et al. 2001; Navarro and Barton 2003; Kirkpatrick and Barton 2006) and examples in a variety of organisms (Wellenreuther and Bernatchez 2018) have shown that inversions can facilitate local adaptation, promote the evolution of genetic incompatibilities and act as barriers between recently diverged species. It is less clear, however, whether fission and fusion rearrangements have a similarly important role in speciation (Rieseberg 2001). These rearrangements do not typically confer the same change in recombination as inversions do, yet there is evidence for increased speciation rates in groups where fissions and fusions happen more often (Bush et al. 1977; Leaché et al. 2016; de Vos et al. 2020). Fissions and fusions could act as barriers to gene flow if hybrid individuals that are heterozygous for a rearrangement suffer from underdominance (heterozygote disadvantage). This will happen when karyotypic heterozygosity generates multivalents at meiosis, which are prone to unbalanced segregation. While there is indeed evidence for fissions and fusions causing underdominance through aneuploidy (Dutrillaux and Rumpler 1977; Castiglia and Capanna 2000; Lukhtanov et al. 2018), models of chromosomal speciation that assume underdominance are paradoxical; for hybrids to suffer from underdominance, the rearrangement must be at high frequency in one population, but how does a rearrangement rise to high frequency if it causes underdominance? Proposed solutions to this paradox include fixation by meiotic drive (White 1968), strong drift in a founder population (Templeton 1981; but see Barton and Charlesworth 1984), and complex rearrangements that evolve in a stepwise manner, where each step has a small fitness effect (White 1978b; Baker and Bickham 1986). This limits the conditions under which underdominant chromosomal speciation can happen, and it is therefore perhaps unsurprising that there are few convincing empirical examples (see Basset et al. 2006, Yannic et al. 2009, and Yoshida et al. 2023).

Not all models of chromosomal speciation require underdominance. For example, fusions could affect gene flow by bringing pre-existing barrier loci onto the same chromosome. Guerrero and Kirkpatrick (2014) showed that for two polymorphic loci maintained by selection-migration balance, a fusion will rise in frequency if it brings two locally adapted alleles into strong linkage disequilibrium (LD). This process has the potential to strengthen the combined effect of barrier loci by reducing recombination between them, thus promoting reproductive isolation. Although Guerrero and Kirkpatrick (2014) do not include underdominance in their model, the process they describe is not mutually exclusive with underdominant chromosomal speciation, and may offer an additional way for fusions to evolve in spite of underdominance.

Fission and fusion rearrangements can also influence the accumulation of reproductive isolation when a barrier to gene flow is highly polygenic. Given such a barrier, the probability that a neutral allele migrates is partly determined by whether it can recombine away from the foreign deleterious alleles that it was introgressed with (Aeschbacher et al. 2017). Fissions and fusions can alter the per-base recombination rates of chromosomes by changing their length and they can therefore influence effective migration. Recently, Martin et al. (2019) showed that recombination rate was the main determinant of the amount of introgression between species of *Heliconius* butterflies, with long fused chromosomes having less introgression than short non-fused ones. These fusions cannot be barriers themselves because they are shared among the species. Instead, because of their length, the fused chromosomes have a low per-base crossover rate (Davey et al. 2017), which reduces effective migration when barrier loci are common. While the fusions in these *Heliconius* butterflies are shared, similar logic applies to a fusion that generates a long chromosome in just one population.

Importantly, a chromosome rearrangement may arise and fix long after a particular species split and so have no role in speciation. Alternatively, if rearrangements are present during the early stages of speciation, they may not have any effect on gene flow. This would be the case if underdominance was weak enough for a rearrangement to be effectively neutral. Moreover, even if rearrangements do have underdominant or recombination modifying effects, there may be barriers of very large effect which have played a much greater role in speciation. It is therefore important to quantify the effect of fission and fusion rearrangements on gene flow, rather than assuming that these conspicuous changes in the genome must play an important role in the speciation process.

4.2.2 Chromosome evolution in butterflies

Most Lepidoptera (moths and butterflies) have similar karyotypes, consisting of around 30 pairs of autosomes and ZW sex chromosomes (de Vos et al. 2020). However, there are notable exceptions. For example, *Pieris* butterflies have a reduced karyotype where chromosomes have undergone substantial reorganisation via inter-chromosomal rearrangements (Hill et al. 2019). There are also taxa with highly variable chromosome counts, such as the butterfly genera *Erebia*, *Lysandra*, *Polyommatus*, and *Leptidea*. In each of these genera it has been suggested that rearrangements have facilitated speciation (Augustijnen et al. 2023; Talavera et al. 2013; Lukhtanov et al. 2005, 2011), although the extent to which rearrangements have affected reproductive isolation remains unclear.

Another group of butterflies in which karyotypes vary is the genus *Brenthis* (Nymphalidae) which consists of four species. While 34 chromosome pairs have been observed in *Brenthis hecate* spermatocytes (de Lesse 1961; Saitoh and Lukhtanov 1988), *B. daphne* and *B. ino* are reported to have only 12 - 14 pairs of chromosomes (Federley 1938; Maeki and Makino 1953; de Lesse 1960; Saitoh 1986, 1987; Saitoh et al. 1989; Saitoh 1991). We recently assembled a *B. ino* reference genome (Mackintosh et al. 2022) with 14 pairs of chromosomes. We found that the genome was highly rearranged compared to the ancestral nymphalid karyotype and that a male individual was heterozygous for a Z-autosome chromosome fusion. These results are consistent with rapid, and likely still ongoing, chromosome evolution in the genus *Brenthis*.

The sister species *B. daphne* and *B. ino* are largely sympatric (Figure 4.1), have differences in larval host plant preference, and are estimated to have split approximately 3 million years ago (Ebdon et al. 2021). Interspecific mating experiments have shown that female *B. daphne* and male *B. ino* can produce fertile offspring, suggesting that reproductive isolation between these species is incomplete (Kitahara 2008, 2012). Additionally, putative F1 hybrids have been observed in Japan (Kitahara 2012). Similar chromosome numbers have been observed for males of either species, 12 - 13 for *B. daphne* and 13 - 14 for *B. ino*, suggesting some intraspecific variation in karyotype, but no large differences between species. However, chromosome numbers will be unchanged by reciprocal translocations or an equal number of chromosome fission

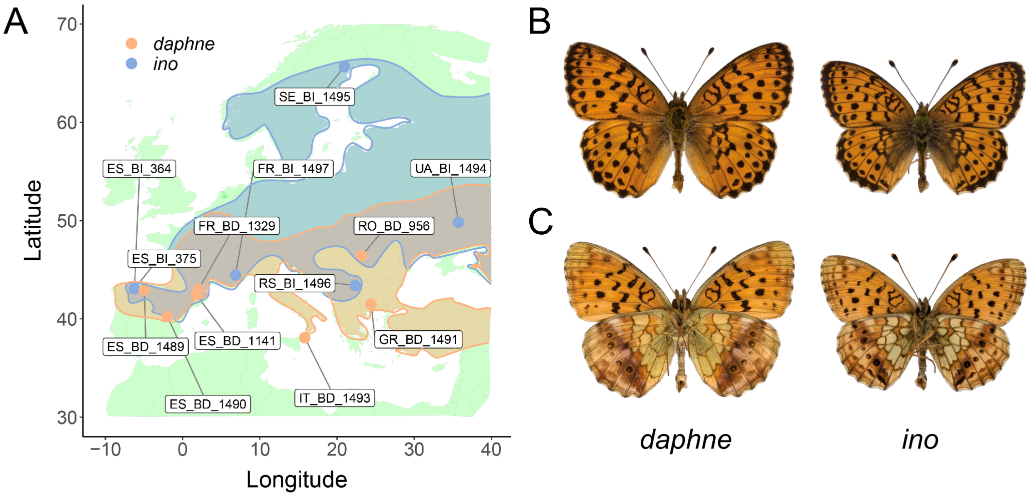

Figure 4.1: **(A)** Sampling locations of *Brenthis daphne* (orange) and *B. ino* (blue) individuals across Europe. Approximate distributions are also shown using the same colour scheme. **(B)** Uppersides of male *B. daphne* and male *B. ino*. **(C)** Undersides of male *B. daphne* and male *B. ino*.

and fusion events. Such “cryptic” rearrangements are best identified by comparing genome assemblies. If *B. daphne* and *B. ino* possess cryptic inter-chromosomal rearrangements, then their recent divergence and potential for ongoing gene flow makes them a useful model for investigating the effects of rearrangements on reproductive isolation.

4.2.3 Overview

Here we show that the genomes of *B. daphne* and *B. ino* differ by multiple fission and fusion rearrangements. More specifically, almost half of the chromosomes are involved in rearrangements, whereas the rest are syntenic. We estimate the demographic history of these species as well as genome-wide variation in effective migration rate (m_e). By intersecting estimates of m_e with chromosome rearrangements, we test whether fissions and fusions have acted as barriers to gene flow. We consider the following scenarios:

- **No effect:** Fission and fusion rearrangements are selectively neutral and have had no effect on the effective rate of gene flow, either directly or indirectly.

- **Underdominance:** Fissions and fusions produce direct, localised barriers to gene flow because early generation hybrids and backcrosses with heterokaryotypes suffer reduced fitness. This would result in decreased post-divergence gene flow on rearranged chromosomes. Assuming that heterokaryotypes still undergo recombination, the reduction in gene flow would be strongest for loci that are closely linked to rearrangement points.

- **Fused barriers:** Fusions are not barriers to gene flow themselves, but have brought individual barrier alleles of large effect into linkage, thus strengthening the barrier effect of these loci. If most fusions put large effect loci into linkage, then this would cause a reduction in gene flow on rearranged chromosomes and the effect would be strongest close to fusion points. This scenario makes no predictions about the effect of chromosome fissions on gene flow.

- **Polygenic barriers:** In the presence of polygenic barriers, fissions and fusions affect gene flow by modifying chromosome lengths and therefore recombination rates. This scenario predicts a negative correlation between gene flow and chromosome length.

4.3 Results

4.3.1 Diversity and divergence

Using our previously published *B. ino* genome assembly (Mackintosh et al. 2022) as a reference, we analysed whole genome sequence data for seven *B. daphne* and six *B. ino* individuals (Figure 4.1; Table C.1). We restricted our analyses to intergenic regions of the genome, as these typically evolve under less selective constraint than genic regions. Consistent with a previous analysis of transcriptomic data (Ebdon et al. 2021), we find that per-site heterozygosity is greater in *B. ino* (0.0111) than in *B. daphne* (0.0043) and that interspecific divergence is considerable (d_{xy} = 0.0228, F_{ST} = 0.4976). We also find evidence of population structure within each species (Figures 4.2A and 4.2B). For example, pairwise F_{ST} is $\sim$ 0.1 for *B. daphne* individuals sampled in different glacial refugia (Iberia, Italy, or the Balkans) and there are similar levels of differentiation between *B. ino* individuals sampled from Iberia and elsewhere in Europe. While this shows that European *B. daphne* and *B. ino* are not panmictic populations, this should only have a small effect on our analyses of long-term diver-

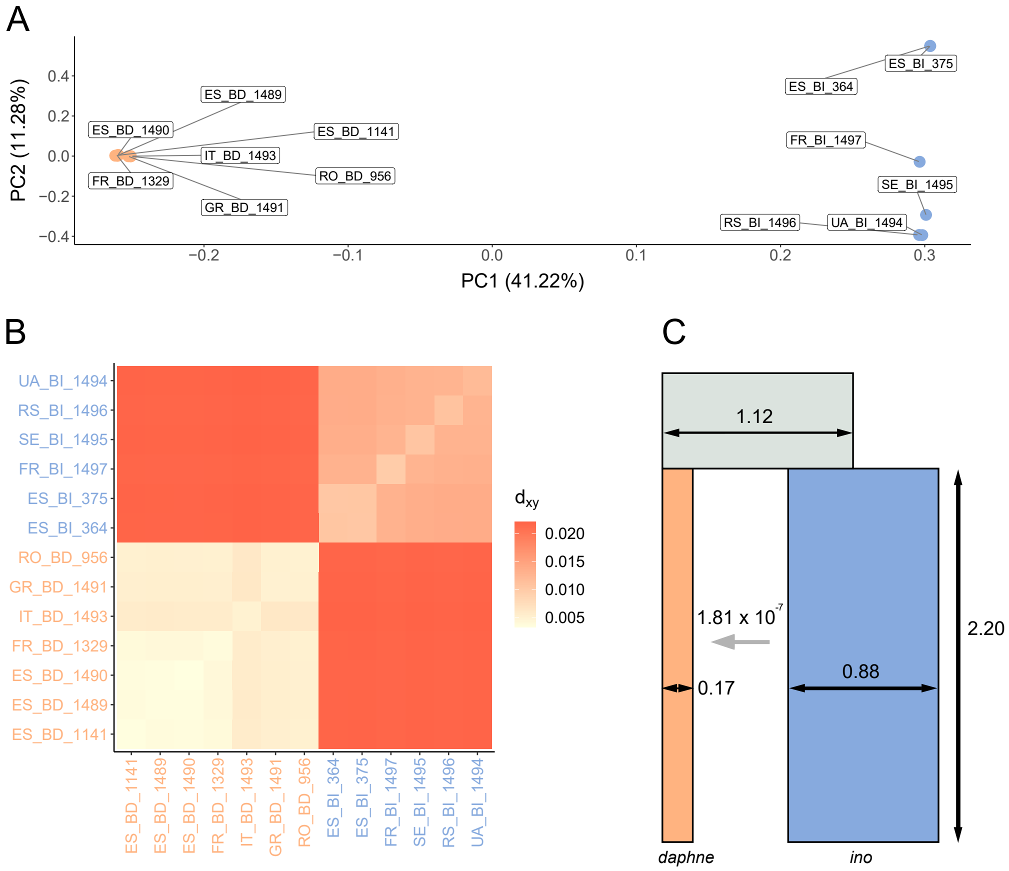

Figure 4.2: Diversity and divergence between *B. daphne* and *B. ino*. **(A)** A PCA of individuals sampled across Europe, with PC1 capturing interspecific variation. Orange points are *B. daphne* individuals and blue points are *B. ino* individuals. The same colour scheme is used in subplots (B) and (C). **(B)** A heatmap showing d_{xy} between pairs of individuals with the diagonal showing heterozygosity within individuals. **(C)** The best fitting demographic model, with parameter values inferred from the genome-wide bSFS. The N_e (indicated by horizontal black arrows) and split time (vertical black arrow) parameter estimates are in units of 10^6 individuals and years respectively. The horizontal grey arrow indicates the direction of gene flow, from *B. ino* to *B. daphne*, forwards in time.

gence and gene flow between the two species (see below).

4.3.2 Demographic history

We use gIMble (Laetsch et al. 2022), a recent implementation of a blockwise likelihood calculation (Lohse et al. 2016), to infer the demographic history of speciation between *B. daphne* and *B. ino*. gIMble calculates the blockwise site frequency spectrum (bSFS) of all possible interspecific pairwise comparisons, i.e. sampling a single diploid genome from each species and tallying mutations in short blocks of sequence (see Methods). We fit three demographic models to the bSFS: strict divergence (*DIV*) and two sce-

Table 4.1: Maximum composite likelihood parameters for three different demographic models. The N_e and split time parameter estimates are in units of 10^6 individuals and years, respectively. The $IM_{\rightarrow Bda}$ model has the highest lnCL.

Model	N_e ***daphne***	N_e ***ino***	N_e ***ancestral***	m_e	**Split time**	**lnCL**
DIV	0.252	0.683	1.433	-	1.183	-2.347×10^8
$IM_{\rightarrow Bda}$	0.171	0.880	1.116	1.811×10^{-7}	2.202	-2.340×10^8
$IM_{\rightarrow Bin}$	0.252	0.683	1.433	0.000	1.183	-2.347×10^8

narios of isolation with migration ($IM_{\rightarrow Bda}$ and $IM_{\rightarrow Bin}$). The *DIV* model has three N_e parameters (*B. daphne*, *B. ino*, ancestral) and a split time parameter. The IM models have an additional parameter, i.e. they assume a constant rate of effective migration (m_e) either from *B. ino* into *B. daphne* forwards in time ($IM_{\rightarrow Bda}$) (Figure 4.2C) or in the opposite direction ($IM_{\rightarrow Bin}$). By optimising the parameters under each model, we found that the $IM_{\rightarrow Bda}$ model fits best (Table 4.1; Figure 4.2C). The *DIV* and $IM_{\rightarrow Bin}$ models converged to the same parameter values and composite likelihood (Table 4.1), i.e. the maximum composite likelihood (MCL) estimate of m_e under the $IM_{\rightarrow Bin}$ model is 0. By contrast, the MCL estimate of m_e from *B. ino* to *B. daphne* under the best fitting ($IM_{\rightarrow Bda}$) model is 1.811×10^{-7}, which is equivalent to 0.124 effective migrants per generation. As a result of this migration, the $IM_{\rightarrow Bda}$ model also has an older split time ($\approx$ 2.2 MY) than the $DIV/IM_{\rightarrow Bin}$ model ($\approx$ 1.2 MY) (Table 4.1).

Given the nesting of models, an *IM* model has to fit the data equally well or better than a *DIV* model because it includes an additional parameter, m_e. To test whether the $IM_{\rightarrow Bda}$ model fits significantly better than *DIV* (see Laetsch et al. 2022), we simulated parametric bootstrap replicates for the MCL estimates under the *DIV* history and optimised both the *DIV* and $IM_{\rightarrow Bda}$ models. The improvement in fit (Δ lnCL) between *DIV* and $IM_{\rightarrow Bda}$ models for parametric bootstrap replicates was far below what we observe in the data (Figure C.1). An IM demographic history, with migration from *B. ino* to *B. daphne*, is therefore well supported.

4.3.3 Synteny

To compare synteny between *B. daphne* and *B. ino*, we generated a chromosome-level assembly for a female *B. daphne* individual, collected in Catalunya, Spain. The assembly is 419.1 Mb in length, with a scaffold N50 of 30.6 Mb and a contig N50 of 13.4 Mb. The *B. daphne* assembly is scaffolded into 13 chromosome-level sequences (hereafter simply referred to as chromosomes) corresponding to 12 autosomes and the Z sex chromosome (Figures 4.3 and C.2). We failed to scaffold the W chromosome which is likely contained within the remaining 35 contigs that total 5.3 Mb.

A pairwise alignment between the *B. daphne* and *B. ino* assemblies shows that only eight chromosomes have one-to-one homology, with the others showing more complex relationships (Figure 4.3). For example, *B. daphne* chromosome 1 is homologous to parts of *B. ino* chromosomes 1, 3, and 8 (Figure 4.3). Altogether, we find that five *B. daphne* chromosomes and six *B. ino* chromosomes are involved in a total of nine inter-chromosomal rearrangements. Hereafter we refer to these chromosomes as rearranged. Additionally, we define rearrangement points as chromosome ends involved in fissions / fusions or sites where alignments on either side connect different *B. daphne* and *B. ino* chromosomes. All nine rearrangements points are supported by both HiC data and contig sequences.

From a single pairwise comparison it is not possible to tell whether a genome possesses a rearrangement in the ancestral or derived state. Therefore, to polarise these rearrangements, we analysed the assemblies alongside a publicly available genome assembly of *Fabriciana adippe* (see Methods). We infer a maximally parsimonious history of rearrangements where the common ancestor of *B. daphne* and *B. ino* had 16 chromosomes, with two fissions and five fusions in the *B. daphne* lineage and two fusions in the *B. ino* lineage. This inferred rearrangement history involves two small ancestral chromosomes (approximately 6.6 and 8.4 Mb), which fused independently to different chromosomes in either species (Figure 4.3).

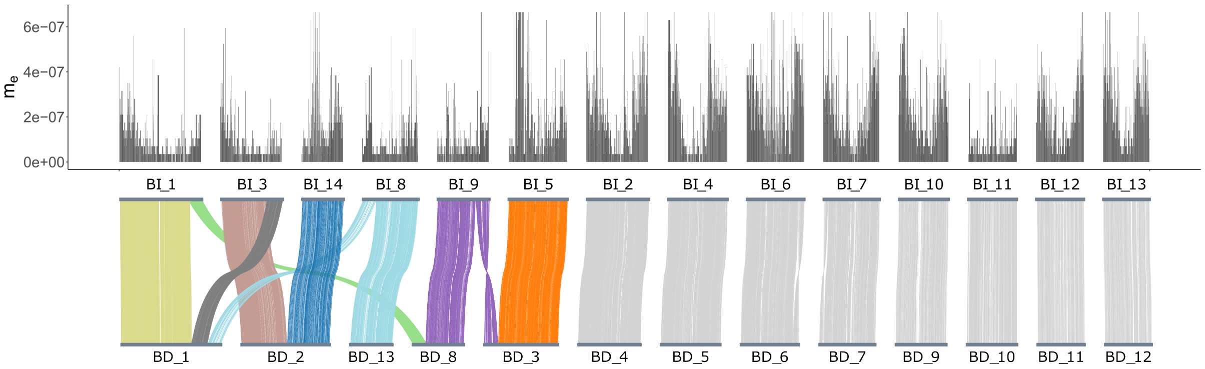

Figure 4.3: A whole genome alignment between *B. daphne* and *B. ino*, with effective migration (m_e) estimates for windows along the *B. ino* genome plotted above. Alignments between non-rearranged chromosomes are coloured in grey. Alignments between rearranged chromosomes are coloured by the inferred chromosomes of the common ancestor of *B. daphne* and *B. ino*. The Z chromosome is labelled as BD_10 in the *B. daphne* genome and BI_11 in the *B. ino* genome.

4.3.4 Variation in m_e across the genome

To investigate the effect of rearrangements on reproductive isolation, we followed the approach of Laetsch et al. (2022) by inferring effective population sizes (N_e) and the effective migration rate (m_e) in windows along the genome. We assume that the species split time is fixed to the MCL estimate under the $IM_{\rightarrow Bda}$ model (Table 4.1). We used simulations to confirm that, given plausible (but conservative) assumptions about recombination, demographic parameters could be inferred for windows containing 30,000 consecutive sequence blocks (Supplementary Note 1; Figure C.3). To infer parameters for the real data, we set up a grid of 67,500 possible parameter value combinations: 15 *B. daphne* N_e values (20,000 - 720,000), 15 *B. ino* N_e values (50,000 - 2,850,000), 15 ancestral N_e values (50,000 - 2,010,000), and 20 m_e values (0 - 6.65×10^{-7}). We identified the best fitting parameter combination for each window (30,000 consecutive blocks, median length = 122 kb). Estimates of local m_e have a long tailed distribution with a peak at 3.5×10^{-8} (Figure 4.4). Consistent with the genome-wide estimate, the mean m_e across windows is 1.845×10^{-7}. We find that m_e is lower on rearranged chromosomes compared to non-rearranged chromosomes (mean $m_e = 1.281 \times 10^{-7}$ vs 2.292×10^{-7} respectively; Figure 4.3; Figure 4.4; one-tailed permutation test $p < 0.005$). This suggests that inter-chromosomal rearrangements are associated with reduced gene flow.

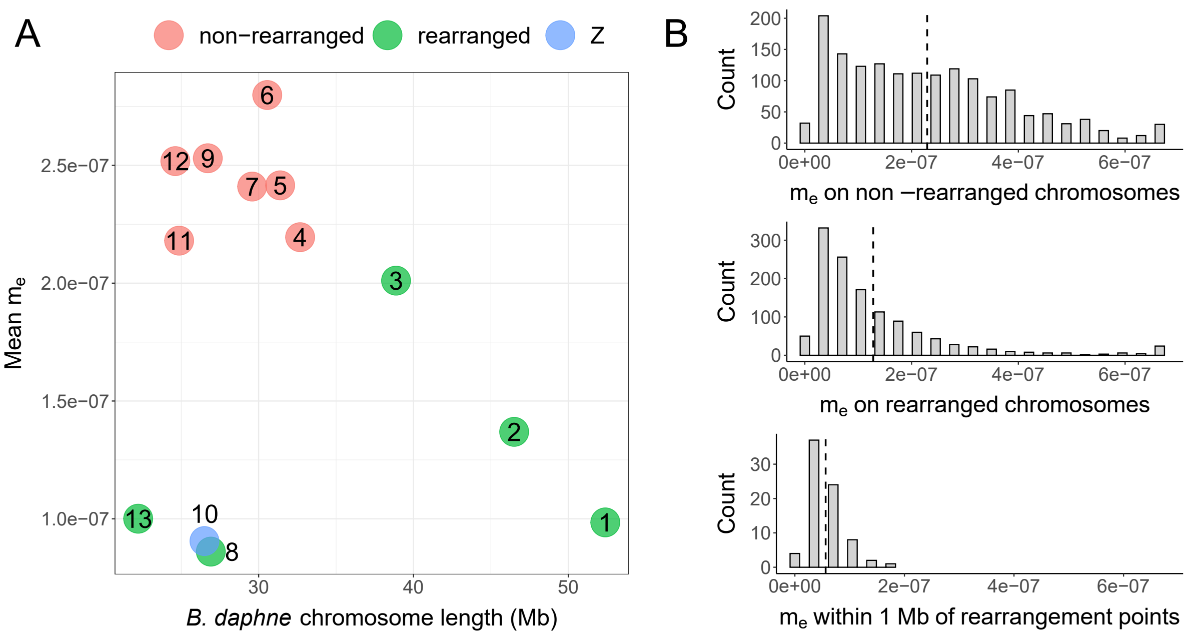

Figure 4.4: Differences in effective migration (m_e) between rearranged and non-rearranged chromosomes. **(A)** Mean m_e for each *B. daphne* chromosome plotted against its length. Points are coloured green if the chromosome is rearranged and red if not. The Z chromosome, which is not rearranged, is coloured blue. **(B)** The distribution of m_e estimates across non-rearranged chromosomes (top), rearranged chromosomes (middle), and within regions near rearrangement points (bottom). For each plot, the mean is plotted as a dashed vertical line.

An alternative approach to estimating m_e for each window is to identify ‘barrier windows’ where there is statistical support for a reduction in gene flow (compared to the background m_e). Following Laetsch et al. (2022), we defined barrier windows as those where $m_e = 0$ has a greater lnCL than $m_e = 1.75 \times 10^{-7}$ (the grid value nearest to the genome-wide m_e estimate). Under this criterion, 23.08% of windows are barriers and these are distributed across all 14 *B. ino* chromosomes. However, the number of barrier windows is not equal among *B. ino* chromosomes, e.g. 48.11% and 4.22% of windows are barriers on chromosome 3 and chromosome 10, respectively. Windows on rearranged chromosomes are twice as often classified as barriers than windows on non-rearranged chromosomes (32.91% vs 15.27%; one-tailed permutation test p $<$ 0.01). The window with the greatest barrier support (Δ lnCL) is located on *B. ino* chromosome 8, with the start of this window being less than 200 kb from a rearrangement point. This alternative, but not independent, estimation of m_e variation provides further evidence for an association between fission and fusion rearrangements and a reduction in gene flow.

Under the best fitting demographic model (Figure 4.2C) *B. daphne* receives gene flow from *B. ino*. As a result, low recombination regions in the *B. daphne* genome are expected to have reduced m_e under the *polygenic barriers* scenario (see Introduction). With this in mind, it is therefore possible that the reduced m_e for rearranged chromosomes is not the result of a direct barrier effect, but instead an indirect consequence of rearrangements producing large *B. daphne* chromosomes with low recombination rates (e.g. *B. daphne* chromosomes 1, 2, and 3; see Figure 4.3). To test this possibility, we assigned each genomic window to a *B. daphne* chromosome using a whole genome alignment (Figure 4.3) and calculated the mean m_e of each *B. daphne* chromosome. There is no significant linear relationship between *B. daphne* chromosome length and mean m_e (Spearman's $\rho_{df=11}$ = -0.0769, p = 0.8065; Figure 4.4). While the largest chromosomes, which happen to be rearranged, do indeed have relatively low m_e, short rearranged chromosomes also have low m_e. Additionally, the Z chromosome (*B. daphne* chromosome 10, *B. ino* chromosome 11), which is not rearranged and is short, has low mean m_e. Chromosome size alone is therefore unlikely to explain the association between chromosome rearrangements and reduced m_e.

If fission and fusion rearrangements act as direct barriers to gene flow, such as in the *fused barriers* and *underdominance* scenarios, then we would expect loci that are closely linked to rearrangement points to have the greatest reduction in m_e. This is because loci that are on the same chromosome but are less closely linked will be more likely to recombine away following introgression. Selection against foreign rearrangements will therefore only have a weak effect on loosely linked loci. We indeed find that genomic windows which are located within 1 Mb of a rearrangement point have a lower m_e (mean = 5.618×10^{-8}) than those located elsewhere on rearranged chromosomes (mean = 1.328×10^{-7}) (Figure 4.4; one-tailed permutation test $p < 0.0005$). All 76 of these windows have estimated m_e values (between 0 and 1.75×10^{-7}; Figure 4.4) that are below the genome-wide estimate (1.811×10^{-7}). Additionally, 59.21% of them are classified as barrier windows. The signal of reduced m_e at closely linked sites provides support for rearrangements having acted as barriers to gene flow.

