

🎓 Welcome to the Research Alumni Portal, Luis Mijangos Araujo!

You will be able to download the finalised version of all thesis submissions that were processed in GRIS here.

Please ensure to include the **completed declaration** (from the Declarations tab), your **completed Inclusion of Publications Statement** (from the Inclusion of Publications Statement tab) in the final version of your thesis that you submit to the Library.

Information on how to submit the final copies of your thesis to the Library is available in the completion email sent to you by the GRS.

Thesis submission for the degree of Doctor of Philosophy

Thesis Title and Abstract

Declarations

Inclusion of Publications Statement

Corrected Thesis and Responses

Thesis Title

Genetic differentiation: from theory to practice

Thesis Abstract

Genetic differentiation is a vital aspect of population genetics and is a direct consequence of evolutionary forces acting on genetic diversity. By interpreting patterns of genetic differentiation, we can detect, infer and estimate the extent to which natural selection, genetic drift and gene flow affect genetic diversity. In this thesis, estimation of genetic differentiation is used as a tool to answer the following questions, three mainly theoretical, and the other an applied study on platypus conservation.

1. Can a form of linked selection termed associative overdominance (AOD) explain lower levels of genetic differentiation between populations (F_{ST}), and higher heterozygosity, than expected under neutrality in experimental populations (*Drosophila melanogaster*) and in a feral population (*Bos taurus*)?
2. Under which circumstances does AOD affect F_{ST} and heterozygosity?
3. Can AOD be detected in natural populations?
4. Do dams restrict gene flow among platypus groups?

AOD is triggered by the occurrence of recessive deleterious mutations that are physically linked and form haplotypes when recombination events are scarce, as in small populations. When haplotypes within an individual contain recessive deleterious mutations at different positions, a heterozygote for two different haplotypes is fitter than either one of the homozygotes. As a result, heterozygosity is higher, and F_{ST} lower than expected under neutrality. Here, using feral, experimental and computer-simulated populations, it is demonstrated how AOD might be prevalent in small po

pulations, and a framework for predicting and detecting AOD is provided.

The extent to