4.4 Discussion

4.4.1 The effect of fission and fusion rearrangements on gene flow

We have shown that the fritillary butterflies *Brenthis daphne* and *B. ino* possess different karyotypes due to multiple fission and fusion rearrangements, and that these rearrangements are associated with reduced m_e. We can therefore reject the *no effect* scenario where rearrangements are only coincidental with speciation.

We considered the possibility that the association between rearrangements and low m_e could be solely driven by the modification of chromosome lengths, and therefore recombination rate, in the presence of polygenic barriers. Indeed fusions in the *B. daphne* population have generated large (up to 52 Mb) chromosomes with presumably low recombination rates and low m_e. However, the fact that small chromosomes that are involved in fissions and fusions have reduced m_e (Figure 4.4) is not well explained by the *polygenic barriers* scenario where rearrangements only modify the size of chromosomes. We do expected recombination rate to play some role in determining variation in m_e across the genome (see below). However, given the small number of chromosomes in the focal *Brenthis* pair, the relationship between chromosome length and m_e variation remains difficult to quantify precisely. Nevertheless, our results – especially the finding of reduced m_e around rearrangement points – are better explained by localised natural selection against introgression around rearrangements. In other words, rearrangements have acted as barriers to gene flow.

The association between rearrangements and m_e that we find is consistent with two scenarios, *underdominance* and *fused barriers*. Under the *underdominance* scenario we would expect rearranged chromosomes to have lower m_e and we would also expect m_e to be further reduced near rearrangement points. We find both of these patterns in our data (Figure 4.4). The expectations under the *fused barriers* scenario are more variable. If the number of initial barrier loci is small, and fusions that put two or more barrier loci into strong LD rise in frequency due to natural selection (Guerrero and Kirkpatrick 2014), then we would indeed expect lower m_e on rearranged chromosomes as well as particularly low m_e around fusion points. However, if there were enough initial barrier loci so that some were in strong LD by chance alone, then the

m_e of barrier loci brought together by a fusion would be unremarkable. We find that all rearranged chromosomes have reduced m_e when compared to other autosomes (Figure 4.4), which can only be explained by the *fused barriers* scenario if fusions always put barrier loci into strong LD, with their combined effects being greater than barrier loci on non-rearranged chromosomes. One way to discern between the *fused barriers* and *underdominance* scenarios would be to compare m_e around fission points, as it is only expected to be reduced in the *underdominance* scenario. However, the two fission events in the *B. daphne* lineage are both followed by fusions, making this test inappropriate. So while the *fused barriers* scenario requires a particular number and distribution of initial barrier loci, it is still consistent with our results. Note, again, that the *fused barriers* and *underdominance* scenarios are not mutually exclusive, and both processes could have contributed to fissions and fusions acting as barriers to gene flow between *B. daphne* and *B. ino*.

4.4.2 The underdominance paradox

Earlier we noted that chromosomal speciation models involving underdominance are often paradoxical (see Introduction). So, how could rearrangements rise to high frequency in the *B. daphne* and *B. ino* populations if heterokaryotypes are selected against? The *fused barriers* scenario is one way in which underdominance could be overcome within a population because this scenario involves natural selection favouring the fusions to enhance local adaptation. Although it can only explain the evolution of fission rearrangements if they were translocations instead. Another solution is that the fitness consequences of heterozygosity for a single fission / fusion are effectively neutral. This is more likely to be the case when chromosomes are holocentric (Lucek et al. 2022), as they are in butterflies (although see Dutrillaux et al. 2022). A single rearrangement could therefore fix in a population and, over time, karyotypes could evolve in a stepwise process. By contrast, heterozygosity for multiple fissions / fusions could have a larger fitness cost due to the difficulty of properly segregating multiple, potentially complex, multivalents (Dutrillaux and Rumpler 1977; Castiglia and Capanna 2000). If *B. daphne* and *B. ino* evolved multiple rearrangements through a stepwise process during a period of allopatry, then rearrangements could act as barriers once the populations came back into contact. This scenario, which has similarities with the stepwise accumulation of Bateson–Dobzhansky–Muller incompatibilities (Dobzhansky 1934), has been previously described by White (1978a), and is known as the chain model (Rieseberg 2001). While the rearrangements between *B. daphne* and *B. ino* are numerous and complex (Figure 4.3), consistent with the chain model, we have not tested whether these populations underwent a period of allopatry followed by secondary contact. There may be enough information in the two-diploid bSFS to fit such a model, but no exact likelihood implementation exists yet (although see Bisschop 2022 and Beeravolu et al. 2018) and so we have had to assume a simpler IM model in order to investigate variation in m_e across the genome. Importantly, if the chain model does apply here, it has only generated partial barriers to gene flow and has not resulted in complete reproductive isolation. If hybrids with heterokaryotypes were sterile, then gene flow would cease across the entire genome. We instead find that gene flow is reduced on rearranged chromosomes, which means that heterokaryotype hybrids must have been able to backcross.

4.4.3 Variation among rearrangements

In our analysis we grouped chromosomes into two categories, rearranged and non-rearranged. While this simplification is convenient, it ignores potentially important variation among rearrangements. For example, rearrangements could vary in their effect on meiosis. While most rearrangements will result in multivalents, particularly complex multivalent chains could cause recombination suppression if crossover formation is physically constrained (Borodin et al. 2019). Rearrangements are also likely to vary in terms of their time of origin, with some arising around the split time of *Brenthis daphne* and *B. ino* ($\approx$ 2.2 MY), affecting gene flow during the early stages of speciation. Others may have arisen much more recently, and so have made a relatively small addition to existing reproductive isolation. It is also possible that some of the rearrangements we have identified are still polymorphic within species (i.e. not fixed between species). Interestingly, a polymorphic rearrangement could act as a barrier to gene flow within a species. An analysis of intraspecific gene flow (Supplementary Note 2; Table C.2) suggests that the rearrangements we have identified only reduce gene flow between species, rather than between different refugial populations of the same species (Figure C.4). Nonetheless, it is likely that at least some rearrangements are polymorphic given variation in chromosome number within both *B. daphne* (de Lesse 1960; Saitoh 1986) and *B. ino* (Federley 1938; Saitoh 1991; Maeki and Makino 1953). We cannot yet infer an evolutionary history for each rearrangement that is detailed enough to capture its time of origin and frequency over time. However, such detailed reconstructions may become a realistic goal as the quality of data and inference methods improve.

4.4.4 Other determinants of m_e variation

We have focused on whether chromosome rearrangements, the most conspicuous genomic difference between these species, have acted as barriers to gene flow. Yet variation in m_e across the genome cannot be explained by rearrangements alone. Firstly, the centres of non-rearranged chromosomes clearly have lower m_e estimates than regions near chromosome ends (Figure 4.3). This can be explained by variation in recombination rate, with crossovers concentrated towards telomeres (Haenel et al. 2018), as neutral alleles are more likely to introgress if they can quickly recombine away from the barrier loci they are linked to. The fact that chromosome centres consistently have lower m_e suggests that there are other barriers to gene flow distributed across the genome, not only rearrangement points. Secondly, the Z chromosome has a considerably lower mean m_e than all other non-rearranged chromosomes (Figure 4.4), which cannot be because of rearrangements or low recombination (the Z recombines more frequently than autosomes in Lepidoptera due to achiasmatic meiosis in females with ZW sex chromosomes; Maeda 1939; Turner and Sheppard 1975). Instead, low m_e on the Z may be a result of recessive barrier loci being exposed to selection in females (Turelli and Orr 1995). Additionally, if the Z evolves faster than autosomal chromosomes (Mongue et al. 2021), then barrier loci, both recessive and dominant, may accumulate faster. Reduced gene flow on the *Brenthis* Z chromosome mirrors findings in other butterfly systems (Xiong et al. 2022; Rosser et al. 2022), as well as in birds (Irwin 2018; Ottenburghs 2022), suggesting that Z chromosomes often accumulate reproductive isolation at a faster rate than autosomes. Given that there are likely many barriers to gene flow between *B. daphne* and *B. ino*, especially on the Z, it may be inaccurate to describe the history of these species as ‘chromosomal speciation’. Instead, fission and fusion rearrangements are likely one of several processes that have promoted reproductive isolation.

4.4.5 Outlook

The particular process we have investigated here, where fissions and fusions act as barriers to gene flow, likely modulates speciation more strongly in certain groups of organisms than in others. For example, the majority of butterfly species have very slow karyotypic evolution and thus speciation will have happened through the accumulation of other genetic barriers. Nevertheless, radiations of butterflies where karyotypes evolve quickly (e.g. the genera *Erebia*, *Lysandra*, and *Polyommatus*) may be partly explained by fissions and fusions acting as barriers to gene flow. This could also be true for other radiations in which karyotypes vary, such as Rock-wallabies (Potter et al. 2017), Morabine grasshoppers (White et al. 1964; Kawakami et al. 2011), and *Carex* sedges (Márquez-Corro et al. 2021). Evidence for fissions and fusions promoting speciation has often been macro-evolutionary, where analyses of large phylogenetic trees have shown an association between rearrangement and diversification rates. By contrast, focusing on a single pair of species, we have shown that fissions and fusions can act as barriers to gene flow and that their effect can be quantified from genomic data.

4.5 Materials and methods

4.5.1 Sampling

Butterflies were collected by hand netting. Individuals collected by KL were flash frozen in a liquid nitrogen dry shipper (Table C.1); those collected by RV and collaborators were dried and, after some days, stored in ethanol at -20°C (Table C.1).

4.5.2 Sequencing

Previously published data - the *B. ino* genome assembly and whole genome sequencing (WGS) data from three individuals (NCBI accessions: GCA_921882275.1; ERX7241006; ERX7249694; ERX7250096) - were used in this study (Table C.1). The sequencing process for generating these data is described in Mackintosh et al. (2022). Additional sequence data - Pacbio long reads, HiC data, and WGS data for ten individuals - were generated for this study (Table C.1).

A high molecular weight (HMW) DNA extraction was performed for *B. daphne* individual ES_BD_1141 (Table C.1), using a salting-out protocol (see Mackintosh et al. 2022 for details). A SMRTbell sequencing library was generated from the HMW extraction by the Exeter Sequencing Service. This was sequenced on three SMRT cells on a Sequel I instrument to generate 20.4 Gb of Pacbio continuous long read (CLR) data.

A second *B. daphne* individual (FR_BD_1329; Table C.1) was used for chromatin conformation capture (HiC) sequencing. The HiC reaction was done using an Arima-HiC kit, following the manufacturer's instructions for flash frozen animal tissue. The Illumina TruSeq library was sequenced on an Illumina NovaSeq 6000 at Edinburgh Genomics, generating 9.9 Gb of paired-end reads.

DNA extractions were performed for nine individuals using a Qiagen DNeasy Blood & Tissue kit, following the manufacturers instructions. TruSeq Nano gel free libraries were prepared from these extractions as well as the HMW extraction of individual ES_BD_1141. All ten libraries were sequenced on a NovaSeq 6000 at Edinburgh Genomics, generating between 10.1 and 40.0 Gb of paired-end reads for each sample.

4.5.3 Genome assembly

A *B. daphne* genome sequence was assembled from the Pacbio long reads (ES_BD_1141), HiC data (FR_BD_1329), and WGS data (ES_BD_1141) using the same pipeline described in Mackintosh et al. (2022) (Hu et al. 2023; Aury and Istace 2021; Laetsch and Blaxter 2017; Guan et al. 2020; Durand et al. 2016; Robinson et al. 2018), with one modification; YaHS (Zhou et al. 2023) was used to scaffold the contig assembly into chromosomes rather than 3d-dna (Dudchenko et al. 2017).

4.5.4 Synteny analysis

To identify rearrangements, the *B. daphne* and *B. ino* assemblies were aligned with minimap2 v2.17 (Li 2018) using the option `-x asm10`. Alignments longer than 50 kb and with a mapping quality of 60 (2563 in total with a mean length of 132 kb) were visualised with minimap2synteny.py. This script (see Data availability) plots the chromosomes of each genome with ribbons connecting regions that align to each other

(Figure 4.3). Fission and fusion rearrangements were identified from the plot and breakpoints were defined using the paf file generated by minimap2.

To polarise rearrangements and infer ancestral chromosomes, the *Brenthis* assemblies were analysed alongside a *Fabriciana adippe* genome assembly (NCBI accession: GCA_ 905404265.1; Lohse et al. 2022b). Single copy orthologues were identified in each genome using BUSCO v5.3.2 (Simão et al. 2015) with the lepidoptera_odb10 dataset. Complete and Fragmented BUSCO genes were analysed with syngraph (`https://github.com/DRL/syngraph`). In brief, syngraph identifies sets of markers, in this case BUSCO genes, that are found on the same chromosome in all three assemblies. Which sets of markers are found together on extant chromosomes is also recorded. Then, given a phylogenetic tree, parsimony is used to estimate the marker content of ancestral chromosomes and the inter-chromosomal rearrangements on each branch.

4.5.5 Variant calling and filtering

Raw WGS reads were adapter and quality trimmed with fastp v0.2.1 (Chen et al. 2018) and aligned to the *B. ino* assembly (GCA_921882275.1) with bwa-mem v0.7.17 (Li 2013). Duplicates were marked using sambamba v0.6.6 (Tarasov et al. 2015). Variants were called with freebayes v1.3.2-dirty (Garrison and Marth 2012), using the following options: `--limit-coverage 250 --use-best-n-alleles 8 --no-population-priors --ploidy 2 --use-mapping-quality --haplotype-length -1`. This generated a VCF file containing unfiltered SNP and indel calls. Note that the `--limit-coverage 250` and `--use-best-n-alleles 8` options are for computational efficiency only and should not affect whether variants are called at a given site.

Variant calls were filtered using gIMble preprocess (Laetsch et al. 2022), with the following options: `--snpgap 2 --min_qual 10 --min_depth 8 --max_depth 3`, where `--max_depth` is in units of mean coverage. This generated a VCF of filtered SNPs, where SNPs were not within two bases of an indel and QUAL scores of SNPs were >= 10. Individual genotypes were set to missing if read depth was below the minimum depth or above the maximum depth. Sites with multiallelic SNPs were retained if they satisfied all other filtering criteria.

Callable sites for each individual were identified with mosdepth v0.3.2 (Pedersen and Quinlan 2017), called through gIMble preprocess. To restrict downstream analyses to intergenic regions of the genome, the callable sites bed file was stripped of sites belonging to genic and/or repeat regions.

4.5.6 Summaries of diversity and divergence

Variants in intergenic regions of autosomal chromosomes, where all individuals had a genotype, were used to generate a PCA with plink v1.90b6.18 (Purcell et al. 2007).

Genome-wide averages of d_{xy} and F_{ST} were calculated from the same set of variants using VCF_stats.py. The denominator for d_{xy} was the total number of autosomal intergenic sites that were callable across all individuals (123 Mb out of a possible 150 Mb).

4.5.7 Demographic modelling with gIMble

To fit a genome-wide demographic model, autosomal variants were analysed with gIMble. Blocks of 64 bases, with a max span of 128 bases, were generated for all interspecific pairwise comparisons. A bSFS with a k_{max} values of 2 was tallied from these blocks. The bSFS contains mutation counts for 81,104,834 interspecific blocks, each of length 64 bases, distributed over 139 Mb of intergenic sequence. Three models (DIV, $IM_{\rightarrow Bda}$, $IM_{\rightarrow Bin}$) were fit to the genome-wide bSFS and the model with the highest lnCL ($IM_{\rightarrow Bda}$) was used for downstream analysis. Absolute parameter estimates were calculated by assuming the *de novo* mutation rate estimate for *Heliconius melpomene* (2.9×10^{-9} mutations per site per generation; Keightley et al. 2015) and a generation time of one year.

Parametric bootstrap simulations were performed with msprime v1.0.2 (Baumdicker et al. 2021), called through gIMble simulate. The simulations were parameterised with the maximum composite likelihood (MCL) DIV values, i.e. the best fitting history without gene flow, and a per-base recombination rate of 8.5×10^{-9} (equivalent to a single crossover per male meiosis for 14 chromosome pairs). A total of 100 replicates were performed. Each simulated bSFS was optimised under the DIV and $IM_{\rightarrow Bda}$ models.

To estimate variation in m_e and N_e across the genome, genomic windows containing 30,000 consecutive blocks were defined. Next, likelihood calculations were generated for a grid of 67,500 parameter combinations using gIMble makegrid. The lnCL of each windowed bSFS was then calculated for every grid-point. The MCL grid-point was recorded for each window. Additionally, MCLs were recorded for each window conditioning on each m_e value, e.g. the MCL of a window considering all grid-points where $m_e = 0$.

Variation in m_e across the Z chromosome was estimated as above, with the following modification: only male individuals (two *B. daphne*, three *B. ino*, Table C.1) were analysed (since females only have a single copy of the Z). Given the smaller number of interspecific comparisons (6 vs 42 for the autosomal analysis), we reduced the number of blocks per window accordingly (4286 consecutive blocks rather than 30,000) to achieve windows of a comparable span.

Demographic models and variation in m_e were also estimated at the intraspecific level (Supplementary Note 2). Individuals within each species were grouped as Iberian if collected in Spain, and Balkan if collected in Serbia, Greece, Romania, or Ukraine. Note that we use the terms Iberian and Balkan to refer to the likely glacial refugia in which populations expanded from. Genome-wide demographic models were fit to the Iberian-Balkan bSFS for each species. For the *B. daphne* analysis, where the genome-wide model suggested post-divergence gene flow, windows of 4286 blocks were defined and a grid of 10,000 parameter values was calculated. Windows were then run across the grid (as described above) to obtain m_e estimates for each window.

4.5.8 Statistical analysis

Permutations were used to test whether differences in m_e between chromosomes were statistically significant. First, a label-switching operation was performed to randomise whether a *B. ino* chromosome was defined as rearranged or non-rearranged, with the rearranged group always consisting of six chromosomes. For each permutation, the differences in mean m_e and barrier window frequency between the randomly defined groups were calculated and used to build null distributions. The observed differences in mean m_e and frequency of barrier windows between rearranged and non-rearranged chromosomes were then compared to these distributions to calculate p-values.

A second permutation test was used to approximate a null distribution for the difference in mean m_e between windows within 1 Mb of rearrangement points, and windows that are elsewhere on rearranged chromosomes. For each permutation, nine points were randomly chosen from rearranged chromosomes and adjacent windows around these points were sampled. The number of adjacent windows sampled for each point was matched to a number of adjacent windows within 1 Mb of a rearrangement point in the real data. Permutations where any window was sampled multiple times were discarded. To avoid under-sampling windows near the ends of chromosomes, adjacent windows were allowed to roll over on to the next rearranged chromosome. The difference in mean m_e between windows adjacent to randomly sampled points and all other windows on rearranged chromosomes, was calculated for each permutation. A total of 100,000 permutations were done to approximate the null distribution. The difference in mean m_e between windows within 1 Mb of rearrangement points and windows that are elsewhere on rearranged chromosomes, was compared to the null distribution to calculate a p-value.

Spearman's ρ was calculated for chromosome length and mean m_e. All analysis were performed in R (R Core Team 2021).

4.6 Data availability

Raw sequencing reads and the *Brenthis daphne* genome assembly are available at the European Nucleotide Archive under project accession PRJEB56310. The scripts VCF_stats.py and minimap2synteny.py, as well as the R code for carrying out at permutation tests are available at https://github.com/A-J-F-Mackintosh/Mackintosh_et_al_2022_Binodaphne.

Do chromosome rearrangements fix by genetic drift or natural selection? Insights from *Brenthis* butterflies

5.1 Abstract

Large-scale chromosome rearrangements, such as fissions and fusions, are a common feature of eukaryote evolution. They can have considerable influence on the evolution of populations, yet it remains unclear exactly how rearrangements become established and eventually fix. Rearrangements could fix by genetic drift if they are weakly deleterious or neutral, or they may instead be favoured by positive natural selection. Here we compare genome assemblies of three closely related *Brenthis* butterfly species and characterise a complex history of fission and fusion rearrangements. An inferred demographic history of these species suggests that rearrangements became fixed in populations with large long-term effective size (N_e), consistent with rearrangements being selectively neutral or only very weakly underdominant. Using a recently developed analytic framework for characterising hard selective sweeps, we find that chromosome fusions are not enriched for evidence of past sweeps compared to other regions of the genome. Nonetheless, we do infer a strong and recent selective sweep around one chromosome fusion in the *B. daphne* genome. Our results suggest that rearrangements in these species likely have weak absolute fitness effects and fix by genetic drift. However, one putative selective sweep raises the possibility that natural selection may sometimes play a role in the fixation of chromosome fusions.

5.2 Introduction

5.2.1 How do chromosome rearrangements fix?

Eukaryotic genomes vary widely in chromosome number and structure, i.e. karyotype. While closely related species often have similar karyotypes, there are also examples of considerable variation in chromosome number within genera (Hipp et al. 2009; Lukhtanov 2015) and even species (John and Hewitt 1970; Searle 1991; Zima et al. 1996). This variation is typically generated through chromosome rearrangements, with chromosome fissions and fusions resulting in increases and decreases in chromosome number, respectively. These rearrangements have been shown to promote speciation as well as influence the rate and distribution of recombination events (Bidau et al. 2001; Davey et al. 2017; Yoshida et al. 2023; Näsvall et al. 2023; Mackintosh et al. 2023), but our understanding of their role in evolution is limited by the fact that we do not know how they rise to high frequency in the first place. Heterozygosity for fissions or fusions can cause improper segregation during meiosis (White 1973), and so it is often suggested that new rearrangements are weakly deleterious and establish through strong genetic drift (Wilson et al. 1975; Bush et al. 1977). An alternative view is that rearrangements become fixed because they are favoured by natural selection (Bickham and Baker 1979; Qumsiyeh and Handal 2022), but there is currently limited empirical evidence to support this.

There are a number of different ways for a fission or fusion to be advantageous. For example, a chromosome fusion can increase linkage disequilibrium (LD) between coadapted alleles, leading to enhanced local adaptation and fixation of the rearrangement (Fisher 1930; Charlesworth 1983; Guerrero and Kirkpatrick 2014). There are examples of fused chromosomes that are enriched for adaptive loci (Wellband et al. 2019; Liu et al. 2022), but it is unclear what fraction of these variants predate the rearrangements and potentially contributed to their fixation. Rearrangements can also have direct effects on gene expression, either through changes in genome positioning within the nucleus (Di Stefano et al. 2020) or if breakpoints occur within a gene body or regulatory element (Harewood and Fraser 2014). While most changes in gene expression are likely deleterious, any beneficial changes could lead to the spread of a rearrangement. Meiotic drive is another mechanism by which chromosome rearrangements could rapidly increase in frequency. This process involves drive alleles that are transmitted to gametes more than 50% of the time and typically occurs within asymmetric meiosis (Pardo-Manuel de Villena and Sapienza 2001b). Chromosome rearrangements with differences in centromere size or form can act as drive alleles which leads to their fixation (Pardo-Manuel de Villena and Sapienza 2001a; Stewart et al. 2019). Although this process is primarily associated with monocentric chromosomes (i.e. those with a single centromere) it has also been suggested to occur in organisms with holocentric chromosomes, such as nematodes and Lepidoptera, where centromeres are not localised (Bureš and Zedek 2014).

While the processes described above are certainly possible, the fixation of fissions and fusions may not be adaptive at all. Instead, a rearrangement could fix entirely through genetic drift (Wright 1941). This may be the case if meiosis is robust to the risk of unbalanced segregation associated with heterokaryotypes (Borodin et al. 2019). Even if a rearrangement does confer a fitness cost, strong drift and inbreeding in small populations could still lead to its fixation (Wright 1941; Lande 1979). Under this scenario, one would expect more rearrangements involving Y/W chromosomes than those involving X/Z chromosomes, due to the approximately three-fold difference in effective population size (N_e). Pennell et al. (2015) tested this prediction and found that Y/W-autosome fusions are indeed significantly more common than X/Z-autosome fusions in fish and squamate reptiles, though not in mammals. They therefore suggest that sex-autosome fusions are often weakly deleterious and fix through genetic drift. While the same could be true for fissions and autosome-autosome fusions, it is unclear whether all of these rearrangements have similar fitness effects.

5.2.2 Inferring selective sweeps

If fissions and fusions rise in frequency due to natural selection, sites that are tightly linked to recent rearrangements will show evidence of selective sweeps. This process, in which a beneficial allele increases rapidly in frequency and nearby alleles ’hitchhike’ with it, leaves a signature in population genomic data that can be used to infer past selection (Maynard Smith and Haigh 1974). A variety of methods have been developed for sweep inference, often making use of different types of genomic data, such as allele frequencies (Nielsen et al. 2005), patterns of haplotype similarity (Garud et al. 2015; Harris and DeGiorgio 2020), or even reconstructed ancestral recombination graphs (Stern et al. 2019; Hejase et al. 2022). One limitation shared by a number of methods is the assumption that the modelled selective sweep has completed very recently. This limits the power to detect and accurately characterise even strong sweeps that happened deeper in time.

Recently, Bisschop et al. (2021) showed that, for small sample sizes, the joint distribution of genealogical branch lengths can be derived under an approximate model of a selective sweep. This allows the calculation of composite likelihoods from mutation configurations in short sequence blocks, in particular the blockwise site frequency spectrum (bSFS; Bunnefeld et al. 2015). Importantly, this analytic framework can be used to infer and characterise sweeps that happened further back in time (i.e. $> 0.1N_e$ but $< 4N_e$ generations ago) by treating the sweep as a discrete event that is both preceded and followed by a neutral coalescent process (Bisschop et al. 2021). This inference method can therefore be used to test whether natural selection has acted on certain regions of the genome, even if the selective events are relatively old.

5.2.3 Overview

Here we use the fast rate of chromosome evolution in *Brenthis* fritillary butterflies to investigate how chromosome fissions and fusions evolve. Previous work has shown that chromosome numbers vary substantially among *Brenthis* species (Saitoh 1986; Saitoh and Lukhtanov 1988; Saitoh 1991; Pazhenkova and Lukhtanov 2019; Mackintosh et al. 2022) and that this variation is due to chromosome rearrangements (Mackintosh et al. 2023) rather than differences in ploidy or supernumerary chromosomes. The genus consists of four species, *B. daphne* (Denis and Schiffermüller, 1775), *B. ino* (Rottemburg, 1775), *B. hecate* (Denis and Schiffermüller, 1775) and *B. mofidii* (Wyatt, 1969), and here we analyse genomic data from the first three. First, we describe a newly generated chromosome-level genome assembly for *B. hecate*. This species has a much larger number of chromosomes (n_c = 34) than *B. daphne* (n_c = 12-13) or *B. ino* (n_c = 13-14), implying a history of rapid rearrangement. Secondly, we compare the genomes of these three *Brenthis* species with publicly available genome assemblies of two other fritillary butterfly species in the tribe Argynnini. Using a maximum parsimony method, we show that almost all rearrangements are confined to the genus *Brenthis*. Thirdly, we use whole genome resequence data for all three *Brenthis* species to estimate their demographic history. This allows inferred rearrangements to be placed within the context of species divergence times and effective population sizes. Finally, we investigate whether chromosome fusions, the most common rearrangement type in our dataset, have fixed through hard selective sweeps. For each of 12 potentially recent chromosome fusions, we use the analytical framework of Bisschop et al. (2021) to estimate support for a hard sweep model as well as the time since the sweep and the strength of selection.

5.3 Materials and methods

5.3.1 Sampling and sequencing

Butterflies were collected by hand netting and frozen from live in a -80 freezer. We performed a high molecular weight (HMW) DNA extraction for one *Brenthis hecate* individual (ES_BH_1412; Table D.1) using a salting-out protocol (see Mackintosh et al. 2022 for details). For four other *B. hecate* individuals (Table D.1), DNA was extracted from Ethanol preserved samples using a Qiagen DNeasy Blood & Tissue kit, following the manufacturer's instructions. Edinburgh Genomics (EG) prepared TruSeq Nano gel free libraries from all five DNA extractions and sequenced them on an Illumina NovaSeq 6000. EG also generated a SMRTbell sequencing library from the HMW DNA and sequenced it on a Pacbio Sequel I instrument. A sixth individual (ES_BH_1411; Table D.1) was used for chromatin conformation capture (HiC) sequencing. EG performed the HiC reaction using an Arima-HiC kit, following the manufacturer's instructions for flash frozen animal tissue, and generated a TruSeq library which was sequenced on an Illumina NovaSeq 6000.

5.3.2 Genome assembly

We generated a reference genome for *B. hecate* by assembling Pacbio continuous long reads with Nextdenovo v2.4.0 (Hu et al. 2023). The contig sequences were polished with Illumina short-reads from the same individual using Hapo-G v1.1 (Aury and Istace 2021). We identified and removed haplotypic duplicates and contigs deriving from other organisms using purge_dups v1.2.5 (Guan et al. 2020) and Blobtools v1.1.1 (Laetsch and Blaxter 2017), respectively. We mapped HiC data to the contigs with bwa-mem v0.7.17 (Li 2013) and then used YaHS v1.1a.2 and juicebox v1.11.08 to scaffold the assembly into chromosome-level sequences (Zhou et al. 2023; Robinson et al. 2018).

5.3.3 Synteny analysis

We compared synteny between five genome assemblies of butterfly species in the tribe Argynini: *Brenthis hecate*, *Brenthis ino* (Mackintosh et al. 2022), *Brenthis daphne* (Mackintosh et al. 2023), *Fabriciana adippe* (Lohse et al. 2022b), and *Boloria selene*

(Lohse et al. 2022a). Pairwise alignment of assemblies were performed with minimap2 v2.17 (Li 2018) and differences in synteny were visualised by plotting high quality alignments (mapping quality of 60 and length >= 50 kb). We found that the genome sequence of *B. selene* has low sequence identity to the other genomes, resulting in few nucleotide alignments. We therefore identified BUSCO genes in all five assemblies (lepidoptera_odb20, BUSCO v5.3.2, Simão et al. 2015) and used the location of these BUSCO genes to visualise synteny between the *B. selene* genome and the others.

We estimated the number of fission and fusion rearrangements across the phylogeny of these species using syngraph (`https://github.com/A-J-F-Mackintosh/syngraph`). We included an additional nymphalid genome assembly in this analysis (*Nymphalis polychloros*, Lohse et al. 2021) as an outgroup. BUSCO genes were used as markers and the minimum number of markers for a rearrangement to be reported was set to five. We used the tabulated output of syngraph, as well as the paf files generated by minimap2, to identity approximate positions of chromosome fusion points.

5.3.4 Fitting a multi-species demographic history

To infer a demographic history for the three *Brenthis* species, we mapped whole genome resequencing (WGS) data to the *F. adippe* reference genome. This included data for five *B. hecate* individuals (Table D.1), as well as seven *B. daphne* and six *B. ino* individuals that were originally analysed in Mackintosh et al. (2023). Individuals were sampled from across the Palearctic (Table D.1, see Figure 1 in Mackintosh et al. 2023), including different glacial refugia.

WGS data were trimmed with fastp v0.2.1 (Chen et al. 2018) and mapped with bwa-mem. Variants were called with freebayes v1.3.2 (Garrison and Marth 2012) and filtered with gIMble preprocess (Laetsch et al. 2022) using the following options: `--snpgap 2 --min_qual 10 --min_depth 8 --max_depth 5`. Here `--snpgap` is the minimum distance a SNP can be from an indel, `--min_qual` is the minimum quality score of a SNP, `--min_depth` is the minimum absolute read depth and `--max_depth` is the maximum read depth relative to the sample-specific mean. We applied an additional filter to remove SNPs where > 70% of individuals were heterozygous, as these are likely due to alignment of paralogous sequence. We annotated genes in the *F. adippe* genome (see Supplementary Methods) and used this to restrict our analysis to fourfold-degenerate (4D) sites.

Given 295,730 SNPs, as well as a total count of 4D sites callable across all individuals (2,487,949), we generated an unfolded three dimensional site frequency spectrum (3D-SFS) using get_3D_SFS.py (see Data accessibility). The ancestral state at each SNP was assigned using the reference (*F. adippe*) allele. After inspection of the 3D-SFS, we chose to fold the data due to an excess of high frequency derived alleles that likely represent polarisation error.

Demographic modelling was performed with fastsimcoal2 (fsc27093) (Excoffier et al. 2021). We fit a model of divergence with gene flow between the three *Brenthis* species which included two split times, six effective population sizes (N_e), and eight asymmetrical effective migration rates (m_e) (16 parameters total, Figure 5.2C, Table D.2). Each N_e and m_e parameter within this model remains constant between speciation events. The parameter estimates with the greatest composite likelihood were recorded as point estimates, and we performed 100 parametric bootstraps to obtain 95% confidence intervals (95% CIs). The lower 95% CIs were calculated by interpolating between the 2nd and 3rd percentiles and the upper 95% CI was calculated by interpolating between the 97th and 98th. Demographic parameter estimates were scaled using a *de novo* mutation rate of 2.9×10^{-9} (Keightley et al. 2015). The fastsimcoal2 commands used are listed in the Supplementary Methods.

5.3.5 Identifying runs of homozygosity

We identified runs of homozygosity (ROH) in each *Brenthis* individual to gain more information about genetic drift in the recent past of these species. To do this, we mapped WGS data for each *Brenthis* species to the corresponding (species-specific) reference genome with bwa-mem. Variants were called within each species using freebayes and filtered with gIMble preprocess: `--snpgap 2 --min_qual 10 --min_depth 8 --max_depth 1.5` (see above for an explanation of these options). We restricted SNPs in the VCF to non-repeat regions where all individuals had a genotype. We then identified runs of homozygosity (ROH) in each individual using plink v1.90b6.18 (Purcell et al. 2007) with the following options: `--homozyg-window-snp 1000` `--homozyg-window-het 10 --homozyg-window-threshold 0.001 --homozyg-kb 100`. See Meyermans et al. (2020) for a description of these options and their effect on ROH identification.

5.3.6 Inferring selective sweeps from blockwise mutation configurations

We fit selective sweep models to 12 chromosome fusions by considering patterns of mutation within 1 Mb of sequence surrounding each rearrangement. Each fusion is private to one of the *Brenthis* species, i.e. we did not include fusions shared by multiple species which likely fixed many generations ago. Two of the 12 fusions were not inferred by the maximum parsimony method described above. Syngraph inferred a single ancient fusion and then a subsequent fission in *B. daphne*. Independent fusions in *B. hecate* and *B. ino* are equally parsimonious and supported by the fact that different chromosome ends are involved in each case. We therefore include these potential fusions in our analysis.

We used the same species-specific filtered VCF files described above as data for inferring selective sweeps. We annotated genes in the assemblies (see Supplementary Methods) and removed SNPs within exons (+/- 10 bases), i.e. we only consider variation within intronic or intergenic sequence. We chose to analyse n = 4 diploids for each species, selecting the set of individuals that minimised pairwise intraspecific F_{st}. We summarised the sequence variation surrounding each fusion in terms of the blockwise site frequency spectrum (bSFS; Bunnefeld et al. 2015), setting a block size so that the average block contained 1.5 segregating sites. We used six_lineage_bSFS.py (see Data accessibility) to record the folded bSFS for six lineages by considering all possible sets of three diploids from n = 4. We then applied a k_{max} value of 2 using format_blocks.py (see Data accessibility), i.e. we recorded exact mutation counts up to a value of 2 in each block and any count greater was summarised as $>$ 2. In summary, each block contains counts of folded singleton, doubleton, and tripleton mutations from a sample of six genomes.

We implemented the sweep inference method of Bisschop et al. (2021) in *Mathematica* (see Data accessibility). In this method, the composite likelihood of a selective sweep is calculated by multiplying the probabilities of observing mutation configurations in short sequence blocks. The probabilities of different mutation configurations (bSFS entries) depend on the parameters of the sweep model (θ, α and T_a, see Main Text) as well as the distance of a block from the sweep centre. For a given point in the genome, we estimated the composite likelihood of a neutral model and a selective sweep model given the bSFS counts in the surrounding 1 Mb region. We normalised the difference in composite likelihood ($\Delta \ln CL$) by the number of blocks to allow comparisons between 1 Mb regions with a different number of blocks. In cases where chromosome fusion points could only be narrowed down to intervals spanning $>$ 5 kb, we sampled points every 5 kb and reported parameter values for the point with the greatest $\Delta \ln CL$ (Table 5.1). As a comparison, we also fit sweep models to points sampled from a non-rearranged chromosome (Figure 5.1). Additional details of the model fitting procedure can be found in the Supplementary Methods.

5.3.7 Simulations

To quantify the power and accuracy of sweep inference based on the bSFS, we performed coalescent simulations with msprime v1.0.2 (Baumdicker et al. 2021) and applied the sweep inference scheme to the simulated data. Three different scenarios were simulated: a strong selective sweep (s = 0.005, T_a = 250,000 generations ago, with N_e = 500,000), neutral evolution (N_e = 500,000), and neutral evolution in a population with a similar demographic history to *B. daphne* (as inferred by fastsimcoal2). The mutation and recombination rates were set to $\mu = r = 2.9 \times 10^{-9}$ per-site per-generation. Each simulation was replicated 100 times, where a single replicate consisted of a 1 Mb sequence sampled for n = 4 diploids.

5.3.8 Statistical analysis

We used resampling tests to evaluate whether chromosome fusions are enriched for selective sweeps when compared to loci elsewhere in the genome. We measured two statistics – the number of fusions with putative sweeps and the sum of $\Delta \ln CL$ across all fusions in each species – and compared these with points sampled from a non-rearranged chromosome (Figure 5.1). Resampling was species-specific, i.e. we sampled the same number of points as fusions analysed for each species. We generated 100,000 random sample sets and calculated one-tailed p-values as the proportion of samples with values greater than our observed statistics.

5.4 Results

5.4.1 A genome assembly of *Brenthis hecate*

We generated a chromosome-level genome assembly for *Brenthis hecate* using a combination of Pacbio long-reads, Illumina short-reads, and HiC data (Figure D.1). The assembly is 408.8 Mb in length, with a scaffold N50 of 12.8 Mb and a contig N50 of 5.9 Mb. Of the 45 sequences in the assembly, 34 are chromosome-level (herafter simply referred to as chromosomes), whereas the remaining 11 are contigs that could not be scaffolded (15 - 104 kb in length, totalling 409 kb). The chromosomes show a bimodal distribution in size (Figure D.1), with seven large chromosomes (21.5 - 29.0 Mb) and 27 smaller chromosomes (6.6 - 14.0 Mb). The number of chromosomes in the *B. hecate* genome assembly (n_C = 34) is consistent with reports of spermatocytes sampled from both France and Siberia (de Lesse 1961; Saitoh and Lukhtanov 1988). The genome sizes of *B. hecate*, *B. daphne* and *B. ino* are all similar: 409, 419, and 412 Mb, respectively.

5.4.2 Synteny between Argynnini butterfly species

To characterise chromosome rearrangements, we performed whole genome alignments between the three *Brenthis* species, as well as genome assemblies from two other fritillary butterfly genera in the tribe Argynnini. The whole genome alignments show that the two outgroup species, *Fabriciana adippe* and *Boloria selene*, have very similar genome / chromosome organisation (Figure 5.1). By contrast, genomes of the *Brenthis* species show evidence for many rearrangements (Figure 5.1).

We next placed fission and fusions events on the phylogeny (*Brenthis sp.* and outgroups) using a maximum parsimony method (see Methods). Of the 53 inferred rearrangements, 50 are found on branches leading to *Brenthis* species or their most recent common ancestors. The branch with the greatest number of inferred rearrangements (11 fissions and 9 fusions) is that leading to the common ancestor of the three *Brenthis* species. Closer to the present, 14 fusion rearrangements are estimated on the branch

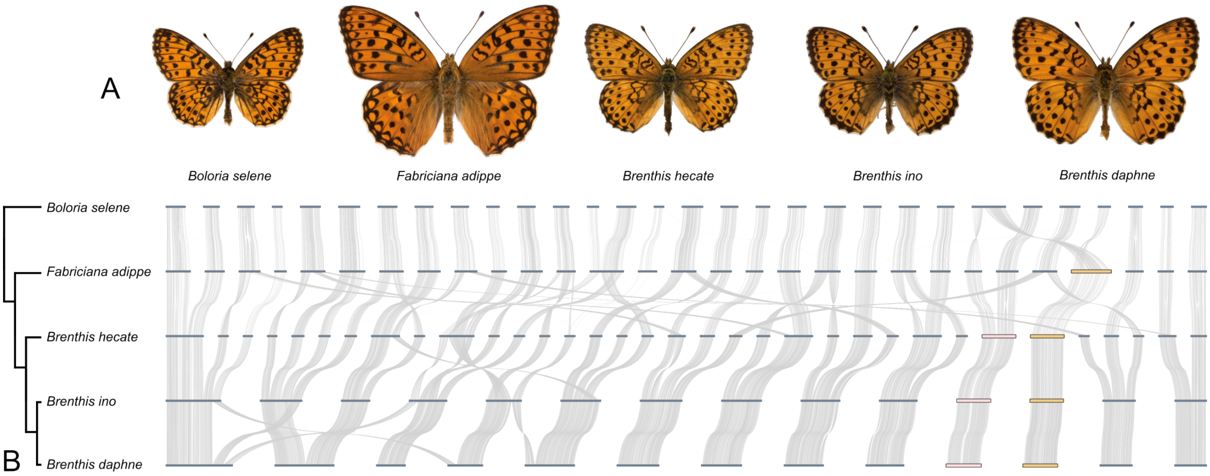

Figure 5.1: Synteny relationships between genomes of three *Brenthis* species, as well as species from two related genera. **(A)** Uppersides of male butterflies representing each of the five species. **(B)** A tree showing phylogenetic relationships between the species. The topology is from Chazot et al. (2021) and the plotted branch lengths are not to scale. Whole genome alignments are shown to the right of the tree. Thick horizontal bars are chromosomes and curved lines are nucleotide alignments, or, in the case of *B. selene* and *F. adippe*, shared BUSCO genes. Two sets of orthologous chromosomes are highlighted towards the right of the plot: an autosomal chromosome shared by all three *Brenthis* species (pink) and the Z chromosome that is shared by all *Brenthis* species and *F. adippe* (orange).

ancestral to *B. daphne* and *B. ino*, while five fissions and two fusions are estimated on the branch leading to *B. hecate*. These rearrangements explain the large difference in chromosome number between these species. We also infer one fission and five fusions on the branch leading to *B. daphne* and three fusions on the branch leading to *B. ino*. Together these rearrangements form a complex history of genome 'reshuffling' that is not seen in the outgroup lineages.

5.4.3 The demographic history of *Brenthis* butterflies

To estimate the timing of rearrangements as well the effective size of the populations in which they fixed, we inferred a multi-species demographic history using allele frequencies in resequenced genomes (see Methods). The best fitting demographic model estimates the *B. daphne* and *B. ino* split at 2.8 MYA and the split with *B. hecate* at 3.2 MYA (Figure 5.2C). These speciation times allow for an estimation of the rearrangement substitution rate per-genome and generation: 3.3×10^{-6}, i.e. one rearrangement every $\sim$ 300 k generations.

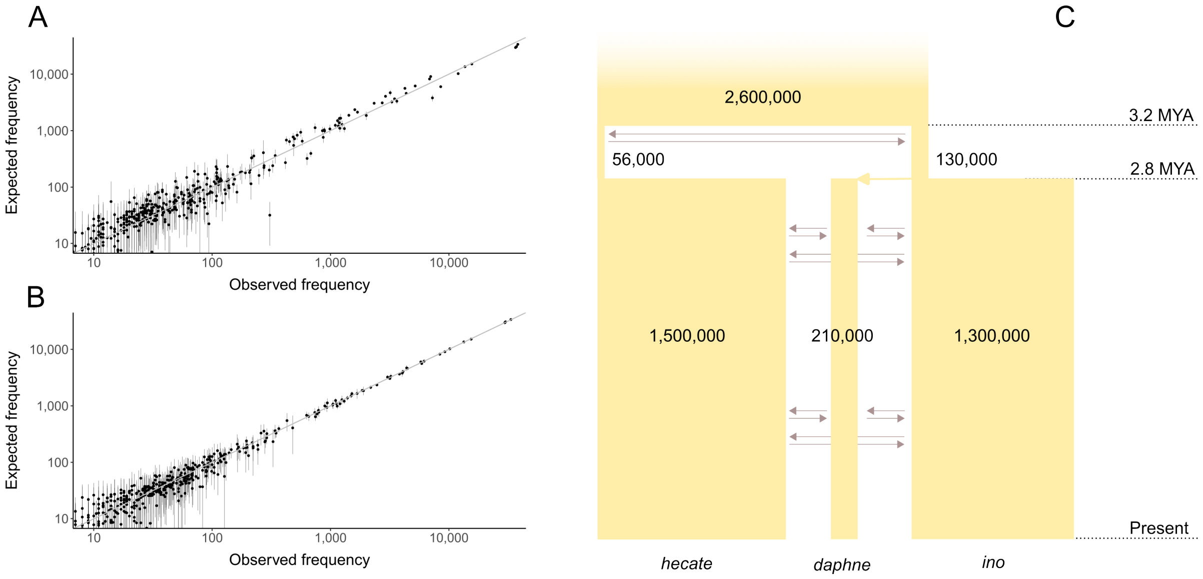

Figure 5.2: A demographic history of divergence and gene flow between three species of *Brenthis* butterfly. **(A)** A comparison between the expected and observed 3D-SFS given the estimated demographic history. Each point represents a single SFS entry, with its position determined by its expected frequency (y-axis) and the frequency observed in the data (x-axis). The diagonal line ($x = y$) represents a perfect fit between the model and the data and errors bars represent 95% CIs estimated from simulation replicates. **(B)** The same as in (A) but observed frequencies are derived from a single simulation. The correlation therefore represents the expected fit when data is simulated under that exact model. **(C)** A schematic representing the estimated demographic history. Each rectangle represents a population with width proportional to effective size (N_e). The N_e of each population is also given to two significant figures. Grey horizontal arrows represent the fact that contemporaneous populations exchange migrants in both directions. The timing of speciation events is shown on the right of the plot in units of a million years (1 generation = 1 year).

Overall genetic diversity in these species ($\sim$ 1% at 4D sites, Table D.3) is typical of butterflies (Mackintosh et al. 2019; Ebdon et al. 2021), suggesting that long-term N_e is on the order of 10^5 or greater. We co-estimated N_e and m_e parameters within the multi-species demographic model, thus taking into account the effect of interspecific gene flow on diversity. We find that N_e estimates of species and ancestral populations vary but are generally high, as expected (Figure 5.2C). The population in which the most rearrangements fixed (the ancestor of *B. daphne* and *B. ino*) has a relatively small N_e (1.3×10^5). By contrast, the population in which the fewest rearrangements fixed (*B. ino*) has a much larger N_e (1.3×10^6). While this may hint at a negative relationship between N_e and rearrangement rate, we cannot meaningfully test this from such a small species tree. Nonetheless, the fact that these species have large effective population sizes, as is typical of insects, suggests that rearrangements do not require extremely small long-term N_e to become fixed. However, this demographic model only partially fits the data (Figures 5.2A and 5.2B). The parameter estimates for ancestral populations also have wide 95% CIs (Table D.2), and therefore the N_e estimates from this model are only approximate (see Discussion).

5.4.4 Runs of homozygosity

The SFS contains information about long-term N_e, whereas regions of the genome that are identical by descent are informative about N_e in the recent past. With this in mind, we searched for runs of homozygosity (ROH) within individual genomes. Large ROH ($>=$ 1 Mb) are generated through recent shared ancestry and should be rare (for $N_e \approx 10^5$) or almost absent (for $N_e \approx 10^6$) in well-mixed populations. For example, the probability that a 1 Mb window is covered by a ROH in a population with effective size equivalent to *B. ino* is 7×10^{-5} when assuming a conservatively low recombination rate of $r = 2.9 \times 10^{-9}$. This corresponds to a probability of only 0.0284 that at least one ROH $>=$ 1 Mb is observed within a 412 Mb genome. Surprisingly, we found that the majority of individuals across all three species (15 of 18) have at least one ROH of this size (Figure 5.3A). Summing the length of these ROH to estimate the inbreeding coefficient F_{roh} reveals that there are several individuals with $F_{\text{roh}} \approx 1/16$ (Figure 5.3), consistent with being the offspring of first cousins. These results suggest that short-term N_e within local populations may be much lower than indicated by overall, i.e. species-wide, levels of diversity or predicted by the SFS-based model of demographic history. Although smaller local populations may promote the fixation of rearrangements through drift, it is less clear whether this would lead to fixation across the entire species range (see Discussion).

5.4.5 Parameter estimates and statistical support for simulated selective sweeps

It is possible that the rearrangements observed in this genus have become fixed through natural selection rather than drift (see Introduction). To test this, we ask whether loci surrounding chromosome fusions show evidence for selective sweeps. The sweep inference presented in Bisschop et al. (2021) calculates the likelihood of a hard selective sweep given mutation counts in short sequence blocks (the bSFS). While Bisschop et al. (2021) used unfolded mutation counts for four lineages, this

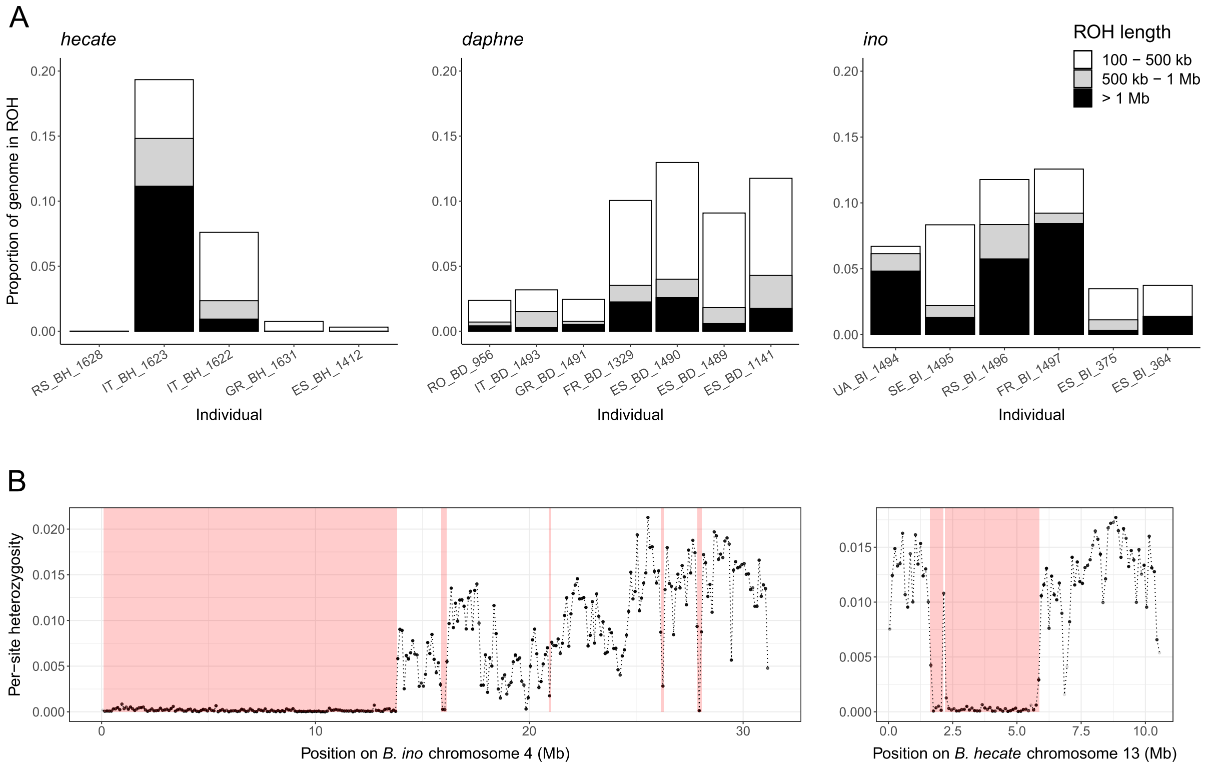

Figure 5.3: Evidence for inbreeding among *Brenthis* butterflies. **(A)** The fraction of the genome covered by runs of homozygosity (ROH) in each *Brenthis* individual. **(B)** Per-site heterozygosity for *B. ino* individual FR_BI_1497 plotted in 100 kb windows across chromosome 4, and the same for *B. hecate* individual IT_BH_1623 across chromosome 13. Red shading shows regions that were identified as ROH.

requires polarisation, i.e. knowledge of ancestral states. We can only obtain this information for genic regions of the genome given the considerable divergence between *Brenthis sp.* and the nearest available outgroup, *F. adippe* ($\sim$ 0.09 at 4D sites). We therefore adapted the composite likelihood based sweep inference to folded mutation counts for six lineages (Figure 5.4A).

We first tested whether this implementation can accurately infer old sweeps. We simulated strong selection ($N_e s = 2,500$, see Methods) and estimated the statistical support for a sweep while also obtaining maximum composite likelihood estimates (MCLE) for three parameters: θ, α and T_a. Here, $\theta = 4N_e * \mu * l$ is the population mutation rate per-block (where l is the block length), $\alpha = \frac{r}{s}\ln[2N_e * s]$ is the rate of recombination relative to the strength of the sweep, and T_a is the timing of the sweep in units of $2N_e$ generations. Across simulations, we find that the statistical support for a sweep – i.e. the increase in composite likelihood ($\Delta \ln CL$) compared to the best fitting neutral

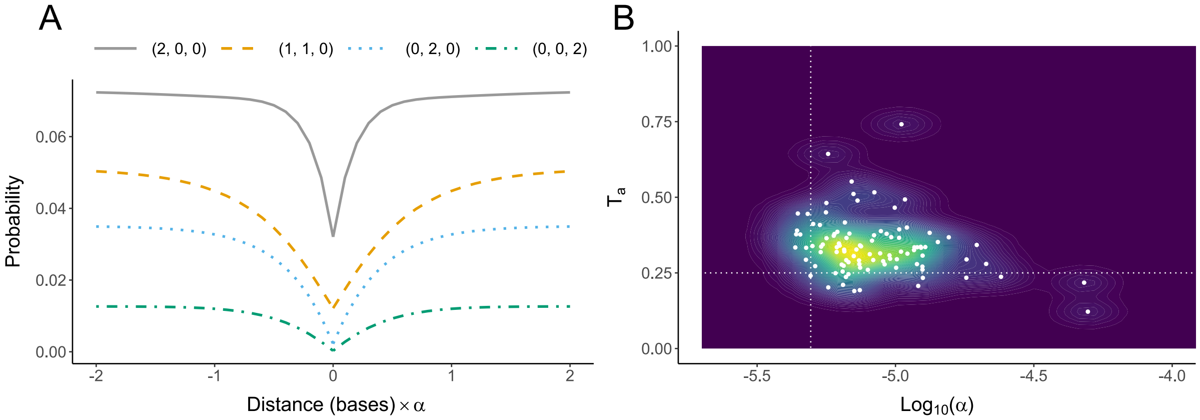

Figure 5.4: Inferring sweeps from the bSFS. **(A)** The probability of observing particular bSFS entries (y-axis) given the distance of a block from the sweep centre (x-axis). In this example, the sweep occurred $0.5N_e$ generations ago (i.e. T_a = 0.25) and the population mutation rate is 0.66 (e.g. $\theta_{per-site}$ = 0.0058 and block length = 113 bases). For a sample of six lineages with folded mutations and counting up to two mutations per branch type, there are 64 total bSFS entries. Each line represents one of these entries, where (i, j, k) denotes a block with i singleton mutations, j doubletons, and k tripletons. The x-axis shows that the effect of a sweep at a particular locus depends on the relative strength of the sweep (α) and the distance of that locus from the sweep centre. **(B)** Parameter estimates for 100 simulation replicates of a selective sweep. The true sweep strength ($Log_{10}(\alpha)$, x-axis) and timing (T_a, y-axis) are shown with dotted vertical and horizontal lines, respectively. Each point represents parameter estimates for a single simulation replicate, and coloured contours show the density of these estimates across all replicates.

model – is always non-zero with a median $\Delta \ln CL$ per-block of 0.018. The per-block mutation rate (θ = 0.66 for l = 113) is well estimated, albeit with a small downward bias (lower quartile, median, upper quartile = 0.62, 0.64, 0.65, respectively). Similarly, the timing (T_a) and strength (α) of the sweep are slightly overestimated and underestimated, respectively (Figure 5.4B). These results show that sweep parameters can be inferred through this method under a simple demographic history.

Repeating this analysis but simulating entirely neutral evolution (see Methods) leads to inferred sweeps that are very weak or, in a minority of cases, strong but very old (Figure D.2) and weakly supported: the median $\Delta \ln CL$ across these simulations is 3.3×10^{-5} and the maximum is 0.002. Given these results, we use thresholds of $Log_{10}(\alpha) < -4$ and $\Delta \ln CL > 0.002$ to define plausible sweep candidates. This α value implies a distortion of genealogical branch lengths across at least 20 kb (Figure 5.4) and corresponds to $s = 1.7 \times 10^{-4}$ given an N_e of 1×10^6 and a recombination rate of 2.9×10^{-9}. At our chosen thresholds, we may discard some weak selective sweeps but false-positives should be rare.

5.4.6 Evidence for an enrichment of selective sweeps around chromosome fusions

We next applied the same inference procedure to a total of 12 potentially recent chromosome fusions, with five, three, and four fusions sampled from *B. daphne*, *B. hecate*, and *B. ino*, respectively. Four fusions show no statistical support for a selective sweep (Table 5.1). The remaining eight fusions have $Log_{10}(\alpha)$ and $\Delta \ln CL$ values that our simulations suggest are unlikely to be observed under neutral evolution (see above). To test whether chromosome fusions have greater statistical support for selective sweeps than other regions of the genome, we fit sweep models across an entire chromosome for each species ($\sim$ 250 points spaced 100 kb apart). We chose the same orthologous chromosome for each species - the only autosome that has not undergone any rearrangements within the genus (Figure 5.1). Summing the $\Delta \ln CL$ of the 12 fusions and comparing this to points sampled from these non-rearranged chromosomes suggests that there is no strong enrichment for signals of selective sweeps (observed = 0.269, one-tailed 95% CIs of permutations = [0, 0.349], one-tailed p-value = 0.161). Similarly, although eight of the 12 fusions show evidence of a selective sweep, this result is not a significant departure from what can be obtained by sampling points from the non-rearranged chromosomes (one-tailed 95% CIs of permutations = [0, 8], one-tailed p-value = 0.060).

The fact that we infer sweeps around some chromosome fusions, but that this is unremarkable when compared to other regions of the genome, suggests a much higher false-positive rate in the real data than in our idealised neutral simulation check. Considering points sampled across the non-rearranged chromosome, we find that 26.9% are classified as sweeps both in *B. hecate* and *B. ino*, although the vast majority of these are old ($T_a > 0.5$, Figure D.2). In *B. daphne* the frequency of inferred sweeps is even higher at 63.2%, and, in contrast to the other species, these sweeps are almost always estimated to be recent ($T_a \approx 0.1$, Figure D.2). A plausible explanation for this is that gene flow into *B. daphne* from *B. ino* has generated genealogical histories that are better explained by a model of recent sweeps than a single panmictic population.

Table 5.1: Maximum composite likelihood parameter estimates for selective sweeps around chromosome fusions in *Brenthis* butterflies. The sweep with the greatest statistical support is highlighted in bold.

Taxon	**Chr**	**Position (Mb)**	$Log_{10}(\alpha)$	*Ta*	$\Delta \ln CL$ **per-block**
B. daphne	1	36.7	-2.29	1.00	.000
B. daphne	1	45.2	-4.71	0.10	.013
B. daphne	**2**	**23.9**	**-5.67**	**0.08**	**.123**
B. daphne	3	7.4	-5.70	0.09	.086
B. daphne	8	6.7	-5.70	0.25	.016
B. hecate	2	6.8	-4.60	0.35	.007
B. hecate	2	19.4	-5.70	0.89	.012
B. hecate	5	15.0	-4.08	1.00	.000
B. ino	1	6.6	-4.43	1.00	.000
B. ino	3	24.1	-5.70	1.00	.008
B. ino	8	22.2	-5.70	1.00	.004
B. ino	9	7.4	-5.03	1.00	.000

Simulating data for a single population that has undergone the long-term demographic history inferred for *B. daphne* (Figure 5.2C) and fitting a sweep model to these data, we recovered false-positive sweeps (32.0% of simulations), albeit with older inferred ages ($T_a \approx 0.5$, Figure D.2). This suggests that gene flow and changes in N_e over time likely explain at least some of the signatures of selective sweeps around chromosome fusions.

5.4.7 Evidence for individual selective sweeps

We next consider the strength of evidence for individual sweeps around chromosome fusions. The fusion with the strongest sweep support is on chromosome 2 of the *B. daphne* genome and has a per-block $\Delta \ln CL$ (Figure 5.5) that is greater than 95% of points sampled from the non-rearranged chromosome. The inferred sweep parameters (θ = 0.85 for l = 210, $Log_{10}(\alpha)$ = -5.7, T_a = 0.079) correspond to an s of 0.0012 and a timing of 56 k generations ago when assuming $\mu = r = 2.9 \times 10^{-9}$. Visualising counts of folded mutation classes shows a scarcity of tripleton and doubleton mutations near the fusion, as well as an overall reduction in diversity (Figure 5.5). In fact, there is a 264 kb region which encompasses the fusion point that does not have a single tripleton mutation (i.e. a mutation shared by three out of six lineages, Figure 5.5).

It is possible that the reduction in diversity around this chromosome fusion has been

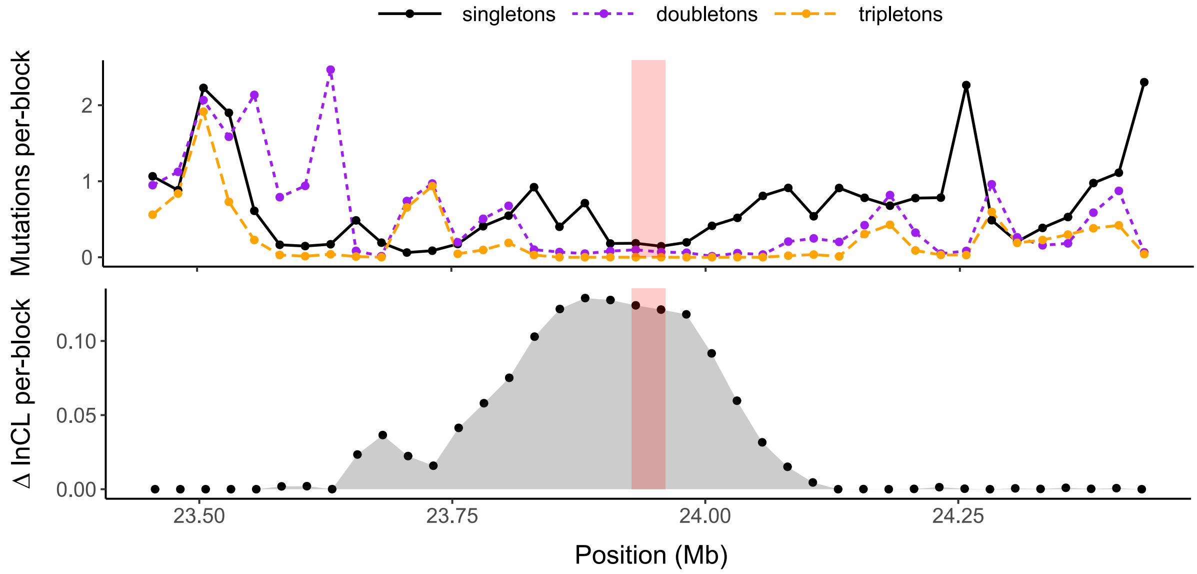

Figure 5.5: A potential selective sweep around a chromosome fusion on *B. daphne* chromosome 2. The top panel shows the frequency (y-axis) of the three folded mutations classes across the 1 Mb of sequence that surrounds the fusion point (x-axis). Mutation frequencies are plotted in 25 kb windows. The bottom panel shows the statistical support for a selective sweep (y-axis) across the same region. Support is measured as the difference in composite likelihood ($\Delta \ln CL$) between a selective sweep model and a neutral model. The models were fit at 40 test points at 25 kb intervals across this region, each represented by a point in the plot. The transparent red bar in the centre of both panels marks a 33 kb region which contains the fusion point

generated by processes other than a selective sweep, e.g. a lower *de novo* mutation rate or background selection. We therefore fit an alternative model to this region in which the fusion point is encompassed by a local reduction in θ that extends for *d* bases in either direction. We find that this model of locally reduced diversity (θ = 1.17, θ_{local} = 0.40, *d* = 400 kb) fits better than the neutral model with a single parameter (θ = 0.55, per-block $\Delta \ln CL$ = 0.110) but not as well as the sweep model (per-block $\Delta \ln CL$ = –0.013). Finally, we also test whether the sweep is supported when considering all seven *B. daphne* genomes, rather than just the four originally analysed. Under this sampling, the inferred sweep is of a similar strength but older (T_a = 0.26) and with reduced statistical support (per-block $\Delta \ln CL$ = 0.057 rather than 0.123). Given the confounding effects of demography, we must interpret patterns of mutation around this chromosome fusion carefully. Nonetheless, our results do raise the possibility that this chromosome fusion has risen in frequency due to positive natural selection.

5.5 Discussion

5.5.1 Patterns of chromosome evolution

Chromosome rearrangements are a fundamental part of eukaryote genome evolution, yet some groups of organisms display a much higher rate of rearrangement than others. We have focused on one such taxon, butterflies in the genus *Brenthis*, and have shown that these species have undergone a complex history of chromosome rearrangement not shared by other closely related genera (Figure 5.1). We find evidence for a large number of chromosome fusions as well as several fissions, with multiple rearrangements occurring between speciation events. We have assumed that chromosomes rearrange through fissions and fusions, rather than translocations. Although small segments of chromosomes appear to have translocated between *B. daphne* and *B. ino* chromosomes (Figure 5.1), the fact that these segments are single chromosomes in *B. hecate* suggests that these are ancestral chromosomes that have fused differentially. Overall, the pattern and tempo of rearrangement appears similar to what has been described in *Melinaea* butterflies (Nymphalidae) (Gauthier et al. 2022) and dissimilar from genera such as *Lysandra* (Lycaenidae) that are dominated by chromosome fissions (Wright et al. 2023). While it is perhaps unsurprising that the mode of rearrangement evolution differs between lineages, there do appear to be some shared features. For example, the Z sex-chromosome is one of just two chromosomes that are not rearranged between *B. hecate* and *B. daphne*, and although we have previously identified a Z-autosome fusion in one *B. ino* haplotype (Mackintosh et al. 2022), this rearrangement is not fixed. The Z is also the only chromosome that has not undergone extensive fissions in *Lysandra sp.*, and it is one of only two chromosomes that are not rearranged between *Melinaea marsaeus* and *M. menophilus* (Gauthier et al. 2022). It therefore seems likely that rearrangements involving the Z-chromosome have different fitness effects than autosomal rearrangements (Wright et al. 2023).

5.5.2 Genetic drift and underdominance

Fissions and fusions are likely to be underdominant, i.e. deleterious when in a heterozygous state, because proper pairing and segregation of chromosomes during meiosis is often impaired (Nunes et al. 2011; Grize et al. 2019, although see Mercer et al. 1992; Borodin et al. 2019). In that case, fixation of these rearrangements is due to strong genetic drift in small populations (Wright 1941). To investigate this possibility in *Brenthis* butterflies, we face a conundrum: the chromosome rearrangements we have investigated likely fixed at different time points spread across millions of years (Figures 5.1 and 5.2) for which we only have information about the long-term coalescent N_e. However, it is the short-term N_e that determines the fixation probability of a new mutation, and our observation of considerable ROH (Figure 5.3) suggests that this may be much lower than our long-term estimates. We therefore explore the fixation probability of a new rearrangement in both contexts.

Our estimates of long-term N_e from the SFS are on the order of $\sim 10^5$, with some variation between species and over time (Figure 5.2C). Given that we have estimated $N_e \sim 10^5$ and the rate of rearrangement fixation as 3.3×10^{-6}, we can use the fixation rate of Lande (1979) to estimate an upper-bound on the heterozygote disadvantage of rearrangements. Although we do not know the *de novo* rearrangement rate, we can assume that it is no higher than one rearrangement per-genome per-generation, as otherwise most individuals would be heterozygous for multiple new fissions or fusions (which we do not observe in our genome assemblies). Under this very conservative assumption, the maximum heterozygote disadvantage is $s = 1.4 \times 10^{-4}$, suggesting that heterozygosity for a fission or fusion has a weak absolute fitness effect in these species.

The above calculation assumes a large panmictic population, which is at odds with our observation of ROH within individual genomes (Figure 5.3). Observations of large long-term N_e yet considerable ROH can be reconciled by considering population structure within species. As an illustration, we consider the simplest possible scenario – a finite-island model (Maruyama 1970) – with ROH providing information about the proportion of recent within-deme coalescence. By fitting this model to three summary statistics (H, d_{xy} and W_{roh}, see Supplementary Methods), we find that a metapopulation with 260 demes, each with an N_e of 3,400 and an m_e of 4×10^{-4}, has the same expected levels of diversity, divergence and ROH as found among *B. ino* individuals. We stress that this calculation assumes the simplest possible model which is unlikely to capture the complex population structure that exists within these dispersive species. Nonetheless, it suggests that local populations of *Brenthis* butterflies may have a short-term N_e that is at least an order of magnitude smaller than overall diversity would suggest.

Is it then possible that population structure has facilitated the fixation of deleterious rearrangements through genetic drift? In the absence of migration, we can perform the same calculation as above for a local population with $N_e = 3,400$, and we find a much higher upper-bound of $s = 0.004$. Fixation in the total population, however, requires low levels of migration so that the rearrangement can still establish locally and spread through the population by extinction and re-colonization events (Lande 1979; Spirito et al. 1993). The m_e values we infer under a finite-island model suggest that migration between demes is high ($4N_em_e > 1$, Table D.3), in which case population structure can only have a weak effect on the fixation probability of an underdominant rearrangement (Slatkin 1981). We therefore conclude that there is not enough population structure in these butterfly species – at least in the very recent past – to allow the fixation of strongly underdominant rearrangements. Furthermore, given that we ignore the effect of selection at linked sites (Maynard Smith and Haigh 1974; Corbett-Detig et al. 2015), we are likely underestimating the short-term N_e of these butterfly species and therefore overestimating the probability that any of the chromosome rearrangements that have fixed in these species have appreciable underdominant fitness effects.

The idea that rearrangements in these species have only very small deleterious fitness effects is further supported by the fact that *B. daphne* and *B. ino* can produce fertile hybrids (Kitahara 2008, 2012) despite their karyotypes differing by as many as nine rearrangements (Figure 5.1). It is not clear how meiosis in these species is so robust to the risk of improper segregation in the presence of heterokaryotypes, although inverted meiosis is one potential explanation that has been described for other butterflies (Lukhtanov et al. 2018, 2020a). While we cannot calculate the exact fitness effects of rearrangements in *Brenthis* butterflies, we can at least rule out the possibility that strongly underdominant rearrangements (e.g. $s > 0.01$) have fixed through genetic drift.

5.5.3 The role of positive natural selection in the fixation of chromosome fusions

The scenario in which fusions are favoured by natural selection would mean that they play a role in adaptation (Yeaman 2013; Guerrero and Kirkpatrick 2014) and/or that they are driving as selfish elements. There is currently little empirical evidence that fusions fix through positive natural selection (although see Stewart et al. 2019), but this is unsurprising given that the majority of identified fusions are relatively old. For example, chromosome 2 of the human genome is the product of a fusion that happened approximately 900 kya (Poszewiecka et al. 2022), corresponding to $\sim 3.6N_e$ generations in the past. Inferring the evolutionary history of such old mutations is challenging given the fact that, on average, all but two lineages in a genealogical tree coalesce within $2N_e$ generations. We have therefore focused on species with recent chromosome fusions and a large long-term N_e (Figure 5.2C), giving us some power to detect the effects of natural selection.

We fit selective sweep models to 12 chromosome fusions and found that the aggregate statistical support is greater than what is found when sampling from a non-rearranged chromosome, but not significantly so. The simplest explanation for this result is that these fusions are selectively neutral and fixed by genetic drift. However, we cannot rule out the possibility that at least some fusions fixed through positive selection but did so $> 2N_e$ generations ago, with only a subtle signal remaining in present-day genome sequence data (Bisschop et al. 2021). We nonetheless interpret this result as evidence against a scenario where the majority of chromosome fusions fix through very strong selection, such as (holocentric) meiotic drive.

One fusion in our dataset, that on *B. daphne* chromosome 2 (Figure 5.5), has greater statistical support for a sweep than 95% of points sampled elsewhere in the genome. However, since we have considered 12 fusions, the probability that at least one meets this threshold by chance is considerable (p = 0.46). This fusion does, however, have greater support for a sweep than all 100 simulations performed under the *B. daphne* demographic history inferred under a multi-species model, and so sequence variation in this region cannot easily be explained by demography alone. The reduction in diversity around this chromosome fusion (Figure 5.5) could be explained by a recombination desert in which background selection continuously erodes diversity. Although this ad-hoc explanation could be applied to almost any inferred sweep, it is at least plausible in this case as the fusion point is in the centre of the chromosome where recombination is typically lowest in butterfly genomes (Shipilina et al. 2022; Palahí i Torres et al. 2022). Additionally, the fact that these fusions act as barriers to gene flow (Mackintosh et al. 2023) is another explanation for the reduction in diversity that we observe. A more general issue is that selective sweep signatures can also be generated by the fixation of deleterious mutations (Johri et al. 2021). This is because mutations with fitness effects s and $-s$ have the same expected fixation time (Maruyama and Kimura 1974). Such mutations would, however, have very different fixation probabilities so long as $2N_es >> 1$. We estimate $2N_es \approx 850$ for the sweep on *B. daphne* chromosome 2, making the fixation of a deleterious chromosome fusion an unlikely explanation for the sweep signature in this region.

Some uncertainty remains as to whether the inferred selective sweep around the chromosome fusion on *B. daphne* chromosome 2 is a true-positive, and we therefore interpret our results as weak evidence for the idea that chromosome fusions primarily fix through positive natural selection. The patterns of mutation around this particular fusion are nonetheless unusual and so warrant further exploration. Ideally, future analyses will jointly model the effects of demography and natural selection on sequence data, which is a long-standing goal in population genomic inference (Przeworski 2002; Jensen et al. 2005; Lauterbur et al. 2022).

5.5.4 Outlook

Knowledge about how genomes change over time is key for our understanding of evolution. Although fission and fusion rearrangements represent just a small fraction of the ways in which genomes can change, we know particularly little about how these drastic mutations become fixed in populations. To address this, we have analysed genome wide variation in *Brenthis* butterflies to infer past demography and natural selection in relation to chromosome rearrangements. Our main findings are that (i) drift is not strong enough to fix considerably underdominant rearrangements, and (ii) there is only weak evidence that chromosome fusions fixed through positive natural selection or meiotic drive. We cannot yet construct a full model of how rearrangements fix in these species, but our results are consistent with rearrangements having small fitness effects and fixing through drift. Clearly, other types of information not contained in genome sequence data are required for a full picture of how rearrangements fix. For example, direct estimates of heterokaryotype fitness (Knief et al. 2016; Luo et al. 2018) and *de novo* rates of rearrangement (Yamaguchi and Mukai 1974) are invaluable for understanding rearrangement evolution. Additionally, while we have focused on rearrangements that are likely to have fixed recently, an alternative strategy would be to identify and analyse the small subset of rearrangements that are still segregating within a species. It is more challenging to collect data on such examples but they could provide information about how rearrangements rise (and fall) in frequency over time. The population genomic analyses presented here represent a first step in understanding how fission and fusion rearrangements fix in *Brenthis* butterflies. We anticipate and look forward to similar investigations in other groups of organisms where chromosome rearrangements are common, which together will illuminate how genomes evolve across the tree of life.

5.6 Data availability

All new sequence data generated in this study and the *Brenthis hecate* genome assembly are available at the European Nucleotide Archive under project accession PR-

JEB62818. Python scripts and *Mathematica* notebooks are available at the following Github repository: https://github.com/A-J-F-Mackintosh/Mackintosh_et_al_2023_rearrangement_fixation.

General discussion

The establishment of new chromosome rearrangements is often suggested to be an important process in evolution (White 1978b; Navarro and Barton 2003; Kirkpatrick and Barton 2006; Lucek et al. 2022). Analysis of genome sequence data has shown that inversion rearrangements facilitate speciation and adpatation (Wellenreuther and Bernatchez 2018), but similar investigations into fission and fusion rearrangements have not yet generated the same strength of evidence. In this thesis I have used comparative and population genomic methods to investigate the role of fission and fusion rearrangements in evolution. Here I briefly discuss the main results of each chapter, as well as alternative methods for investigating chromosome rearrangements and population history.

6.1 Methods for investigating chromosome evolution

In chapter 3 I focused on the problem of inferring past inter-chromosomal rearrangements given multiple genome sequences and the phylogeny that relates them. I show that parsimony methods such as `syngraph` will underestimate rearrangements and infer incorrect ALGs as the rearrangement rate per-branch approaches the number of chromosomes. Given this result, how confident can we be in the rearrangements inferred between *Brenthis* butterflies in chapters 4 and 5? Reassuringly, the number of fissions and fusions estimated on each external branch is small (4 - 7) compared to the number of chromosomes (13 - 34), meaning that recent rearrangements are likely to be well estimated. However, the fact that the rearrangement histories inferred in chapters 4 and 5 differ slightly (a fission in *B. daphne* was instead estimated as a fusion in *B. ino* when the *B. hecate* genome was included in chapter 5) shows that the method is sensitive to the species included. Ideally, it would be possible to calculate a measure of confidence for each inferred rearrangement. A potential extension of `syngraph` would therefore be to sample rearrangements histories in proportion to their likelihood or posterior probability (Miklós and Tannier 2010; Miklós and Smith 2015).

Considering a single genome assembly per-species means that only chromosome variation between species can be identified. In chapter 2 I used HiC sequence data to identify a Z-autosome chromosome fusion (i.e. a neo-Z) in a heterozygous state (Figure 2.2). This shows that there are chromosome rearrangements segregating within *B. ino*, but also demonstrates that this type of sequence data could be used to identify rearrangements across a population. Importantly, population-level HiC data would give information about which individuals possess certain rearrangements, whether they are in a heterozygous or homozygous state, and their haplotypes. This would open up new possibilities for investigating recent chromosome rearrangements. For example, given this type of data, the selective sweep inference method used in chapter 5 could be adapted to (i) include phase information and (ii) fit a partial selective sweep model where haplotypes that lack the rearrangement always recombine out. Generating population-scale rearrangement data seems like a natural next step for improving our understanding of rearrangements in evolution (Kim et al. 2022; Liao et al. 2023; Orteu et al. 2023).

6.2 Methods for inferring the evolutionary history of populations

In chapter 4 I used gIMble (Laetsch et al. 2022) to investigate historic barriers to gene flow between *B. ino* and *B. daphne*, whereas in chapter 5 I adapted the selective sweep inference scheme of Bisschop et al. (2021) and applied it to loci surrounding chromosome fusions in *B. ino*, *B. daphne*, and *B. hecate*. Both of the methods derive the joint distribution of branch lengths using generating functions (Lohse et al. 2011) and calculate likelihoods given mutation configurations in short sequence blocks. Estimating m_e variation across the genome with gIMble is a significant step forward from genome-scans for barrier loci that rely on summary statistics (e.g. F_{st} and d_{xy}). Likewise, the method of Bisschop et al. (2021) can infer much older selective sweeps than SFS-based approaches (Nielsen et al. 2005) and can also estimate the timing of selection. These methods are therefore powerful tools for learning about the evolutionary history of populations, yet they have several limitations.

One assumption of these methods is that there is a single genealogy underlying the mutations in each sequence block, i.e. no intra-block recombination. Choosing a block length therefore involves a compromise between minimising the bias introduced by recombination (Figure C.3) and maximising the number of linked mutations. Although the bSFS retains useful linkage information not included in the SFS, it does not capture the long-range linkage disequilibrium that is required for inference of very recent population history. There are also limits to the complexity of models for which blockwise likelihoods can be calculated analytically. Although multiple exponentially distributed processes can be included in a given model (e.g. coalescence, migration, mutation), it is currently difficult to include more than one discrete event (e.g. a bottleneck or the merging of two populations) (Bisschop 2022, but see Lohse and Frantz 2014).

An alternative to the blockwise methods used in this thesis are those that directly infer the ARG (Rasmussen et al. 2014; Kelleher et al. 2019; Speidel et al. 2019). One can fit models to reconstructed ARGs (Stern et al. 2019), or instead focus on non-parametric summaries of ARG features, such as the distribution of coalescent times (Li and Durbin 2011). In chapter 5 I inferred a recent selective sweep around a chromosome fusion on *B. daphne* chromosome 2. Can a reconstructed ARG provide more information about past selection at this region of the genome? Interestingly, an inferred marginal genealogy (Figure 6.1) shows two geographically clustered bursts of recent coalescence, with the two remaining lineages taking $\sim 5N_e$ generations to coalesce. This is inconsistent with a hard selective sweep but could be the result of a soft sweep where the selected mutation has segregated at low frequency (Messer and Petrov 2013).

The above example shows that reconstructed ARGs contain useful information that is harder to obtain from blockwise inference methods. However, a fundamental issue is that we must assume a specific model when inferring an ARG that will usually differ from the true population history. ARGweaver, for example, samples ARGs from a posterior probability distribution while using a neutral coalescent prior (Rasmussen et al. 2014). It is worth considering how dependent the reconstructed marginal genealogy in Figure 6.1 is on this prior, and whether very different results would be obtained by inference under a model including natural selection or population structure (Hubisz et al. 2020).

Although the blockwise likelihood methods used in this thesis have limitations, it is not yet clear that reconstructing ARGs is a better alternative. One interesting idea would be to combine these approaches by applying the method of Lohse et al. (2011) to mu-

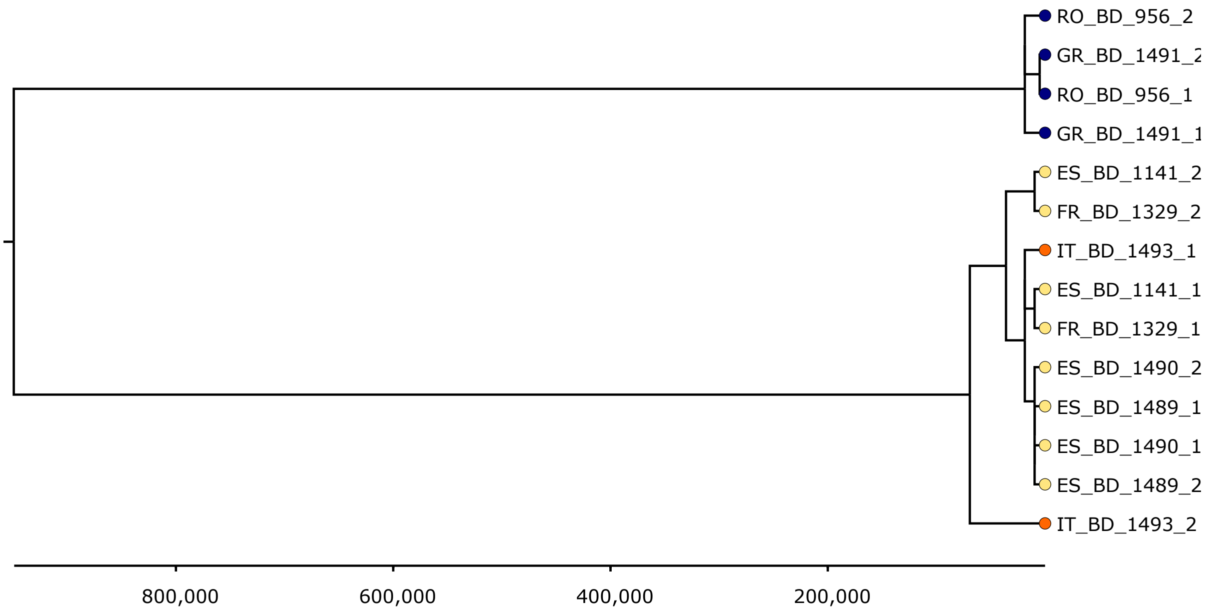

Figure 6.1: A marginal genealogy inferred by ARGweaver for position 23,943,769 on *B. daphne* chromosome 2 (see Appendix for details). The x-axis shows the time from the present, with ticks at intervals of approximately $1N_e$ generations. Tips are coloured by sampling location, with individuals sampled in Western Europe, Italy and Eastern Europe coloured in light yellow, orange and dark purple, respectively.

tations within marginal genealogies (inferred by an ARG reconstruction method), rather than short sequence blocks. This would provide access to the information contained within genealogies with large spans and therefore recent coalescent times, while also reducing the bias introduced by intra-block recombination.

6.3 The role of fission and fusion rearrangements in *Brenthis* butterflies

In chapter 4 I showed that fission and fusion rearrangements are associated with reduced post-divergence gene flow between *B. ino* and *B. daphne*. Previous research has typically focused on the outcomes of speciation, such as rates of diversification (Augustijnen et al. 2023) or present-day reproductive isolation (Yoshida et al. 2023). By contrast, I explicitly modelled the cessation of gene flow across the genome, allowing direct insight into the role of these rearrangements in the speciation process. Although the overall evidence that fissions and fusions have promoted speciation between *B. ino* and *B. daphne* is strong, the exact mechanisms preventing gene flow are unclear. Possibilities include underdominance due to meitoic breakdown, increased linkage disequilibrium between barrier loci, and crossover suppression.

Although it is challenging to discern between the above possibilities, heterozygosity for multiple rearrangements has been associated with reduced fertility in many taxa (Mercer et al. 1992; Castiglia and Capanna 2000; Yoshida et al. 2023), including butterflies (Lukhtanov et al. 2020b). Some level of meiotic breakdown in hybrids between *B. ino* and *B. daphne* therefore seems likely, but requires an explanation of how rearrangements with potentially underdominant fitness effects became fixed in each population. Given that there are a total of nine rearrangements between *B. ino* and *B. daphne*, it is possible that each has a weak fitness effect, allowing fixation, but that together they result in significant underdominance in early generation hybrids. This putative model of chromosomal speciation is supported by previous results showing that reproductive isolation will accumulate quickest when there are many rearrangements with underdominant fitness effects on the order of $1/N_e$ (Walsh 1982). Moreover, Walsh (1982) also showed that speciation due to a single rearrangement with strong underdominance is very unlikely, even when N_e is low.

Accepting the above model would mean that speciation involving fissions and fusions may be limited to the small subset of taxa where multiple rearrangements are able to accumulate during the early stages of divergence. However, chromosomal speciation could be much more common if (underdominant) rearrangements spread through natural selection or meiotic drive, rather than drift alone (Hedrick 1981; Walsh 1982). In chapter 5 I investigated this possibility. While there is certainly some evidence for a selective sweep coinciding with one chromosome fusion (Figures 5.5 and 6.1), the other 11 fusions only show small departures from a neutral model of evolution. As the exact timing of these rearrangements is uncertain, the selection / meiotic drive events could simply be too old to be detectable. Nonetheless, the results in chapter 5 are consistent with fissions and fusions having weak fitness effects and fixing through drift in large effective populations.

In summary, by investigating chromosome fissions and fusions in *Brenthis* butterflies I have demonstrated that these rearrangements do, at least sometimes, promote speciation in nature. There is still uncertainty around the details of exactly how this happens, but one promising scenario is that the step-wise accumulation of rearrangements eventually leads to underdominance in early generation hybrids. If this is the principle mode by which chromosomal speciation happens, then it must be rare, leaving fissions and fusions with only a minor role in evolution. However, we are still a long way from having a complete understanding of fission and fusion rearrangements in evolution, and this will only be obtained once investigations are performed across many different groups of organisms. Here I have provided a genomics-based framework for undertaking such research.

Bibliography

Adam Z, Sankoff D. 2008. The ABCs of MGR with DCJ. Evolutionary Bioinformatics. 4:69–74.

Aeschbacher S, Selby JP, Willis JH, Coop G. 2017. Population-genomic inference of the strength and timing of selection against gene flow. Proceedings of the National Academy of Sciences. 114:7061–7066.

Ahola V, Lehtonen R, Somervuo P, Salmela L, Koskinen P, Rastas P, Välimäki N, Paulin L, Kvist J, Wahlberg N et al. 2014. The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. Nature Communications. 5:4737.

Allio R, Schomaker-Bastos A, Romiguier J, Prosdocimi F, Nabholz B, Delsuc F. 2020. Mitofinder: Efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. Molecular Ecology Resources. 20:892–905.

Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, Fang Q, Xie D, Feng S, Stiller J et al. 2020. Progressive cactus is a multiple-genome aligner for the thousand-genome era. Nature. 587:246–251.

Augustijnen H, Baetscher L, Cesanek M, Chkhartishvili T, Dincă V, Iankoshvili G, Ogawa K, Vila R, Klopfstein S, de Vos J et al. 2023. A macroevolutionary role for chromosomal fusion and fission in *Erebia* butterflies. bioRxiv. Unpublished. doi: 10.1101/2023.01.16.524200.

Aury JM, Istace B. 2021. Hapo-G, haplotype-aware polishing of genome assemblies with accurate reads. NAR Genomics and Bioinformatics. 3. lqab034.

Baker RJ, Bickham JW. 1986. Speciation by monobrachial centric fusions. Proceedings of the National Academy of Sciences. 83:8245–8248.

Baril T, Imrie R, Hayward A. 2021. TobyBaril/EarlGrey: Earl Grey v1.2. Zenodo. https://doi.org/10.5281/zenodo.5718734.

Baril T, Imrie RM, Hayward A. 2022. Earl Grey: a fully automated user-friendly transposable element annotation and analysis pipeline. bioRxiv. Unpublished. doi: 10.1101/2022.06.30.498289.

Barnett DW, Garrison EK, Quinlan AR, Strömberg MP, Marth GT. 2011. BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics. 27:1691–1692.

Barton NH. 1998. The effect of hitch-hiking on neutral genealogies. Genetics Research. 72:123–133.

Barton NH, Charlesworth B. 1984. Genetic revolutions, founder effects, and speciation. Annual Review of Ecology and Systematics. 15:133–164.

Barton NH, Hewitt G. 1981. A chromosomal cline in the grasshopper *Podisma pedestris*. Evolution. pp. 1008–1018.

Basset P, Yannic G, Brünner H, Hausser J. 2006. Restricted gene flow at specific parts of the shrew genome in chromosomal hybrid zones. Evolution. 60:1718–1730.

Baumdicker F, Bisschop G, Goldstein D, Gower G, Ragsdale AP, Tsambos G, Zhu S, Eldon B, Ellerman EC, Galloway JG et al. 2021. Efficient ancestry and mutation simulation with msprime 1.0. Genetics. 220. iyab229.

Beeravolu CR, Hickerson MJ, Frantz LA, Lohse K. 2018. ABLE: blockwise site frequency spectra for inferring complex population histories and recombination. Genome biology. 19:1–16.

Bickham J, Baker R. 1979. Canalization model of chromosomal evolution., In: Models and methodologies in evolutionary theory, Pittsburgh, Carnegie Museum of Natural History. pp. 70–84.

Bidau CJ, Giménez MD, Palmer CL, Searle JB. 2001. The effects of robertsonian fusions on chiasma frequency and distribution in the house mouse (*Mus musculus domesticus*) from a hybrid zone in northern Scotland. Heredity. 87:305–313.

Bisschop G. 2022. Graph-based algorithms for Laplace transformed coalescence time distributions. PLOS Computational Biology. 18:1–13.

Bisschop G, Lohse K, Setter D. 2021. Sweeps in time: leveraging the joint distribution of branch lengths. Genetics. 219:iyab119.

Borodin P, Torgasheva A, Fedyk S, Chetnicki W, Pavlova S, Searle J. 2019. Meiosis and fertility associated with chromosomal heterozygosity, In: Shrews, chromosomes and speciation, Cambridge University Press. pp. 217–270.

Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvák Z, Levin HL, Macfarlan TS et al. 2018. Ten things you should know about transposable elements. Genome Biology. 19:199.

Bourque G, Pevzner PA. 2002. Genome-scale evolution: reconstructing gene orders in the ancestral species. Genome research. 12:26–36.

Boveri T. 1904. *Ergebnisse über die Konstitution der chromatischen Substanz des Zellkerns*. Verlag von Gustav Fischer in Jena.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using diamond. Nature Methods. 12:59–60.

Bunnefeld L, Frantz LAF, Lohse K. 2015. Inferring Bottlenecks from Genome-Wide Samples of Short Sequence Blocks. Genetics. 201:1157–1169.

Bureš P, Zedek F. 2014. Holokinetic drive: centromere drive in chromosomes without centromeres. Evolution. 68:2412–2420.

Bush GL. 1975. Modes of animal speciation. Annual Review of Ecology and Systematics. 6:339–364.

Bush GL, Case S, Wilson A, Patton JL. 1977. Rapid speciation and chromosomal evolution in mammals. Proceedings of the National Academy of Sciences. 74:3942–3946.

Bush ZD, Naftaly AF, Dinwiddie D, Albers C, Hillers KJ, Libuda DE. 2023. De novo genome assemblies reveal structural variations between laboratory and natural isolates of *C. elegans*. bioRxiv. pp. 2023–01. Unpublished. doi: 10.1101/2023.01.13.523974.

Capilla L, Medarde N, Alemany-Schmidt A, Oliver-Bonet M, Ventura J, Ruiz-Herrera A. 2014. Genetic recombination variation in wild Robertsonian mice: on the role of chromosomal fusions and Prdm9 allelic background. Proceedings of the Royal Society B: Biological Sciences. 281:20140297.

Castiglia R, Capanna E. 2000. Contact zone between chromosomal races of *Mus musculus domesticus*. 2. Fertility and segregation in laboratory-reared and wild mice heterozygous for multiple robertsonian rearrangements. Heredity. 85:147–156.

Charlesworth B. 1983. Models of the evolution of some genetic systems. Proceedings of the Royal society of London. Series B. Biological sciences. 219:265–279.

Chazot N, Condamine FL, Dudas G, Peña C, Kodandaramaiah U, Matos-Maraví P, Aduse-Poku K, Elias M, Warren AD, Lohman DJ et al. 2021. Conserved ancestral tropical niche but different continental histories explain the latitudinal diversity gradient in brush-footed butterflies. Nature communications. 12:5717.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 34:i884–i890.

Corbett-Detig RB, Hartl DL. 2012. Population genomics of inversion polymorphisms in *Drosophila melanogaster*. PLoS genetics. 8:e1003056.

Corbett-Detig RB, Hartl DL, Sackton TB. 2015. Natural selection constrains neutral diversity across a wide range of species. PLoS biology. 13:e1002112.

Cotton JA, Bennuru S, Grote A, Harsha B, Tracey A, Beech R, Doyle SR, Dunn M, Hotopp JCD, Holroyd N et al. 2016. The genome of *Onchocerca volvulus*, agent of river blindness. Nature Microbiology. 2:1–12.

Coyne JA, Meyers W, Crittenden AP, Sniegowski P. 1993. The fertility effects of pericentric inversions in *Drosophila melanogaster*. Genetics. 134:487–496.

Craddock E. 1970. Chromosome number variation in a stick insect *Didymuria violescens* (Leach). Science. 167:1380–1382.

Damas J, Kim J, Farré M, Griffin DK, Larkin DM. 2018. Reconstruction of avian ancestral karyotypes reveals differences in the evolutionary history of macro-and microchromosomes. Genome biology. 19:1–16.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM et al. 2021. Twelve years of SAMtools and BCFtools. GigaScience. 10. giab008.

Dapporto L, Cini A, Vodă R, Dincă V, Wiemers M, Menchetti M, Magini G, Talavera G, Shreeve T, Bonelli S et al. 2019. Integrating three comprehensive data sets shows that mitochondrial DNA variation is linked to species traits and paleogeographic events in European butterflies. Molecular Ecology Resources. 19:1623–1636.

DasGupta B, Jiang T, Kannan S, Li M, Sweedyk Z. 1997. On the complexity and approximation of syntenic distance. In: Proceedings of the first annual international conference on computational molecular biology. pp. 99–108.

Davey JW, Barker SL, Rastas PM, Pinharanda A, Martin SH, Durbin R, McMillan WO, Merrill RM, Jiggins CD. 2017. No evidence for maintenance of a sympatric *Heliconius* species barrier by chromosomal inversions. Evolution letters. 1:138–154.

Davisson MT, Akeson EC. 1993. Recombination suppression by heterozygous Robertsonian chromosomes in the mouse. Genetics. 133:649–667.

de Lesse H. 1960. Spéciation et variation chromosomique chez les lépidoptères rhopalocères. Ann. Soc. Nat., Zool.. pp. 1–223.

de Lesse H. 1961. Signification supraspécifique des formules chromosomiques chez les lépidoptères. Bulletin de la Société entomologique de France. 66:71–83.

de Vos JM, Augustijnen H, Bätscher L, Lucek K. 2020. Speciation through chromosomal fusion and fission in Lepidoptera. Philosophical Transactions of the Royal Society B: Biological Sciences. 375:20190539.

Di Stefano M, Di Giovanni F, Pozharskaia V, Gomar-Alba M, Baù D, Carey LB, Marti-Renom MA, Mendoza M. 2020. Impact of chromosome fusions on 3D genome organization and gene expression in budding yeast. Genetics. 214:651–667.

Dobzhansky T. 1934. Studies on hybrid sterility. Zeitschrift für Zellforschung und mikroskopische Anatomie. 21:169–223.

Dobzhansky T, Epling C. 1948. The suppression of crossing over in inversion heterozygotes of *Drosophila pseudoobscura*. Proceedings of the National Academy of Sciences. 34:137–141.

Doyle SR, Tracey A, Laing R, Holroyd N, Bartley D, Bazant W, Beasley H, Beech R, Britton C, Brooks K et al. 2020. Genomic and transcriptomic variation defines the chromosome-scale assembly of *Haemonchus contortus*, a model gastrointestinal worm. Communications biology. 3:656.

Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP et al. 2017. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science. 356:92–95.

Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL. 2016. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Systems. 3:95–98.

Dutheil JY, Ganapathy G, Hobolth A, Mailund T, Uyenoyama MK, Schierup MH. 2009. Ancestral population genomics: the coalescent hidden Markov model approach. Genetics. 183:259–274.

Dutrillaux B, Dutrillaux AM, McClure M, Gèze M, Elias M, Bed'Hom B. 2022. Improved basic cytogenetics challenges holocentricity of butterfly chromosomes. bioRxiv. Unpublished. doi: 10.1101/2022.03.11.484012.

Dutrillaux B, Rumpler Y. 1977. Chromosomal evolution in Malagasy lemurs. Cytogenetic and Genome Research. 18:197–211.

Ebdon S, Laetsch DR, Dapporto L, Hayward A, Ritchie MG, Dincă V, Vila R, Lohse K. 2021. The Pleistocene species pump past its prime: Evidence from European butterfly sister species. Molecular Ecology. 30:3575–3589.

Edge P, Bafna V, Bansal V. 2017. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. Genome Research. 27:801–812.

Emms DM, Kelly S. 2019. Orthofinder: phylogenetic orthology inference for comparative genomics. Genome biology. 20:1–14.

Escudero M, Marques A, Lucek K, Hipp AL. 2023. Genomic hotspots of chromosome rearrangements explain conserved synteny despite high rates of chromosome evolution in a holocentric lineage. Molecular Ecology. .

Excoffier L, Marchi N, Marques DA, Matthey-Doret R, Gouy A, Sousa VC. 2021. fastsimcoal2: demographic inference under complex evolutionary scenarios. Bioinformatics. 37:4882–4885.

Farré M, Kim J, Proskuryakova AA, Zhang Y, Kulemzina AI, Li Q, Zhou Y, Xiong Y, Johnson JL, Perelman PL et al. 2019. Evolution of gene regulation in ruminants differs between evolutionary breakpoint regions and homologous synteny blocks. Genome research. 29:576–589.

Federley H. 1938. Chromosomenzahlen Finnländischer Lepidopteren. Hereditas. 24:397–464.

Feijão P, Araujo E. 2016. Fast ancestral gene order reconstruction of genomes with unequal gene content. BMC bioinformatics. 17:187–200.

Felsenstein J. 1988. Phylogenies from molecular sequences: inference and reliability. Annual review of genetics. 22:521–565.

Feng B, Zhou L, Tang J. 2017. Ancestral genome reconstruction on whole genome level. Current Genomics. 18:306–315.

Ferretti V, Nadeau JH, Sankoff D. 1996. Original synteny. In: Combinatorial pattern matching: 7th annual symposium, cpm 96 laguna beach, california, june 10–12, 1996 proceedings 7. pp. 159–167. Springer.

Fertin G, Labarre A, Rusu I, Vialette S, Tannier E. 2009. *Combinatorics of genome rearrangements*. MIT press.

Feuk L, MacDonald JR, Tang T, Carson AR, Li M, Rao G, Khaja R, Scherer SW. 2005. Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. PLoS genetics. 1:e56.

Feulner P, De-Kayne R. 2017. Genome evolution, structural rearrangements and speciation. J Evol Biol. 30:1488–1490.

Fisher RA. 1923. On the dominance ratio. Proceedings of the royal society of Edinburgh. 42:321–341.

Fisher RA. 1930. *The genetical theory of natural selection*. Oxford University Press.

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020. Repeatmodeler2 for automated genomic discovery of transposable element families. Proceedings of the National Academy of Sciences. 117:9451–9457.

Fu YX. 1995. Statistical properties of segregating sites. Theoretical population biology. 48:172–197.

Garrison E, Kronenberg ZN, Dawson ET, Pedersen BS, Prins P. 2021. Vcflib and tools for processing the VCF variant call format. bioRxiv. Unpublished. doi: 10.1101/2021.05.21.445151.

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. ArXiv e-prints. Unpublished.. .

Garud NR, Messer PW, Buzbas EO, Petrov DA. 2015. Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. PLoS genetics. 11:e1005004.

Gauthier J, Meier J, Legeai F, McClure M, Whibley A, Bretaudeau A, Boulain H, Parrinello H, Mugford ST, Durbin R et al. 2022. First chromosome scale genomes of ithomiine butterflies (Nymphalidae: Ithomiini): comparative models for mimicry genetic studies. Molecular Ecology Resources. 23:872–885.

Gilbert C, Schaack S, Pace II JK, Brindley PJ, Feschotte C. 2010. A role for host–parasite interactions in the horizontal transfer of transposons across phyla. Nature. 464:1347–1350.

Girgis HZ. 2015. Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. BMC bioinformatics. 16:1–19.

Goel M, Sun H, Jiao WB, Schneeberger K. 2019. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome biology. 20:1–13.

Gonzalez de la Rosa PM, Thomson M, Trivedi U, Tracey A, Tandonnet S, Blaxter M. 2021. A telomere-to-telomere assembly of *Oscheius tipulae* and the evolution of rhabditid nematode chromosomes. G3. 11:jkaa020.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MHY et al. 2010. A draft sequence of the Neandertal genome. science. 328:710–722.

Gremme G, Steinbiss S, Kurtz S. 2013. Genometools: A comprehensive software library for efficient processing of structured genome annotations. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 10:645–656.

Griffiths RC, Marjoram P. 1997. An ancestral recombination graph. Institute for Mathematics and its Applications. 87:257.

Grize SA, Wilwert E, Searle JB, Lindholm AK. 2019. Measurements of hybrid fertility and a test of mate preference for two house mouse races with massive chromosomal divergence. BMC evolutionary biology. 19:1–15.

Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. 2020. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics. 36:2896–2898.

Guerrero RF, Kirkpatrick M. 2014. Local adaptation and the evolution of chromosome fusions. Evolution. 68:2747–2756.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS genetics. 5:e1000695.

Haenel Q, Laurentino TG, Roesti M, Berner D. 2018. Meta-analysis of chromosome-scale crossover rate variation in eukaryotes and its significance to evolutionary genomics. Molecular ecology. 27:2477–2497.

Harewood L, Fraser P. 2014. The impact of chromosomal rearrangements on regulation of gene expression. Human Molecular Genetics. 23:R76–R82.

Harris AM, DeGiorgio M. 2020. A likelihood approach for uncovering selective sweep signatures from haplotype data. Molecular biology and evolution. 37:3023–3046.

Hedrick PW. 1981. The establishment of chromosomal variants. Evolution. pp. 322–332.

Hejase HA, Mo Z, Campagna L, Siepel A. 2022. A deep-learning approach for inference of selective sweeps from the ancestral recombination graph. Molecular Biology and Evolution.

39:msab332.

Hill J, Rastas P, Hornett EA, Neethiraj R, Clark N, Morehouse N, de la Paz Celorio-Mancera M, Cols JC, Dircksen H, Meslin C et al. 2019. Unprecedented reorganization of holocentric chromosomes provides insights into the enigma of lepidopteran chromosome evolution. Science advances. 5:eaau3648.

Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. Genetics Research. 8:269–294.

Hillier LW, Miller RD, Baird SE, Chinwalla A, Fulton LA, Koboldt DC, Waterston RH. 2007. Comparison of *C. elegans* and *C. briggsae* genome sequences reveals extensive conservation of chromosome organization and synteny. PLoS biology. 5:e167.

Hipp AL, Rothrock PE, Roalson EH. 2009. The evolution of chromosome arrangements in *Carex* (Cyperaceae). Botanical Review. 75:96–109.

Hoff K, Lange S, Lomsadze A, Borodovsky M, Stanke M. 2015. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. Bioinformatics. 32:767–769.

Hoff K, Lomsadze A, Borodovsky M, Stanke M. 2019. Gene prediction: Methods and protocols.

Hora KH, Marec F, Roessingh P, Menken SB. 2019. Limited intrinsic postzygotic reproductive isolation despite chromosomal rearrangements between closely related sympatric species of small ermine moths (Lepidoptera: Yponomeutidae). Biological Journal of the Linnean Society. 128:44–58.

Hu F, Zhou L, Tang J. 2013. Reconstructing ancestral genomic orders using binary encoding and probabilistic models. In: Bioinformatics research and applications: 9th international symposium, ISBRA 2013, Charlotte, NC, USA, may 20-22, 2013. proceedings 9. pp. 17–27. Springer.

Hu J, Wang Z, Sun Z, Hu B, Ayoola AO, Liang F, Li J, Sandoval JR, Cooper DN, Ye K et al. 2023. An efficient error correction and accurate assembly tool for noisy long reads. bioRxiv. pp. 2023–03. Unpublished. doi: 10.1101/2023.03.09.531669.

Hubisz MJ, Williams AL, Siepel A. 2020. Mapping gene flow between ancient hominins through demography-aware inference of the ancestral recombination graph. PLoS genetics. 16:e1008895.

Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, Smit AF, Wheeler TJ. 2015. The Dfam database of repetitive DNA families. Nucleic Acids Research. 44:D81–D89.

Hudson RR. 1983a. Properties of a neutral allele model with intragenic recombination. Theoretical population biology. 23:183–201.

Hudson RR. 1983b. Properties of a neutral allele model with intragenic recombination. Theoretical population biology. 23:183–201.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics. 18:337–338.

Hudson RR, Kaplan NL. 1988. The coalescent process in models with selection and recombination. Genetics. 120:831–840.

Irwin DE. 2018. Sex chromosomes and speciation in birds and other ZW systems. Molecular Ecology. 27:3831–3851.

Ivancevic AM, Kortschak RD, Bertozzi T, Adelson DL. 2018. Horizontal transfer of bovb and l1 retrotransposons in eukaryotes. Genome Biology. 19:85.

Jay P, Chouteau M, Whibley A, Bastide H, Parrinello H, Llaurens V, Joron M. 2021. Mutation load at a mimicry supergene sheds new light on the evolution of inversion polymorphisms. Nature genetics. 53:288–293.

Jay P, Whibley A, Frézal L, de Cara MÁR, Nowell RW, Mallet J, Dasmahapatra KK, Joron M. 2018. Supergene evolution triggered by the introgression of a chromosomal inversion. Current Biology. 28:1839–1845.

Jensen JD, Kim Y, DuMont VB, Aquadro CF, Bustamante CD. 2005. Distinguishing between selective sweeps and demography using dna polymorphism data. Genetics. 170:1401–1410.

Jiao WB, Schneeberger K. 2020. Chromosome-level assemblies of multiple *Arabidopsis* genomes reveal hotspots of rearrangements with altered evolutionary dynamics. Nature Communications. 11:989.

John B, Hewitt G. 1970. Inter-population sex chromosome polymorphism in the grasshopper *Podisma pedestris*: I. fundamental facts. Chromosoma. 31:291–308.

Johri P, Charlesworth B, Howell EK, Lynch M, Jensen JD. 2021. Revisiting the notion of deleterious sweeps. Genetics. 219:iyab094.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase update, a database of eukaryotic repetitive elements. Cytogenetic and Genome Research. 110:462–467.

Jühling F, Pütz J, Bernt M, Donath A, Middendorf M, Florentz C, Stadler PF. 2011. Improved systematic tRNA gene annotation allows new insights into the evolution of mitochondrial tRNA structures and into the mechanisms of mitochondrial genome rearrangements. Nucleic Acids Research. 40:2833–2845.

Kamm J, Terhorst J, Durbin R, Song YS. 2020. Efficiently inferring the demographic history of many populations with allele count data. Journal of the American Statistical Association.

115:1472–1487.

Kanzaki N, Tsai IJ, Tanaka R, Hunt VL, Liu D, Tsuyama K, Maeda Y, Namai S, Kumagai R, Tracey A et al. 2018. Biology and genome of a newly discovered sibling species of *Caenorhabditis elegans*. Nature communications. 9:3216.

Kawakami T, Butlin RK, Cooper SJ. 2011. Chromosomal speciation revisited: modes of diversification in Australian morabine grasshoppers (*Vandiemenella*, *viatica* species group). Insects. 2:49–61.

Keightley PD, Pinharanda A, Ness RW, Simpson F, Dasmahapatra KK, Mallet J, Davey JW, Jiggins CD. 2015. Estimation of the spontaneous mutation rate in *Heliconius melpomene*. Molecular biology and evolution. 32:239–243.

Kelleher J, Wong Y, Wohns AW, Fadil C, Albers PK, McVean G. 2019. Inferring whole-genome histories in large population datasets. Nature genetics. 51:1330–1338.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology. 37:907–915.

Kim J, Farré M, Auvil L, Capitanu B, Larkin DM, Ma J, Lewin HA. 2017. Reconstruction and evolutionary history of eutherian chromosomes. Proceedings of the National Academy of Sciences. 114:E5379–E5388.

Kim KW, De-Kayne R, Gordon IJ, Omufwoko KS, Martins DJ, Ffrench-Constant R, Martin SH. 2022. Stepwise evolution of a butterfly supergene via duplication and inversion. Philosophical Transactions of the Royal Society B. 377:20210207.

Kimura M. 1957. Some problems of stochastic processes in genetics. The Annals of Mathematical Statistics. pp. 882–901.

Kingman JF. 1982. On the genealogy of large populations. Journal of applied probability. 19:27–43.

Kirkpatrick M, Barton N. 2006. Chromosome Inversions, Local Adaptation and Speciation. Genetics. 173:419–434.

Kitahara H. 2008. Interspecific hybrid between *Brenthis daphne* and *B. ino* (Lepidoptera, Nymphalidae). The Lepidopterological Society of Japan. 59:144–148.

Kitahara H. 2012. Artificial interspecific and natural hybrids between sympatric *Brenthis daphne* and *B. ino* (Lepidoptera, Nymphalidae) in Nagano Prefecture, Japan. The Lepidopterological Society of Japan. 63:142–150.

Knief U, Hemmrich-Stanisak G, Wittig M, Franke A, Griffith SC, Kempenaers B, Forstmeier W. 2016. Fitness consequences of polymorphic inversions in the zebra finch genome. Genome biology. 17:1–22.

Koch EL, Morales HE, Larsson J, Westram AM, Faria R, Lemmon AR, Lemmon EM, Johannesson K, Butlin RK. 2021. Genetic variation for adaptive traits is associated with polymorphic inversions in *Littorina saxatilis*. Evolution letters. 5:196–213.

Kokot M, Długosz M, Deorowicz S. 2017. KMC 3: counting and manipulating k-mer statistics. Bioinformatics. 33:2759–2761.

Küpper C, Stocks M, Risse JE, Dos Remedios N, Farrell LL, McRae SB, Morgan TC, Karlionova N, Pinchuk P, Verkuil YI et al. 2016. A supergene determines highly divergent male reproductive morphs in the ruff. Nature genetics. 48:79–83.

Laetsch D, Blaxter M. 2017. Blobtools: Interrogation of genome assemblies. F1000Research. 6.

Laetsch DR, Bisschop G, Martin SH, Aeschbacher S, Setter D, Lohse K. 2022. Demographically explicit scans for barriers to gene flow using gIMble. bioRxiv. Unpublished. doi: 10.1101/2022.10.27.514110.

Lande R. 1979. Effective deme sizes during long-term evolution estimated from rates of chromosomal rearrangement. Evolution. pp. 234–251.

Lapierre M, Lambert A, Achaz G. 2017. Accuracy of demographic inferences from the site frequency spectrum: the case of the Yoruba population. Genetics. 206:439–449.

Lauterbur ME, Munch K, Enard D. 2022. Versatile detection of diverse selective sweeps with flex-sweep. bioRxiv. pp. 2022–11. Unpublished. doi: 10.1101/2022.11.15.516494.

Leaché AD, Banbury BL, Linkem CW, de Oca ANM. 2016. Phylogenomics of a rapid radiation: is chromosomal evolution linked to increased diversification in north american spiny lizards (Genus *Sceloporus*)? BMC evolutionary biology. 16:1–16.

Lee KZ, Eizinger A, Nandakumar R, Schuster SC, Sommer RJ. 2003. Limited microsynteny between the genomes of *Pristionchus pacificus* and *Caenorhabditis elegans*. Nucleic acids research. 31:2553–2560.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 34:3094–3100.

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. Nature. 475:493–496.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25:2078–2079.

Li Y, Liu H, Steenwyk JL, LaBella AL, Harrison MC, Groenewald M, Zhou X, Shen XX, Zhao

T, Hittinger CT et al. 2022. Contrasting modes of macro and microsynteny evolution in a eukaryotic subphylum. Current Biology. 32:5335–5343.

Liao WW, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, Lu S, Lucas JK, Monlong J, Abel HJ et al. 2023. A draft human pangenome reference. Nature. 617:312–324.

Liben-Nowell D. 2001. On the structure of syntenic distance. Journal of Computational Biology. 8:53–67.

Liu X, Fu YX. 2020. Stairway Plot 2: demographic history inference with folded SNP frequency spectra. Genome biology. 21:1–9.

Liu Z, Roesti M, Marques D, Hiltbrunner M, Saladin V, Peichel CL. 2022. Chromosomal fusions facilitate adaptation to divergent environments in threespine stickleback. Molecular biology and evolution. 39:msab358.

Lohse K, Chmelik M, Martin SH, Barton NH. 2016. Efficient Strategies for Calculating Blockwise Likelihoods Under the Coalescent. Genetics. 202:775–786.

Lohse K, Frantz LA. 2014. Neandertal admixture in Eurasia confirmed by maximum-likelihood analysis of three genomes. Genetics. 196:1241–1251.

Lohse K, Harrison RJ, Barton NH. 2011. A general method for calculating likelihoods under the coalescent process. Genetics. 189:977–987.

Lohse K, Laetsch D, Vila R, Darwin Tree of Life Consortium et al. 2021. The genome sequence of the large tortoiseshell, *Nymphalis polychloros* (linnaeus, 1758). Wellcome Open Research. 6:238.

Lohse K, Setter D, Darwin Tree of Life Consortium et al. 2022a. The genome sequence of the small pearl-bordered fritillary butterfly, *Boloria selene* (schiffermüller, 1775). Wellcome Open Research. 7:76.

Lohse K, Vila R, Hayward A, Laetsch DR, Wahlberg N, Darwin Tree of Life Consortium et al. 2022b. The genome sequence of the high brown fritillary, *Fabriciana adippe* (Dennis & Schiffermüller, 1775). Wellcome Open Research. 7:298.

Lomsadze A, Burns PD, Borodovsky M. 2014. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Research. 42:e119–e119.

Lucek K, Augustijnen H, Escudero M. 2022. A holocentric twist to chromosomal speciation? Trends in Ecology & Evolution. 37:655–662.

Lukhtanov VA. 2015. The blue butterfly *Polyommatus (Plebicula) atlanticus* (Lepidoptera, Lycaenidae) holds the record of the highest number of chromosomes in the non-polyploid eukaryotic organisms. Comparative Cytogenetics. 9:683–690.

Lukhtanov VA, Dantchenko AV, Khakimov FR, Sharafutdinov D, Pazhenkova EA. 2020a. Karyotype evolution and flexible (conventional versus inverted) meiosis in insects with holocentric chromosomes: a case study based on *Polyommatus* butterflies. Biological Journal of the Linnean Society. 130:683–699.

Lukhtanov VA, Dincă V, Friberg M, Šíchová J, Olofsson M, Vila R, Marec F, Wiklund C. 2018. Versatility of multivalent orientation, inverted meiosis, and rescued fitness in holocentric chromosomal hybrids. Proceedings of the National Academy of Sciences. 115:E9610–E9619.

Lukhtanov VA, Dincă V, Friberg M, Vila R, Wiklund C. 2020b. Incomplete sterility of chromosomal hybrids: implications for karyotype evolution and homoploid hybrid speciation. Frontiers in genetics. 11:583827.

Lukhtanov VA, Dincă V, Talavera G, Vila R. 2011. Unprecedented within-species chromosome number cline in the Wood White butterfly *Leptidea sinapis* and its significance for karyotype evolution and speciation. BMC evolutionary biology. 11:1–11.

Lukhtanov VA, Kandul NP, Plotkin JB, Dantchenko AV, Haig D, Pierce NE. 2005. Reinforcement of pre-zygotic isolation and karyotype evolution in Agrodiaetus butterflies. Nature. 436:385–389.

Lundberg M, Mackintosh A, Petri A, Bensch S. 2023. Inversions maintain differences between migratory phenotypes of a songbird. Nature Communications. 14:452.

Luo J, Sun X, Cormack BP, Boeke JD. 2018. Karyotype engineering by chromosome fusion leads to reproductive isolation in yeast. Nature. 560:392–396.

Lysak MA. 2022. Celebrating Mendel, McClintock, and Darlington: On end-to-end chromosome fusions and nested chromosome fusions. The Plant Cell. 34:2475–2491.

Ma J. 2010. A probabilistic framework for inferring ancestral genomic orders. In: 2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 179–184. IEEE.

Mackintosh A, Laetsch DR, Baril T, Foster RG, Dincă V, Vila R, Hayward A, Lohse K. 2022. The genome sequence of the lesser marbled fritillary, *Brenthis ino*, and evidence for a segregating neo-Z chromosome. G3. 12:jkac069.

Mackintosh A, Laetsch DR, Hayward A, Charlesworth B, Waterfall M, Vila R, Lohse K. 2019. The determinants of genetic diversity in butterflies. Nature Communications. 10:3466.

Mackintosh A, Vila R, Laetsch DR, Hayward A, Martin SH, Lohse K. 2023. Chromosome fissions and fusions act as barriers to gene flow between *Brenthis* fritillary butterflies. Molecular Biology and Evolution. 40:msad043.

Maeda T. 1939. Chiasma studies in the silk worm, *Bombyx mori* l. Jpn. J. Genet.. 15:118–127.

Maeki K, Makino S. 1953. Chromosome numbers of some Japanese Rhopalocera. The Lepidopterists' News. 7:36–38.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Molecular Biology and Evolution. 38:4647–4654.

Márquez-Corro JI, Martín-Bravo S, Jiménez-Mejías P, Hipp AL, Spalink D, Naczi RF, Roalson EH, Luceño M, Escudero M. 2021. Macroevolutionary insights into sedges (*Carex*: Cyperaceae): The effects of rapid chromosome number evolution on lineage diversification. Journal of Systematics and Evolution. 59:776–790.

Martin SH, Davey JW, Salazar C, Jiggins CD. 2019. Recombination rate variation shapes barriers to introgression across butterfly genomes. PLoS biology. 17:e2006288.

Maruyama T. 1970. Effective number of alleles in a subdivided population. Theoretical population biology. 1:273–306.

Maruyama T, Kimura M. 1974. A note on the speed of gene frequency changes in reverse directions in a finite population. Evolution. pp. 161–163.

Maynard Smith J, Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genetics Research. 23:23–35.

McVean GA, Cardin NJ. 2005. Approximating the coalescent with recombination. Philosophical Transactions of the Royal Society B: Biological Sciences. 360:1387–1393.

Melters DP, Paliulis LV, Korf IF, Chan SW. 2012. Holocentric chromosomes: convergent evolution, meiotic adaptations, and genomic analysis. Chromosome Research. 20:579–593.

Mercer S, Wallace B, Searle J. 1992. Male common shrews (*Sorex araneus*) with long meiotic chain configurations can be fertile: implications for chromosomal models of speciation. Cytogenetic and Genome Research. 60:68–73.

Mérot C, Berdan EL, Cayuela H, Djambazian H, Ferchaud AL, Laporte M, Normandeau E, Ragoussis J, Wellenreuther M, Bernatchez L. 2021. Locally adaptive inversions modulate genetic variation at different geographic scales in a seaweed fly. Molecular Biology and Evolution. 38:3953–3971.

Messer PW. 2013. SLiM: simulating evolution with selection and linkage. Genetics. 194:1037–1039.

Messer PW, Petrov DA. 2013. Population genomics of rapid adaptation by soft selective sweeps. Trends in ecology & evolution. 28:659–669.

Meyermans R, Gorssen W, Buys N, Janssens S. 2020. How to study runs of homozygosity using PLINK? A guide for analyzing medium density SNP data in livestock and pet species. BMC genomics. 21:1–14.

Miklós I, Smith H. 2015. Sampling and counting genome rearrangement scenarios. BMC bioinformatics. 16:1–14.

Miklós I, Tannier E. 2010. Bayesian sampling of genomic rearrangement scenarios via double cut and join. Bioinformatics. 26:3012–3019.

Mongue AJ, Hansen ME, Walters JR. 2021. Support for faster and more adaptive Z chromosome evolution in two divergent lepidopteran lineages. Evolution. 76:332–345.

Moran PAP. 1958. Random processes in genetics. In: Mathematical proceedings of the cambridge philosophical society. volume 54. pp. 60–71. Cambridge University Press.

Moshe A, Wygoda E, Ecker N, Loewenthal G, Avram O, Israeli O, Hazkani-Covo E, Pe'er I, Pupko T. 2022. An approximate bayesian computation approach for modeling genome rearrangements. Molecular Biology and Evolution. 39:msac231.

Muffato M, Louis A, Nguyen NTT, Lucas J, Berthelot C, Roest Crollius H. 2023. Reconstruction of hundreds of reference ancestral genomes across the eukaryotic kingdom. Nature Ecology & Evolution. 7:355–366.

Muller HJ. 1940. Bearing of the *Drosophila* work on systematics. The new systematics,. pp. 185–268.

Murat F, Zhang R, Guizard S, Gavranović H, Flores R, Steinbach D, Quesneville H, Tannier E, Salse J. 2015. Karyotype and gene order evolution from reconstructed extinct ancestors highlight contrasts in genome plasticity of modern rosid crops. Genome biology and evolution. 7:735–749.

Nagylaki T. 1982. Geographical invariance in population genetics. Journal of Theoretical Biology. 99:159–172.

Narain Y, Fredga K. 1997. Meiosis and fertility in common shrews, *Sorex araneus* from a chromosomal hybrid zone in central Sweden. Cytogenetic and Genome Research. 78:253–259.

Näsvall K, Boman J, Höök L, Vila R, Wiklund C, Backström N. 2023. Nascent evolution of recombination rate differences as a consequence of chromosomal rearrangements. PLoS genetics. 19:e1010717.

Navarro A, Barton NH. 2003. Accumulating postzygotic isolation genes in parapatry: A new twist on chromosomal speciation. Evolution. 57:447–459.

Nemetschke L, Eberhardt AG, Viney ME, Streit A. 2010. A genetic map of the animal-parasitic nematode *Strongyloides ratti*. Molecular and Biochemical Parasitology. 169:124–127.

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. 2005. Genomic scans for selective sweeps using SNP data. Genome research. 15:1566–1575.

Noor MA, Grams KL, Bertucci LA, Reiland J. 2001. Chromosomal inversions and the reproductive isolation of species. Proceedings of the National Academy of Sciences. 98:12084–12088.

Nunes A, Catalan J, Lopez J, Ramalhinho M, Mathias M, Britton-Davidian J. 2011. Fertility assessment in hybrids between monobrachially homologous Rb races of the house mouse from the island of Madeira: implications for modes of chromosomal evolution. Heredity. 106:348–356.

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Research. 27:824–834.

Orteu A, Kucka M, Katili E, Ngumbao C, Gordon IJ, Ng'iru I, Talavera G, Warren IA, Collins S, ffrench Constant RH et al. 2023. Transposable element insertions are associated with batesian mimicry in the pantropical butterfly *Hypolimnas misippus*. bioRxiv. pp. 2023–07.

Ostevik KL, Samuk K, Rieseberg LH. 2020. Ancestral reconstruction of karyotypes reveals an exceptional rate of nonrandom chromosomal evolution in sunflower. Genetics. 214:1031–1045.

Ottenburghs J. 2022. Avian introgression patterns are consistent with Haldane's rule. Journal of Heredity. 113:363–370.

Otto SP, Whitton J. 2000. Polyploid incidence and evolution. Annual review of genetics. 34:401–437.

Ou S, Jiang N. 2019. Ltr_finder_parallel: parallelization of ltr_finder enabling rapid identification of long terminal repeat retrotransposons. Mobile DNA. 10:48.

Palahí i Torres A, Höök L, Näsvall K, Shipilina D, Wiklund C, Vila R, Pruisscher P, Backström N. 2022. The fine-scale recombination rate variation and associations with genomic features in a butterfly. bioRxiv. pp. 2022–11. Unpublished. doi: 10.1101/2022.11.02.514807.

Pardo-Manuel de Villena F, Sapienza C. 2001a. Female meiosis drives karyotypic evolution in mammals. Genetics. 159:1179–1189.

Pardo-Manuel de Villena F, Sapienza C. 2001b. Nonrandom segregation during meiosis: the unfairness of females. Mammalian Genome. 12:331–339.

Passarge E, Horsthemke B, Farber RA. 1999. Incorrect use of the term synteny. Nature genetics. 23:387–387.

Pazhenkova EA, Lukhtanov VA. 2019. Nuclear genes (but not mitochondrial DNA barcodes) reveal real species: Evidence from the *Brenthis* fritillary butterflies (Lepidoptera, Nymphalidae). Journal of Zoological Systematics and Evolutionary Research. 57:298–313.

Pazhenkova EA, Lukhtanov VA. 2023. Chromosomal conservatism vs chromosomal megaevolution: enigma of karyotypic evolution in lepidoptera. Chromosome Research. 31:16.

Pedersen BS, Quinlan AR. 2017. Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics. 34:867–868.

Pennell MW, Kirkpatrick M, Otto SP, Vamosi JC, Peichel CL, Valenzuela N, Kitano J. 2015. Y fuse? sex chromosome fusions in fishes and reptiles. PLOS Genetics. 11:1–17.

Perrin A, Varré JS, Blanquart S, Ouangraoua A. 2015. ProCARs: Progressive reconstruction of ancestral gene orders. BMC genomics. 16:1–11.

Platt, Roy N. I, Blanco-Berdugo L, Ray DA. 2016. Accurate Transposable Element Annotation Is Vital When Analyzing New Genome Assemblies. Genome Biology and Evolution. 8:403–410.

Poszewiecka B, Gogolewski K, Stankiewicz P, Gambin A. 2022. Revised time estimation of the ancestral human chromosome 2 fusion. BMC genomics. 23:1–16.

Potter S, Bragg JG, Blom MP, Deakin JE, Kirkpatrick M, Eldridge MD, Moritz C. 2017. Chromosomal speciation in the genomics era: disentangling phylogenetic evolution of rock-wallabies. Frontiers in Genetics. 8:10.

Potter S, Bragg JG, Turakulov R, Eldridge MD, Deakin J, Kirkpatrick M, Edwards RJ, Moritz C. 2022. Limited introgression between rock-wallabies with extensive chromosomal rearrangements. Molecular Biology and Evolution. 39:msab333.

Przeworski M. 2002. The signature of positive selection at randomly chosen loci. Genetics. 160:1179–1189.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American journal of human genetics. 81:559–575.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26:841–842.

Qumsiyeh MB, Handal EN. 2022. Adaptive nature of chromosome variation in placental mammals and applicability to domestication and invasiveness. Hystrix, the Italian Journal of Mammalogy. 33:102–106.

R Core Team. 2021. R: A language and environment for statistical computing.

Ranallo-Benavidez TR, Jaron KS, Schatz MC. 2020. Genomescope 2.0 and smudgeplot for reference-free profiling of polyploid genomes. Nature Communications. 11:1432.

Ranz JM, Maurin D, Chan YS, Von Grotthuss M, Hillier LW, Roote J, Ashburner M, Bergman CM. 2007. Principles of genome evolution in the *Drosophila melanogaster* species group.

PLoS biology. 5:e152.

Rasmussen MD, Hubisz MJ, Gronau I, Siepel A. 2014. Genome-wide inference of ancestral recombination graphs. PLoS genetics. 10:e1004342.

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 28:i333–i339.

Renwick JH. 1971. The mapping of human chromosomes. Annual review of genetics. 5:81–120.

Rhie A, Walenz BP, Koren S, Phillippy AM. 2020. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biology. 21:245.

Rieseberg LH. 2001. Chromosomal rearrangements and speciation. Trends in Ecology & Evolution. 16:351–358.

Robertson WRB. 1916. Chromosome studies. I. Taxonomic relationships shown in the chromosomes of Tettigidae and Acrididae: V-shaped chromosomes and their significance in Acrididae, Locustidae, and Gryllidae: Chromosomes and variation. Journal of Morphology. 27:179.

Robinson JT, Turner D, Durand NC, Thorvaldsdóttir H, Mesirov JP, Aiden EL. 2018. Juicebox.js provides a cloud-based visualization system for Hi-C data. Cell Systems. 6:256–258.e1.

Robinson R. 1971. *Lepidoptera Genetics: International Series of Monographs in Pure and Applied Biology: Zoology*. volume 46. Elsevier.

Rödelsperger C, Meyer JM, Prabh N, Lanz C, Bemm F, Sommer RJ. 2017. Single-molecule sequencing reveals the chromosome-scale genomic architecture of the nematode model organism *Pristionchus pacificus*. Cell Reports. 21:834–844.

Rosser N, Edelman NB, Queste LM, Nelson M, Seixas F, Dasmahapatra KK, Mallet J. 2022. Complex basis of hybrid female sterility and Haldane's rule in *Heliconius* butterflies: Z-linkage and epistasis. Molecular Ecology. 31:959–977.

Rozas J, Aguade M. 1994. Gene conversion is involved in the transfer of genetic information between naturally occurring inversions of *Drosophila*. Proceedings of the National Academy of Sciences. 91:11517–11521.

RStudio Team. 2020. Rstudio: Integrated development environment for R.

Rubino F, Creevey C. 2014. Mgkit: Metagenomic framework for the study of microbial communities. Figshare. Poste.

Ruckman SN, Jonika MM, Casola C, Blackmon H. 2020. Chromosome number evolves at equal rates in holocentric and monocentric clades. PLoS genetics. 16:e1009076.

Saitoh K. 1986. On the haploid chromosome number of *Brenthis daphne iwatensis* (Lepidoptera Nymphalidae) from Hama-koshimizu, Hokkaido. Tyo To Ga. 37:101–102.

Saitoh K. 1987. A note on the haploid chromosome number of *Brenthis ino* (Rottemburg, 1775) from Finland (Lepidoptera, Nymphalidae). Nota lepidopterologica. 10:131–132.

Saitoh K. 1991. Chromosome number of *Brenthis ino* (Rottemburg, 1775) from Sweden (Lepidoptera, Nymphalidae). Nota lepidopterologica. 14:241–243.

Saitoh K, Abe A, Kumagai Y, Hiroshi O. 1989. Chromosomes of the fritillaries of the genus *Brenthis* (Lepidoptera, Nymphalidae) from Japan II. A chromosome survey in males of *Brenthis ino mashuensis* (Kono, 1931). Tyo To Ga. 40:253–257.

Saitoh K, Lukhtanov V. 1988. Some chromosomal aspects of *Brenthis hecate* [Denis & Schiffermüller], 1775 from South Altai, USSR (Lepidoptera, Nymphalidae). Nota Lepidopterologica. 11:234–236.

Sankoff D. 2003. Rearrangements and chromosomal evolution. Current opinion in genetics & development. 13:583–587.

Schultz DT, Haddock SH, Bredeson JV, Green RE, Simakov O, Rokhsar DS. 2023. Ancient gene linkages support ctenophores as sister to other animals. Nature. pp. 1–8.

Searle JB. 1991. A hybrid zone comprising staggered chromosomal clines in the house mouse (*Mus musculus domesticus*). Proceedings: Biological Sciences. 246:47–52.

Setter D. 2023. Breakups and Hookups: a Markov model for karyotype evolution. bioRxiv. pp. 2023–08. Unpublished. doi: 10.1101/2023.08.15.553394.

Setter D, Mousset S, Cheng X, Nielsen R, DeGiorgio M, Hermisson J. 2020. VolcanoFinder: genomic scans for adaptive introgression. PLoS Genetics. 16:e1008867.

Shi T, Zhang X, Hou Y, Jiang Y, Jia C, Lai Q, Dan X, Feng J, Feng J, Ma T et al. 2023. The super-pangenome of *Populus* unveil genomic facets for adaptation and diversification in widespread forest trees. bioRxiv. pp. 2023–07.

Shipilina D, Näsvall K, Höök L, Vila R, Talavera G, Backström N. 2022. Linkage mapping and genome annotation give novel insights into gene family expansions and regional recombination rate variation in the painted lady (*Vanessa cardui*) butterfly. Genomics. 114:110481.

Simakov O, Bredeson J, Berkoff K, Marletaz F, Mitros T, Schultz DT, O'Connell BL, Dear P, Martinez DE, Steele RE et al. 2022. Deeply conserved synteny and the evolution of metazoan chromosomes. Science advances. 8:eabi5884.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 31:3210–3212.

Slatkin M. 1981. Fixation probabilities and fixation times in a subdivided population. Evolution.

pp. 477–488.

Smit A, Hubley R, Green P. 2015. Repeatmasker open-4.0. http://www.repeatmasker.org.

Song B, Marco-Sola S, Moreto M, Johnson L, Buckler ES, Stitzer MC. 2022. Anchorwave: Sensitive alignment of genomes with high sequence diversity, extensive structural polymorphism, and whole-genome duplication. Proceedings of the National Academy of Sciences. 119:e2113075119.

Speidel L, Forest M, Shi S, Myers SR. 2019. A method for genome-wide genealogy estimation for thousands of samples. Nature genetics. 51:1321–1329.

Spirito F, Rizzoni M, Rossi C. 1993. The establishment of underdominant chromosomal rearrangements in multi-deme systems with local extinction and colonization. Theoretical population biology. 44:80–94.

Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 24:637–644.

Stanke M, Schöffmann O, Morgenstern B, Waack S. 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics. 7:62.

Stern AJ, Wilton PR, Nielsen R. 2019. An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. PLoS Genetics. 15:e1008384.

Stevens L, Moya ND, Tanny RE, Gibson SB, Tracey A, Na H, Chitrakar R, Dekker J, Walhout AJ, Baugh LR et al. 2022. Chromosome-level reference genomes for two strains of *Caenorhabditis briggsae*: an improved platform for comparative genomics. Genome biology and evolution. 14:evac042.

Stevens L, Rooke S, Falzon LC, Machuka EM, Momanyi K, Murungi MK, Njoroge SM, Odinga CO, Ogendo A, Ogola J et al. 2020. The genome of *Caenorhabditis bovis*. Current Biology. 30:1023–1031.

Stewart NB, Ahmed-Braimah YH, Cerne DG, McAllister BF. 2019. Female meiotic drive preferentially segregates derived metacentric chromosomes in *Drosophila*. bioRxiv. p. 638684. Unpublished. doi: 10.1101/638684.

Streicher JW, of Life WSIT, Darwin Tree of Life Consortium et al. 2021. The genome sequence of the common toad, *Bufo bufo* (Linnaeus, 1758). Wellcome Open Research. 6.

Strobeck C. 1987. Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. Genetics. 117:149–153.

Sturtevant AH. 1921. A case of rearrangement of genes in *Drosophila*. Proceedings of the

National Academy of Sciences of the United States of America. 7:235–237.

Sukumaran J, Holder MT. 2010. Dendropy: a Python library for phylogenetic computing. Bioinformatics. 26:1569–1571.

Sultana T, Zamborlini A, Cristofari G, Lesage P. 2017. Integration site selection by retroviruses and transposable elements in eukaryotes. Nature Reviews Genetics. 18:292–308.

Sun S, Shinya R, Dayi M, Yoshida A, Sternberg PW, Kikuchi T. 2020. Telomere-to-telomere genome assembly of *Bursaphelenchus okinawaensis* strain SH1. Microbiology Resource Announcements. 9:10–1128.

Sutton WS. 1903. The chromosomes in heredity. The Biological Bulletin. 4:231–250.

Tajima F. 1983. Evolutionary relationship of dna sequences in finite populations. Genetics. 105:437–460.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 123:585–595.

Talavera G, Lukhtanov VA, Rieppel L, Pierce NE, Vila R. 2013. In the shadow of phylogenetic uncertainty: The recent diversification of *Lysandra* butterflies through chromosomal change. Molecular Phylogenetics and Evolution. 69:469–478.

Tandonnet S, Koutsovoulos GD, Adams S, Cloarec D, Parihar M, Blaxter ML, Pires-daSilva A. 2019. Chromosome-wide evolution and sex determination in the three-sexed nematode *Auanema rhodensis*. G3: Genes, Genomes, Genetics. 9:1211–1230.

Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. 2015. Sambamba: fast processing of NGS alignment formats. Bioinformatics. 31:2032–2034.

Templeton AR. 1981. Mechanisms of speciation - a population genetic approach. Annual Review of Ecology and Systematics. 12:23–48.

Tesler G. 2002. Grimm: genome rearrangements web server. Bioinformatics. 18:492–493.

Teterina AA, Willis JH, Phillips PC. 2020. Chromosome-level assembly of the *Caenorhabditis remanei* genome reveals conserved patterns of nematode genome organization. Genetics. 214:769–780.

Tracey A, Foster JM, Paulini M, Grote A, Mattick J, Tsai YC, Chung M, Cotton JA, Clark TA, Geber A et al. 2020. Nearly complete genome sequence of *Brugia malayi* strain FR3. Microbiology Resource Announcements. 9:10–1128.

Turelli M, Orr HA. 1995. The dominance theory of Haldane's rule. Genetics. 140:389–402.

Turner J, Sheppard P. 1975. Absence of crossing-over in female butterflies (*Heliconius*). Heredity. 34:265–269.

Twyford AD, Friedman J. 2015. Adaptive divergence in the monkey flower *Mimulus guttatus* is maintained by a chromosomal inversion. Evolution. 69:1476–1486.

Vila R, Hayward A, Lohse K, Wright C, Darwin Tree of Life Consortium et al. 2021. The genome sequence of the Glanville fritillary, *Melitaea cinxia* (Linnaeus, 1758) [version 1; peer review: 1 approved]. Wellcome Open Research. 6.

Wahrman J, Goitein R, Nevo E. 1969. Mole rat *Spalax*: evolutionary significance of chromosome variation. Science. 164:82–84.

Wakeley J. 1999. Nonequilibrium migration in human history. Genetics. 153:1863–1871.

Wallau GL, Ortiz MF, Loreto ELS. 2012. Horizontal Transposon Transfer in Eukarya: Detection, Bias, and Perspectives. Genome Biology and Evolution. 4:801–811.

Walsh JB. 1982. Rate of accumulation of reproductive isolation by chromosome rearrangements. The American Naturalist. 120:510–532.

Wang J, Veronezi GM, Kang Y, Zagoskin M, O'Toole ET, Davis RE. 2020. Comprehensive chromosome end remodeling during programmed dna elimination. Current Biology. 30:3397–3413.

Wellband K, Mérot C, Linnansaari T, Elliott J, Curry RA, Bernatchez L. 2019. Chromosomal fusion and life history-associated genomic variation contribute to within-river local adaptation of atlantic salmon. Molecular Ecology. 28:1439–1459.

Wellenreuther M, Bernatchez L. 2018. Eco-evolutionary genomics of chromosomal inversions. Trends in Ecology & Evolution. 33:427–440.

White M. 1978a. Chain processes in chromosomal speciation. Systematic Zoology. 27:285–298.

White M, Carson H, Cheney J. 1964. Chromosomal races in the Australian grasshopper *Moraba viatica* in a zone of geographic overlap. Evolution. pp. 417–429.

White MJD. 1968. Models of speciation. Science. 159:1065–1070.

White MJD. 1973. *Animal cytology & evolution*. Cambridge university press.

White MJD. 1978b. *Modes of speciation*. W. H. Freeman San Francisco.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J et al. 2019. Welcome to the tidyverse. Journal of Open Source Software. 4:1686.

Wilson A, Bush G, Case S, King M. 1975. Social structuring of mammalian populations and rate of chromosomal evolution. Proceedings of the National Academy of Sciences. 72:5061–5065.

Wong WY, Simakov O. 2018. RepeatCraft: a meta-pipeline for repetitive element defragmentation and annotation. Bioinformatics. 35:1051–1052.

Woolsey CI. 1915. Linkage of chromosomes correlated with reduction in numbers among the species of a genus, also within a species of the Locustidae. The Biological Bulletin. 28:163–186.

Wright CJ, Stevens L, Mackintosh A, Lawniczak M, Blaxter M. 2023. Chromosome evolution in lepidoptera. bioRxiv. pp. 2023–05. Unpublished. doi: 10.1101/2023.05.12.540473.

Wright S. 1941. On the probability of fixation of reciprocal translocations. The American Naturalist. 75:513–522.

Xiong T, Li X, Yago M, Mallet J. 2022. Admixture of evolutionary rates across a butterfly hybrid zone. Elife. 11:e78135.

Xu Z, Wang H. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Research. 35:W265–W268.

Yamaguchi O, Mukai T. 1974. Variation of spontaneous occurrence rates of chromosomal aberrations in the second chromosomes of *Drosophila melanogaster*. Genetics. 78:1209–1221.

Yannic G, Basset P, Hausser J. 2009. Chromosomal rearrangements and gene flow over time in an inter-specific hybrid zone of the *Sorex araneus* group. Heredity. 102:616–625.

Yeaman S. 2013. Genomic rearrangements and the evolution of clusters of locally adaptive loci. Proceedings of the National Academy of Sciences. 110:E1743–E1751.

Yoshida K, Kitano J. 2021. Tempo and mode in karyotype evolution revealed by a probabilistic model incorporating both chromosome number and morphology. PLoS genetics. 17:e1009502.

Yoshida K, Rödelsperger C, Röseler W, Riebesell M, Sun S, Kikuchi T, Sommer RJ. 2023. Chromosome fusions repatterned recombination rate and facilitated reproductive isolation during *Pristionchus* nematode speciation. Nature Ecology & Evolution. pp. 1–16.

Zhao T, Schranz ME. 2019. Network-based microsynteny analysis identifies major differences and genomic outliers in mammalian and angiosperm genomes. Proceedings of the National Academy of Sciences. 116:2165–2174.

Zheng C, Sankoff D. 2011. On the pathgroups approach to rapid small phylogeny. BMC bioinformatics. 12:1–9.

Zhou C, McCarthy SA, Durbin R. 2023. YaHS: yet another Hi-C scaffolding tool. Bioinformatics. 39:btac808.

Zima J, Fedyk S, Fredga K, Hausser J, Mishta A, Searle JB, Volobouev VT, Wójcik JM. 1996. The list of the chromosome races of the common shrew (*Sorex araneus*). Hereditas. 125:97–107.

Supplementary Materials for Chapter 2

A.1 Supplementary figures

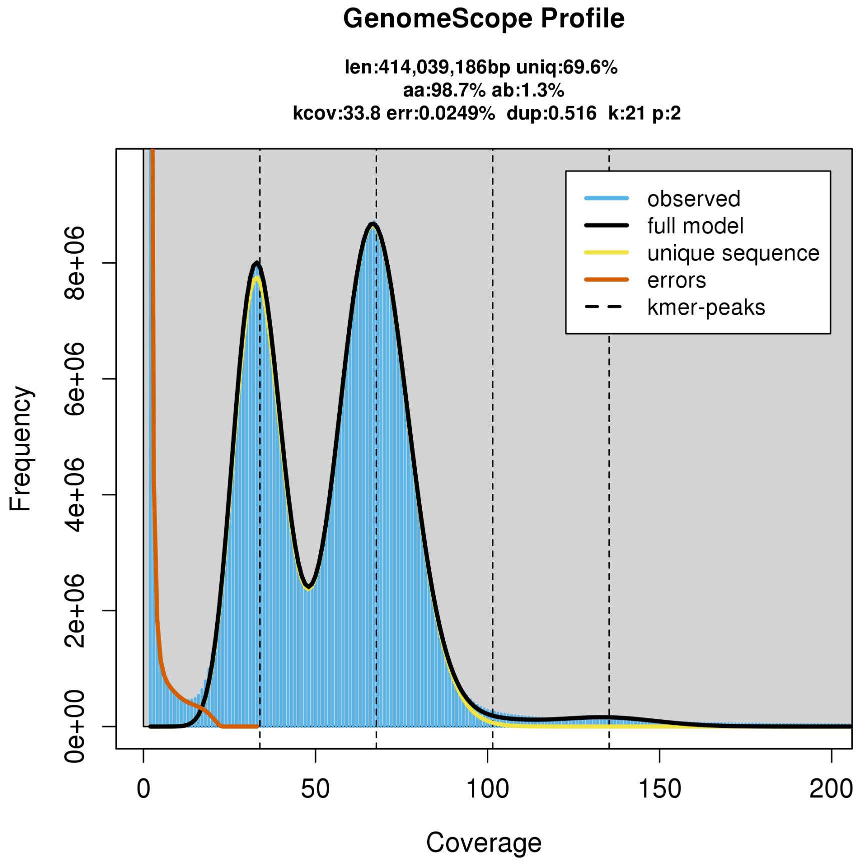

Figure A.1: Kmer spectrum and Genomescope parameter estimates for SO_BI_364. Coverage (x-axis) corresponds to the number of times a kmer is observed in the reads and frequency (y-axis) is the number of kmers with that coverage.

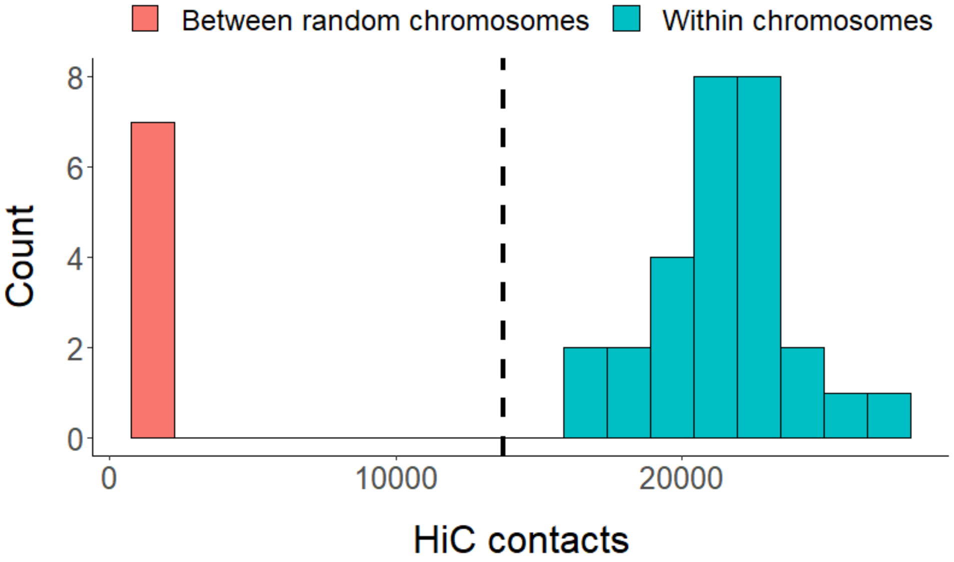

Figure A.2: A histogram of HiC contact frequency within and between chromosomes. In red, the number of contacts spanning fusion points of randomly fused chromosome pairs, where either read is within 5Mb of the fusion point and each chromosome has been randomly fused once. In blue, the number of contacts spanning arbitrary points within chromosomes, where reads are again within 5Mb and two independent points are sampled per chromosome. The dashed line represents the number of contacts spanning the putative neo-Z fusion between chromosomes 11 and 13. The frequency of contacts supporting the neo-Z is consistent with heterozygosity.

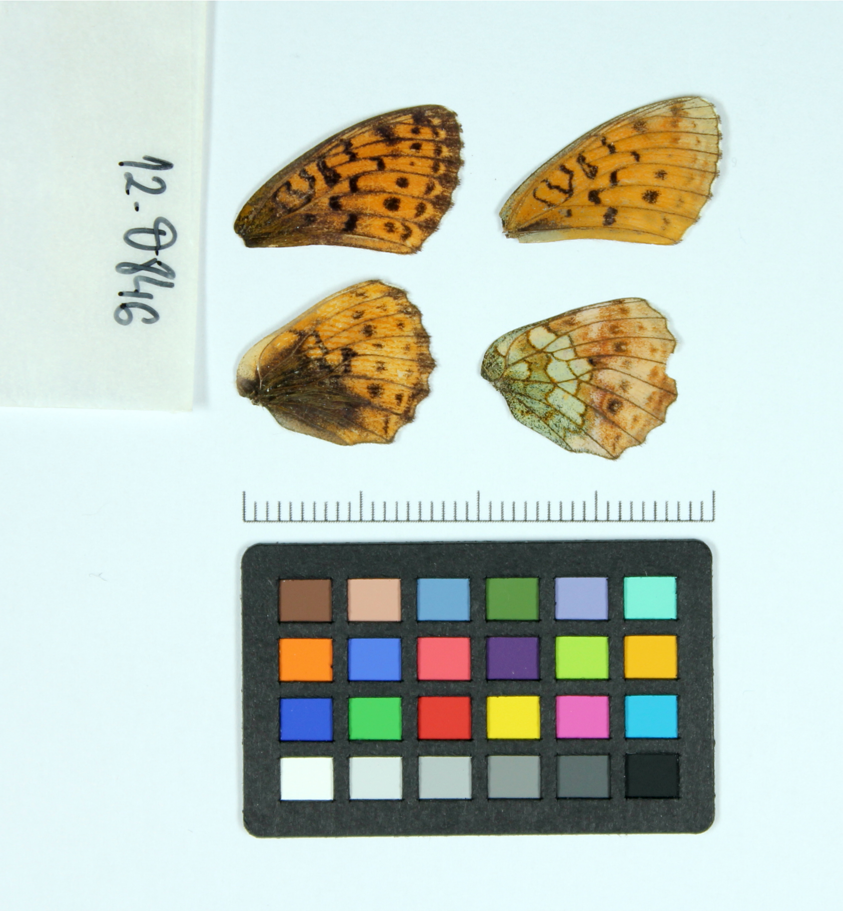

Figure A.3: Wings of the female specimen FR_BI_1497. Top-left: dorsal forewing. Top-right: ventral forewing. Bottom-left: dorsal hindwing. Bottom-right: ventral hindwing

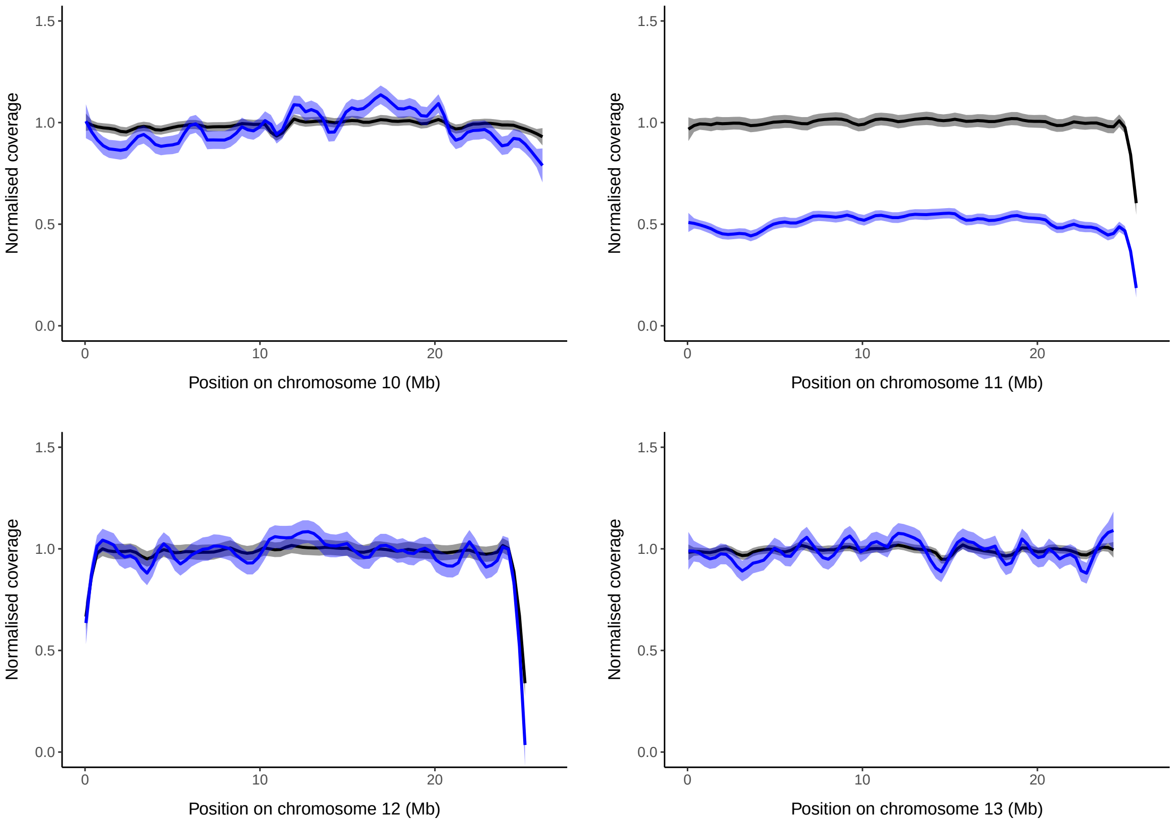

Figure A.4: Normalised coverage on four chromosomes. The black line represents normalised coverage of WGS reads from the male used for genome assembly (SO_BI_364), while the blue line is normalised coverage of WGS reads from a female individual (FR_BI_1497). On chromosome 11 the male has full coverage whereas the female has half coverage, consistent with expectations for Z-linked chromosomes in Lepidoptera. On the other three chromosomes, as well as the ten not shown, both individuals have full normalised coverage.

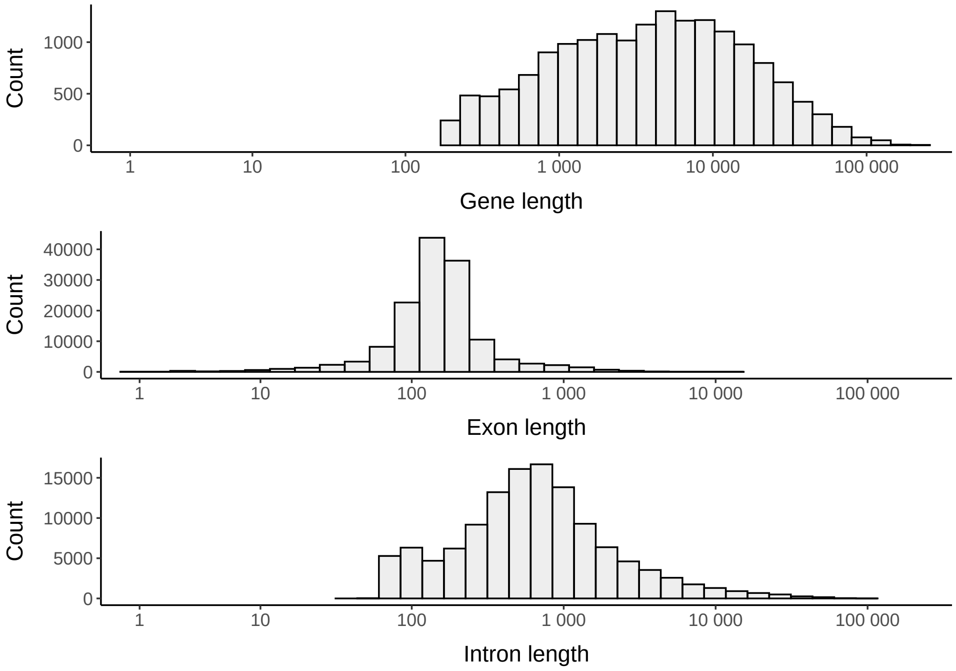

Figure A.5: The distribution of gene, exon, and intron lengths, plotted on a log_{10} scale.

A.2 Supplementary tables

Table A.1: Sampling locations and other metadata for *B. ino* samples used in this study.

Sample	Date	Sex	Locality	Region	Country	Lat	Long	Altitude	Collector	Data	Analyses
SO_BI_364	05/07/2017	Male	Somiedo, Braña de Mumian	Asturias	Spain	43.068	-6.24	1420	KL	Pacbio CLR + Illumina WGS	Contig assembly + Z identification
SO_BI_375	05/07/2017	Male	Somiedo, Braña de Mumian	Asturias	Spain	43.068	-6.24	1420	KL	Illumina HiC + Illumina WGS	Scaffolding + Karyotyping
SO_BI_376	05/07/2017	n/a	Somiedo, Braña de Mumian	Asturias	Spain	43.068	-6.24	1420	KL	Illumina RNA-seq	Gene annotation
FR_BI_1497 (RV-coll12O846)	11/08/2012	Female	Larche (Les Marmottes)	Alpes-de-Haute-Provence	France	44.446	6.851	1680	VD & Raluca Vodă	Illumina WGS	Z identification

Table A.2: Annotated transposable elements

Repeat class	No. elements	Total length (Mb)	Percentage of genome (%)	No. distinct classifications
Retroelement	116,256	47.36	11.49	930
SINE	26,242	6.53	1.58	23
LINE	59,472	25.92	6.29	611
Penelope	25,145	6.89	1.67	34
LTR element	5,397	8.02	1.95	262
DNA transposon	36,793	11.48	2.79	598
Rolling Circle	258,724	73.34	17.81	307
Unclassified	74,790	23.74	5.76	339
Other	21	0.00	0.00	3
Total	486,584	155.92	37.85	2,177

Supplementary Materials for Chapter 3

B.1 Supplementary methods

The heuristic algorithm of Ferretti et al. (1996) finds the syntenic distance between two genomes using fission, fusion and translocation rearrangements. Their algorithm allows non-reciprocal translocations, which are unidentifiable from a fission and subsequent fusion. In this work we only consider reciprocal translocations, as they are more likely to be identifiable from fissions and fusions (Figure 3.5). We therefore modified the algorithm of Ferretti et al. (1996) to only include reciprocal translocations. Here we describe the algorithm through a simple example.

Consider two genomes, G_A and G_B, that consist of chromosome containing orthologous markers (lower case letters). We can write these genomes as

G_A: [a b c d e f] [g h i] [j k] [l m n o p q r]

G_B: [a b c q r] [g h i j k] [d e f l m] [n o p]

Using the compact representation of Ferretti et al. (1996), we can instead write the chromosomes of G_A in terms of the G_B chromosomes that they share markers with.

G_A: [B_1 B_3] [B_2] [B_2] [B_3 B_4 B_1] [a b c d e f] [g h i] [j k] [l m n o p q r]

To transform G_A into G_B, we first identify labels (highlighted in bold below) that are only present once and implement a fission to generate a new chromosome containing that label.

[B_1 B_3] [B_2] [B_2] [B_3 $\mathbf{B_4}$ B_1] [a b c d e f] [g h i] [j k] [l m n o p q r]

↓ fission ↓

[B_1 B_3] [B_2] [B_2] [B_3 B_1] [$\mathbf{B_4}$] [a b c d e f] [g h i] [j k] [l m q r] [n o p]

If there are no more labels present only once, we choose a label that is present twice. If the chromosomes that share that label also share at least one other label, then a translocation is implemented.

[$\mathbf{B_1}$ B_3] [B_2] [B_2] [B_3 $\mathbf{B_1}$] [B_4] [a b c d e f] [g h i] [j k] [l m q r] [n o p]

↓ translocation ↓

[$\mathbf{B_1}$] [B_2] [B_2] [B_3] [B_4] [a b c q r] [g h i] [j k] [d e f l m] [n o p]

If we have chosen another label that is present twice, but the chromosomes that share that label do not share any other labels, then a fusion is implemented.

[B_1] [$\mathbf{B_2}$] [$\mathbf{B_2}$] [B_3] [B_4] [a b c q r] [g h i] [j k] [d e f l m] [n o p]

↓ fusion ↓

[B_1] [$\mathbf{B_2}$] [B_3] [B_4] [a b c q r] [g h i j k] [d e f l m] [n o p]

We have now recovered G_B through rearrangement and so there are no more steps. The syntenic distance is three, and the putative rearrangements history involves one fission, one translocation and one fusion.

In the above example we did not encounter any instances where there were labels shared by > 2 chromosomes. In this case, a fusion is implemented involving the two chromosomes with the greatest intersection of labels.

B.2 Supplementary tables

Table B.1: Nematode genomes analysed in this study. The first 14 taxa listed were originally analysed by Gonzalez de la Rosa et al. (2021), whereas the *Pristionchus exspectatus* genome was not. In some cases more recent assemblies were used than in Gonzalez de la Rosa et al. (2021).

Taxon	Chr.	NCBI accession	Study
Ascaris suum	24	GCA_013433145.1	(Wang et al. 2020)
Auanema rhodensis	7	GCA_947366455.1	(Tandonnet et al. 2019)
Brugia malayi	5	GCA_000002995.5	(Tracey et al. 2020)
Bursaphelenchus okinawaensis	6	GCA_904067145.1	(Sun et al. 2020)
Caenorhabditis briggsae	6	GCA_021491975.1	(Stevens et al. 2022)
Caenorhabditis elegans	6	GCA_028201515.1	(Bush et al. 2023)
Caenorhabditis inopinata	6	GCA_003052745.1	(Kanzaki et al. 2018)
Caenorhabditis nigoni	6	GCA_027920645.1	NA
Caenorhabditis remanei	6	GCA_010183535.1	(Teterina et al. 2020)
Haemonchus contortus	6	GCA_000469685.2	(Doyle et al. 2020)
Onchocerca volvulus	4	GCA_000499405.2	(Cotton et al. 2016)
Oscheius tipulae	6	GCA_013425905.1	(Gonzalez de la Rosa et al. 2021)
Pristionchus pacificus	6	GCA_000180635.4	(Rödelsperger et al. 2017)
Strongyloides ratti	3	GCA_001040885.1	(Nemetschke et al. 2010)
Pristionchus exspectatus	6	GCA_911812115.1	(Yoshida et al. 2023)

Supplementary Materials for Chapter 4

C.1 Supplementary note 1

The ability to accurately infer effective demographic parameters from the bSFS depends on a number of variables. A single block only contains information about a single genealogy, so many blocks are required to make accurate inference and statistical power depends on the amount of recombination between blocks. However, recombination within blocks can introduce bias, because the analytic calculation for the bSFS assumes no recombination within blocks. So recombination involves a trade-off between power and potential for bias in parameter and model estimates.

With this in mind, we used gIMble to investigate how recombination affects our ability to estimate demographic parameters in windows across the genome. We simulated windows of equivalent size as those in the real data under the best fitting genome-wide demographic model (Figure 4.2C). Each simulation contained 30,030 blocks of 64 bases, equivalent to a window of length 45.76 kb split up into 715 blocks. The bSFS was calculated by recording the mutation counts for all 42 possible pairwise comparisons (6 × 7 unphased diploid individuals from *B. ino* and *B. daphne* respectively).

We performed 100 replicate simulations for six different per-base recombination rates (r): 1×10^{-9}, 5×10^{-9}, 9×10^{-9}, 1.3×10^{-8}, 1.7×10^{-8}, 2.1×10^{-8}. We estimated demographic parameters (N_e and m_e) for each replicate, while fixing the split time (T), analogous to the the window-wise analysis on the real data. However, we used free optimisation rather than a grid, because the former provides finer parameter estimates. Importantly, it is possible to identify simulation replicates which have not converged to their maximum composite lnCL (MCL), because the parameters they were simulated under are known. This task is much more challenging with the *Brenthis* data (because the parameters are unknown) and so requires the use of a grid.

Comparing results under different recombination rates (Figure C.3), we find that there is little power to accurately estimate m_e, the parameter we are most interested in, when the recombination rate is 1×10^{-9}. As recombination increases, estimates of m_e become closer to the true value (1.811×10^{-7}). However, m_e is often underestimated at higher recombination rates (Figure C.3). For example, the mean estimate of m_e when $r = 2.1 \times 10^{-8}$ is 1.195×10^{-7}.

This power analysis on simulated data shows that given plausible recombination rates, our analyses of m_e in windows (of 30,000 consecutive blocks) have reasonable power, even though estimates suffer from some downward bias. Although we lack estimates of recombination rate in *B. daphne* and *B. ino*, we expect the mean crossover rate to be approximately 8.5×10^{-9} (equivalent to a single crossover per male meiosis for 14 chromosome pairs). In addition, windows in the *Brenthis* dataset often span much greater distances than 45.76 kb (median window span 122 kb) because genic and repetitive regions are removed. This increases the amount of between-block recombination and therefore power.

C.2 Supplementary note 2

Throughout our analyses we have assumed that the arrangement of chromosomes found in each genome assembly is representative of each species, i.e. rearrangements are fixed between species. Although all chromosome rearrangements we have identified are homozygous in the reference assemblies, it is still possible that a small number of rearrangements are polymorphic within species. A potential consequence of polymorphic rearrangements is that they act as barriers to gene flow within a species. We tested for this possibility by repeating our demographic analysis at the intraspecific level.

We fit genome-wide demographic models to estimate the divergence history of populations that currently occupy different glacial refugia. We inferred that *B. ino* from Iberian and Balkan populations ($F_{\mathrm{ST}} = 0.118$) split approximately 459 kya without post-divergence gene flow (Table C.2). By contrast, *B. daphne* Iberian and Balkan populations ($F_{\mathrm{ST}} = 0.112$) likely split more recently (331 kya) and with considerable post-divergence gene flow ($m_e = 1.072 \times 10^{-5}$) from Iberian to the Balkan populations forwards in time (Table C.2).

We then estimated variation in m_e between Iberian and Balkan populations of *B. daphne* across the genome. The distribution of m_e estimates on non-rearranged chromosomes, rearranged chromosomes, and within 1 Mb of rearrangement points, are all very similar (Figure C.4). There are no statistically significant differences between their means (permutation tests, see Methods). This is in stark contrast to the interspecific results (Figure 4.4), thus demonstrating that these rearrangements are barriers to gene flow between species but not between refugial populations within *B. daphne*.

C.3 Supplementary figures

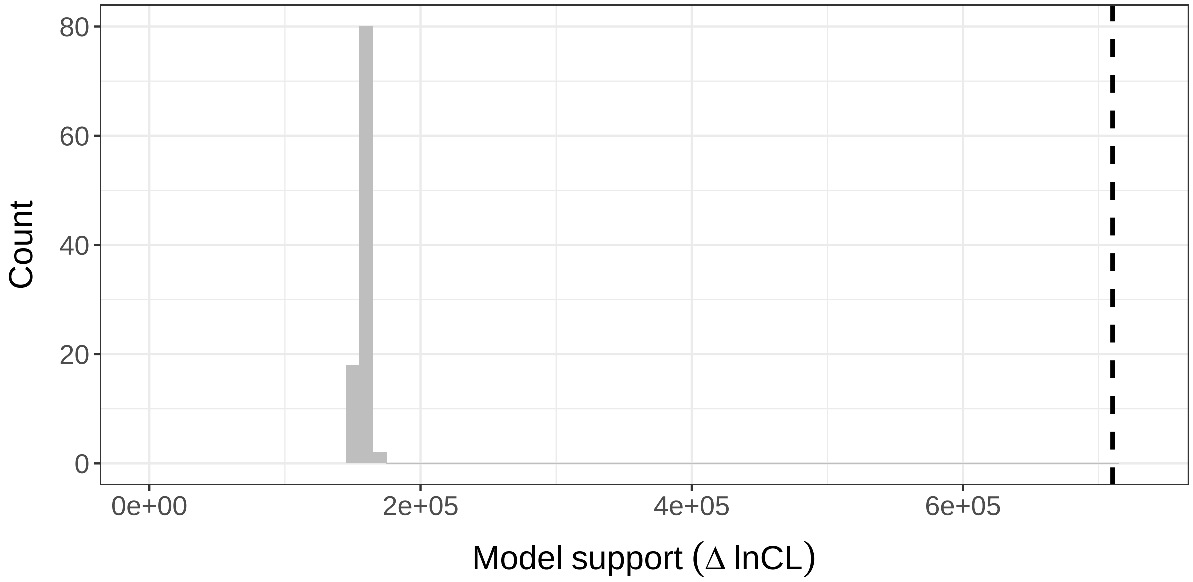

Figure C.1: A histogram showing the improvement in model support (Δ lnCL) between the *DIV* and $IM_{\rightarrow Bda}$ models for 100 parametric bootstrap replicates, each simulated under the same *DIV* history. The improvement in fit for the real data is marked with a dashed vertical line.

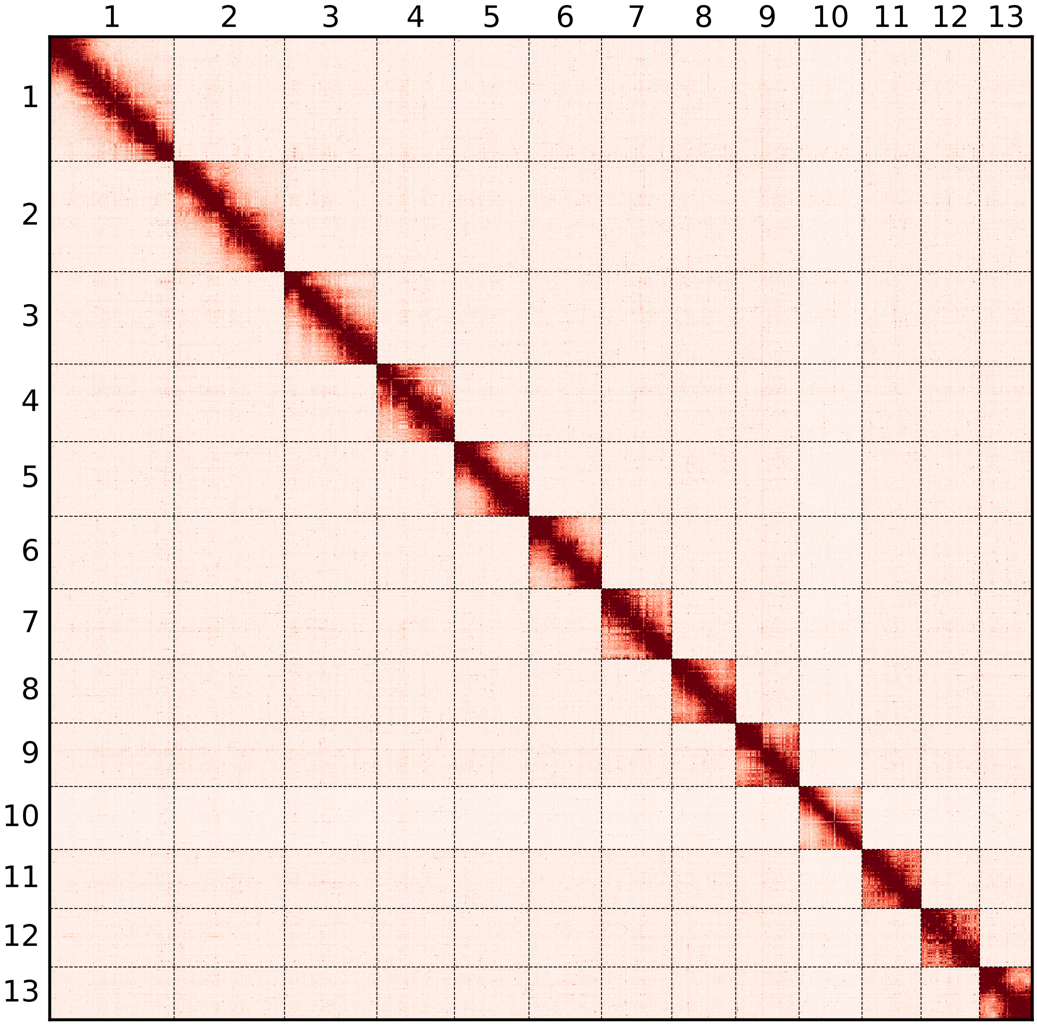

Figure C.2: A HiC contact heatmap showing the 13 *Brenthis daphne* chromosomes.

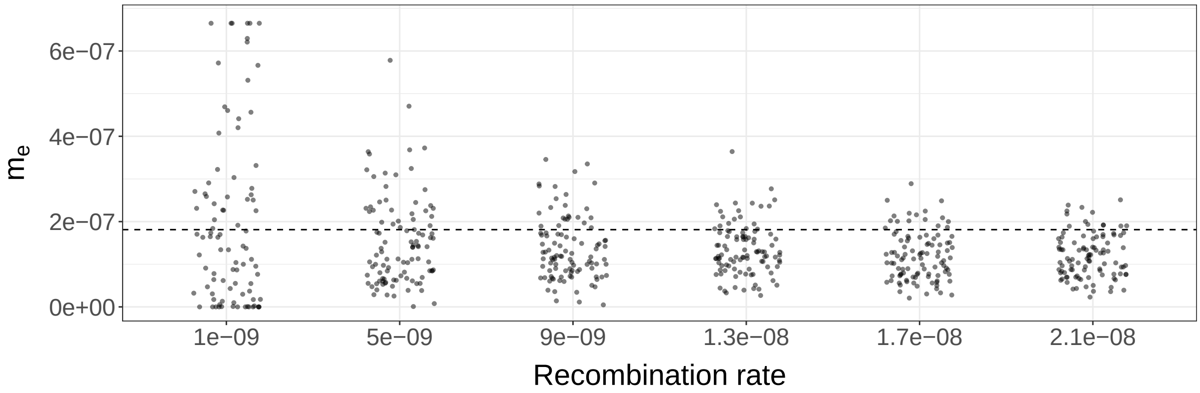

Figure C.3: Estimates of effective migration rate (m_e) from simulations with different recombination rates. Simulation replicates are plotted as jittered points around the recombination rate that they were simulated under. The simulated m_e (1.811×10^{-7}) is plotted as a dashed horizontal line.

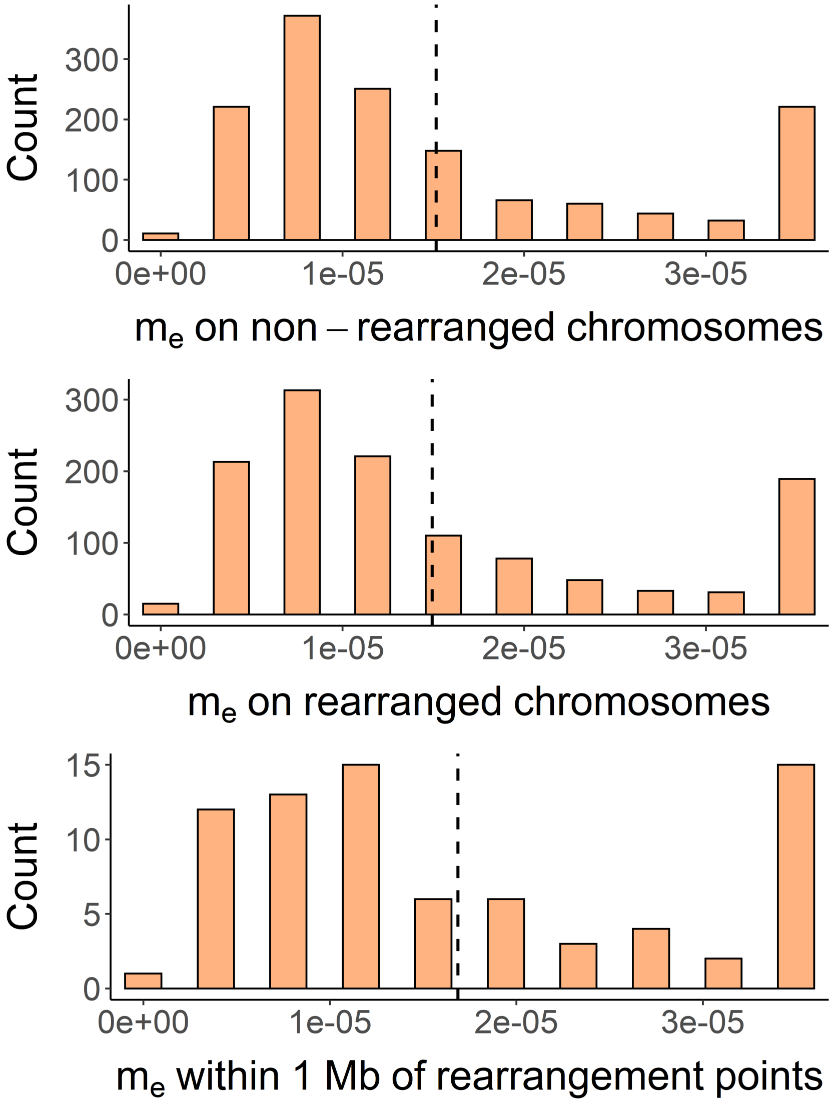

Figure C.4: The distribution of m_e estimates between Iberian and Balkan *B. daphne* populations for genomic windows from non-rearranged and rearranged chromosomes, as well as within 1 Mb of rearrangement points. Vertical dashed lines represent the mean of each distribution. The high frequency of windows with a maximum m_e value included in the grid reflects the long tail of the m_e distribution.

C.4 Supplementary tables

Table C.1: Sampling locations and other metadata for individuals used in this study. In the data column, the source of the data is denoted as TS (This Study) or M2022 (Mackintosh et al. 2022).

Sample	Date	Species	Sex	Locality	Region	Country	Lat	Long	Collector	Data
ES_BD_1141	25/7/2018	daphne	Female	Meranges, La Cerdanya	Catalunya	Spain	42.435	1.797	RV, Sabina Vila	Pacbio, WGS; TS
ES_BD_1489	29/7/2009	daphne	Female	Prioro	Castile and León	Spain	42.937	-4.964	RV	WGS; TS
ES_BD_1490	24/7/2008	daphne	Female	Uña	Castile-La Mancha	Spain	40.231	-1.960	RV	WGS; TS
FR_BD_1329	15/7/2019	daphne	Female	D620	Aude	France	42.997	2.053	KL	WGS, HiC; TS
GR_BD_1491	26/7/2013	daphne	Female	Rhodopi, Frakto Forest	Drama	Greece	41.504	24.400	RV	WGS; TS
IT_BD_1493	26/6/2012	daphne	Male	Saguccio, Aspromonte	Aspromonte	Italy	38.080	15.830	RV	WGS; TS
RO_BD_956	17/7/2018	daphne	Male	Pin1000m, Lupsa, Apuseni Mt.	Alba	Romania	46.416	23.192	KL, AH, DL, RV	WGS; TS
ES_BI_364	05/07/2017	ino	Male	Somiedo, Braña de Mumian	Asturias	Spain	43.068	-6.24	KL	WGS, ref. genome; M2022
ES_BI_375	05/07/2017	ino	Male	Somiedo, Braña de Mumian	Asturias	Spain	43.068	-6.24	KL	WGS; M2022
FR_BI_1497	11/08/2012	ino	Female	Larche (Les Marmottes)	Alpes-de-Haute-Provence	France	44.446	6.851	Vlad Dincă, Raluca Vodă	WGS; M2022
RS_BI_1496	29/6/2014	ino	Male	Čeganica	-	Serbia	43.396	22.368	RV	WGS; TS
SE_BI_1495	13/7/2016	ino	Female	Älvsbyn	Norrbotten	Sweden	65.668	20.955	RV	WGS; TS
UA_BI_1494	20/6/2014	ino	Female	Kruglyanka, Novaya Vodolaga	Kharkiv oblast	Ukraine	49.817	35.733	RV	WGS; TS

Table C.2: Maximum composite likelihood parameters for intraspecific demographic models. Parameter estimates and log composite likelihoods (lnCL) are shown for *IM* models of divergence between Iberian and Balkan populations of *B. ino* and *B. daphne*. The N_e and split time parameter estimates are in units of 10^6 individuals and years, respectively.

Species	**Model**	N_e ***Balkans***	N_e ***Iberia***	N_e ***ancestral***	m_e	**Split time**	**lnCL**
ino	$IM_{\rightarrow Balk}$	1.025	0.700	1.019	5.949×10^{-20}	0.459	-22,993,925
ino	$IM_{\rightarrow Iber}$	1.025	0.700	1.019	0	0.459	-22,993,925
daphne	$IM_{\rightarrow Balk}$	0.101	0.102	1.090	1.072×10^{-5}	0.331	-17,408,294
daphne	$IM_{\rightarrow Iber}$	0.222	0.012	1.239	3.510×10^{-5}	0.646	-17,422,413

Supplementary Materials for Chapter 5

D.1 Supplementary methods

D.1.1 Gene annotation

We annotated genes in three genome assemblies: *F. adippe*, *B. daphne*, and *B. hecate* (note that *B. ino* already has a gene annotation from Mackintosh et al. 2022). This was done so that the SFS-based demographic modelling could be restricted to putatively neutral fourfold-degenerate (4D) sites in the *F. adippe* genome, and that exonic regions in the *Brenthis* genomes could be excluded when fitting sweep models. We masked repeats in the *B. daphne* and *B. hecate* genomes using Red (Girgis 2015) with default parameters. A repeat-masked version of the *F. adippe* assembly was kindly supplied by Tobias Baril (personal communication), having been repeat annotated with EarlGrey v1.2 (Baril et al. 2021, 2022).

RNA-seq data was generated for an *F. adippe* and *B. hecate* individual and kindly shared with us by Sam Ebdon (Table D.1). RNA extractions, library preparations, and sequencing were performed alongside datasets generated for Ebdon et al. (2021). We also accessed the RNA-seq dataset for *B. daphne* from Ebdon et al. (2021). We next mapped species-specific RNA-seq reads to the assemblies with HISAT2 v2.1.0 (Kim et al. 2019). The repeat-masked assemblies and RNA-seq alignments were used as input for gene annotation with braker2.1.5 (Stanke et al. 2006, 2008; Li et al. 2009; Barnett et al. 2011; Lomsadze et al. 2014; Buchfink et al. 2015; Hoff et al. 2015, 2019). We used GenomeTools v1.6.1 (Gremme et al. 2013) to format gff3 and bed files for each annotation. Finally, 4D sites in the *F. adippe* genome assembly were identified with partition_cds.py (see Data accessibility).

D.1.2 Fitting a multi-species demographic model with fastsimcoal2

We fit a single demographic model to the folded 3D_SFS using fastsimcoal2 (version fsc27093). We chose to fit a complex model (Figure 5.2) and then quantify the uncertainty in parameter estimates through parametric bootstraps (Table D.2). To obtain maximum composite likelihood estimates for each parameter we performed the following optimisation command ten times:

```
fsc27093 -t brenthis.tpl -n 1000000 -m -e brenthis.est -M -L 30 -c 50 -B 50
```

This corresponds to parameter optimisation where each likelihood estimate is approximated using 1,000,000 coalescent simulations and parameters are optimised through Brent's algorithm across 30 rounds.

Given the parameter estimates with the greatest composite likelihood (Table D.2), we performed 100 parametric bootstrap simulations with the following command:

```
fsc27093 -i brenthis.par -n100 -c 5 -B 5 -j -m -s0 -x -I -q --multiSFS.
```

Each simulation consists of 619 loci of length 4 kb with mutation and recombination rates of $\mu = r = 2.9 \times 10^{-9}$. This number and length of loci corresponds to the total amount of data used to generate the observed SFS, as well as the approximate level of linkage given that reads only map to genic regions of the *F. adippe* genome. Parameter estimates were then obtained for each simulated SFS using the same optimisation procedure as described above. These estimates were used to estimate 95% CIs (Table D.2).

D.1.3 Strategies for fitting sweep models to the bSFS

We fit hard selective sweep models using the method of Bisschop et al. (2021). For each analysis we used data from 1 Mb of sequence and therefore thousands of short sequence blocks. Each composite likelihood calculation requires the probability of observing the mutation configuration (bSFS entry) of each block given its distance from the sweep centre. Instead of performing these calculations repeatedly, which would be prohibitively slow, we generated a grid with dimensions corresponding to θ, $\alpha * distance$ and T_a, in which, each element contains the exact probabilities of all 64 possible bSFS entries. The probability of a bSFS entry for a particular parameter combination and distance from the sweep centre can then be obtained through linear interpolation between points in the grid. The grid contained 15 θ points between 0.1 and 1.5, 47 $\alpha * distance$ points between 0 and 12.0, 11 T_a points between 0 and 1.0, and therefore 7755 parameter combinations in total. This places a limit on the age of sweeps that can be inferred ($T_a = 1$, i.e. $2N_e$ generations ago). When a sweep is weak, many blocks will be $\alpha * distance > 12$ away from the sweep centre and therefore outside of the grid. However, probabilities at such a high $\alpha * distance$ are effectively the same as under a neutral model, and we approximate the probability as such.

For a given point in the genome and the blockwise data in the surrounding 1 Mb, we optimised the parameters of the sweep model using the Nelder Mead algorithm in *Mathematica*. We repeated the optimisation three times with different random seeds and retained the parameters with the greatest likelihood. We set a minimum $Log_{10}(\alpha)$ value of -5.7. This corresponds to a very strong sweep where $\alpha \times$ 500 kb = 1. Sweeps with smaller α values than this would be unlikely to show a spatial pattern across 1 Mb and so cannot be identified reliably.

We fit two other models to the same data. The first is a neutral model with a single parameter, θ. The second is a model with a central region $2 * d$ bases in size where θ is reduced relative to a background value. Unlike the sweep model, these models do not include any distortion in genealogical branch lengths. Code for fitting all three of these models to bSFS data can be found in the *Mathematica* notebook titled brenthis_sweeps_chromosome_scan.nb (see Data accessibility).

D.1.4 Fitting a finite-island model

We tested whether a finite-island model (Maruyama 1970) could explain levels of overall genetic diversity and ROH in each *Brenthis* species. This model consists of local populations (i.e. demes) of effective size N_e, where the effective migration rate m_e is the per-generation probability that a lineage migrates out of a deme. The N_e of each deme largely determines the chance of very recent common ancestry. By contrast, the longer term rate of coalescence and therefore the overall levels of genetic diversity and divergence are a function of the effective migration rate (m_e) and number of demes.

We fit this finite-island model using three summary statistics: per-site heterozygosity (H), pairwise intraspecific divergence (d_{xy}) between individuals sampled from different demes in Europe, and the proportion of 1 Mb windows covered by a ROH (which we call W_{roh}). These statistics provide information about the rate of coalescence within and between demes (H and d_{xy}) as well as the rate of very recent within-deme coalescence (W_{roh}). For a given species, we averaged these statistics across all individuals/pairwise comparisons. We used the expected time of coalescence for lineages sampled within and between demes (Nagylaki 1982; Strobeck 1987; Wakeley 1999) to calculate the expected H and d_{xy}, respectively. We estimated the probability of observing a 1 Mb window covered by a ROH as the probability that, for two lineages sampled from the same deme, the first event backwards in time is coalescence rather than migration or recombination within the window. These calculations assume equal recombination and mutation rates ($\mu = r = 2.9 \times 10^{-9}$). We then inferred the parameters of the finite-island model (number of demes, N_e and m_e) as those for which the expected H, d_{xy} and W_{roh} match the data. Model fitting was performed in a *Mathematica* notebook (finite_island_model.nb, see Data accessibility).

D.2 Supplementary figures

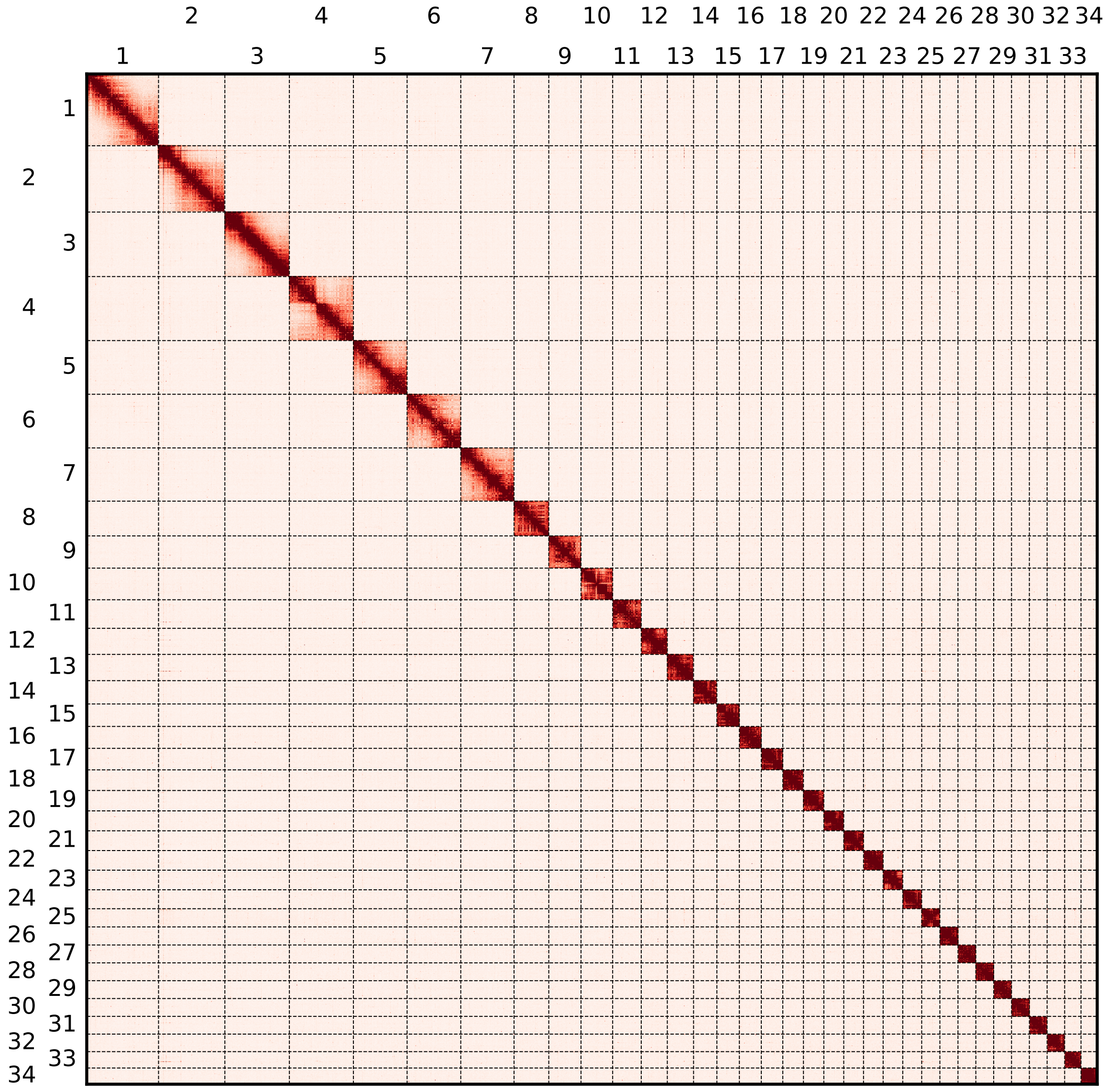

Figure D.1: A HiC contact heatmap showing the 34 *Brenthis hecate* chromosomes.

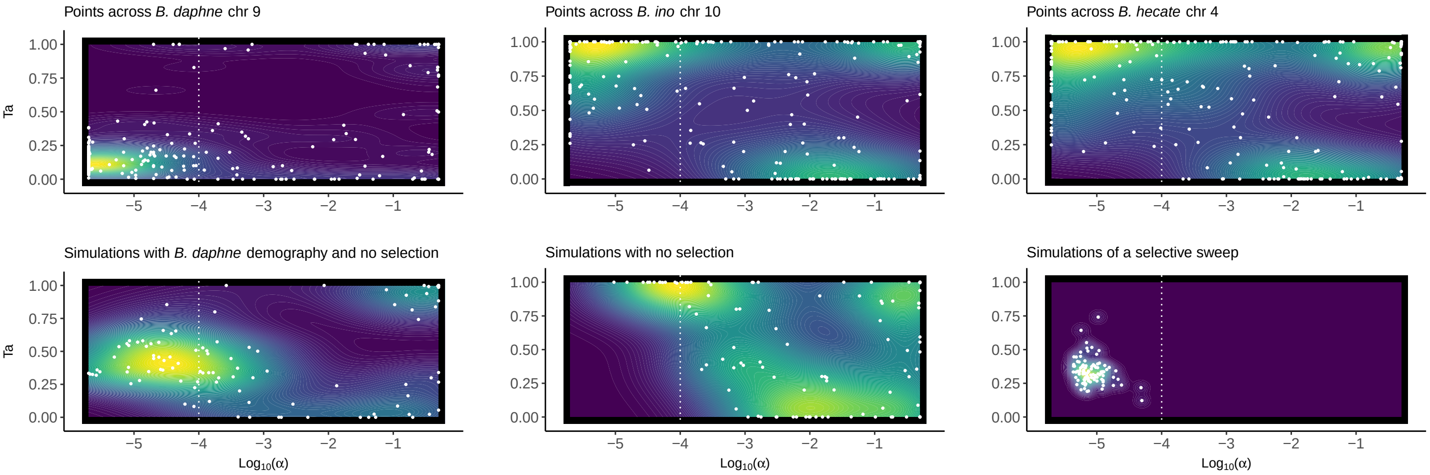

Figure D.2: Parameters of inferred selective sweeps. Plots show the estimated strength of sweeps ($Log_{10}(\alpha)$, x-axis) and their estimated timing (T_a, y-axis). Within a plot, each white point represents parameter estimates for a test site in the genome or a single simulation, whereas coloured contours show the density of these estimates across multiple points/simulations. The top plots show inferred sweep parameters for points sampled across the same (orthologous) chromosome, for *B. daphne*, *B. ino*, and *B. hecate*. The bottom plots show inferred sweep parameters for simulations. Each plot has a vertical dashed line at $Log_{10}(\alpha)$ = -4, as points to the left of this can be considered putative selective sweeps (see Main Text).

D.3 Supplementary tables

Table D.1: Sampling locations and other metadata for the individuals used to generate new sequence data in this study.

Sample	Preservation	Date	Species	Sex	Locality	Region	Country	Lat	Long	Collector	Data
ES_BH_1411	Liquid nitrogen	6/6/2019	*B. hecate*	Male	Segura de la Sierra, Jaén	Andalucia	Spain	38.263	-2.615	RV	RNA-seq, HiC
ES_BH_1412	Liquid nitrogen	10/6/2019	*B. hecate*	Male	Ablanque	Castille-La Mancha	Spain	40.927	-2.189	RV	Pacbio, WGS
IT_BH_1622	Ethanol	14/7/2013	*B. hecate*	Female	Borgo Olivi	Treviso	Italy	46.024	12.280	L. Dapporto, R. Vodă	WGS
IT_BH_1623	Ethanol	22/7/2013	*B. hecate*	Male	Sasso Tetto	Macerata	Italy	43.007	13.232	L. Dapporto	WGS
RS_BH_1628	Ethanol	27/6/2014	*B. hecate*	Male	Divcibare, Mt. Maljen	-	Serbia	44.122	20.015	R. Vodă, V. Dincă	WGS
GR_BH_1631	Ethanol	3/7/2014	*B. hecate*	Female	Granitis	East Macedonia and Thrace	Greece	41.308	23.905	R. Vodă, V. Dincă	WGS
RO_FA_934	Liquid nitrogen	17/7/2018	*F. adippe*	Male	Pin1000m, Lupsa, Apuseni Mt.	Alba	Romania	46.416	23.192	KL, RV, Alex Hayward, Dominik R. Laetsch	RNA-seq

Table D.2: Maximum composite likelihood parameter estimates for a demographic model describing the divergence history of three *Brenthis* species. Values are given to three significant digits. Lower and upper 95% confidence intervals (CIs) were calculated from parametric bootstrap simulations. For some parameters (*) the point estimate fall outside of the 95% CIs. The → of each m_e parameter denotes the direction of migration backwards in time.

Parameter	**Lower 95% CI**	**Point estimate**	**Upper 95% CI**
N_e *daph*	185,000	212,000 *	208,000
N_e *ino*	1,140,000	1,260,000 *	1,230,000
N_e *hec*	1,330,000	1,460,000 *	1,450,000
N_e *daph + ino*	99,500	130,000	859,000
N_e *hec ancestral*	18,800	55,600	123,000
N_e *daph + ino + hec*	1,800,000	2,560,000	4,730,000
Split *daph + ino*	2,360,000	2,790,000	3,000,000
Split *daph + ino + hec*	3,030,000	3,200,000	8,520,000
m_e *daph →ino*	1.58×10^{-7}	1.68×10^{-7}	2.16×10^{-7}
m_e *daph →hec*	9.42×10^{-9}	1.13×10^{-8}	2.43×10^{-8}
m_e *ino →daph*	2.33×10^{-9}	4.48×10^{-9}	1.28×10^{-8}
m_e *ino →hec*	6.89×10^{-9}	6.03×10^{-9} *	1.29×10^{-8}
m_e *hec →daph*	4.91×10^{-9}	5.26×10^{-9}	9.37×10^{-9}
m_e *hec →ino*	1.65×10^{-8}	1.59×10^{-8} *	2.41×10^{-8}
m_e *daph + ino →hec ancestral*	5.43×10^{-9}	4.33×10^{-7}	6.51×10^{-7}
m_e *hec ancestral →daph + ino*	2.39×10^{-8}	3.53×10^{-7}	5.65×10^{-7}

Table D.3: Summary statistics for each species and corresponding parameter estimates under a finite-island model. H is per-4D-site heterozygosity, d_{xy} is pairwise intraspecific 4D site divergence between individuals sampled from different demes in Europe, and W_{roh} is the proportion of 1 Mb windows covered by a ROH. Estimates of the number of demes, N_e and m_e are given to two significant figures.

Species	**H**	d_{xy}	W_{roh}	**Demes**	N_e	m_e
B. daphne	0.0044	0.0048	0.0044	20	19,000	1.3×10^{-4}
B. ino	0.010	0.012	0.022	260	3,400	3.9×10^{-4}
B. hecate	0.0098	0.013	0.013	130	6,500	1.4×10^{-4}

Appendix

Here I provide the methods used to generate Figures 1.1 and 6.1.

I performed coalescent simulations to approximate the distribution of segregating sites for a sample of two lineages (Figure 1.1C). I considered two different demographic histories, the bottleneck history in Figure 1.1A and a history of constant population size (N_e = 894) with equivalent overall diversity. The demographic histories were simulated 100,000 times with msprime v1.0.2 (Baumdicker et al. 2021). Each simulation consisted of a 5 kb sequence with the mutation and recombination rates both set to 1×10^{-7} per-base per-generation. I recorded the exact counts of segregating sites for each simulation and grouped those that had > 8 together (Figure 1.1C).

I reconstructed a marginal genealogy around a chromosome fusion point on *Brenthis daphne* chromosome 2 (Figure 6.1) using ARGweaver v1.0 (Rasmussen et al. 2014). I used the filtered VCF file generated in chapter 5 (used for ROH identification and sweep inference in *B. daphne*) as input. I included all SNPs, regardless of whether genotypes were missing for some individuals. I ran ARGweaver with the following command:

```
arg-sample --vcf brenthis_daphne.vcf.gz
--region brenthis_daphne.ES_BD_1141.chromosome_2:23443769-24443769 --unphased
--mutrate 2.9e-9 --recombrate 2.9e-9 --maxtime 25e6 --popsize 2e5 --verbose 2
--output arg-sample.sweep --compress-seq 3 --resample-window-iters 1 --iters 1000
--resample-window 3000
```

I plotted the marginal genealogy from the 1000th MCMC iteration (Figure 6.1), spanning sites 23,943,621 - 23,943,926 on *B. daphne* chromosome 2, as this is centre of the 33 kb region which contains the fusion point.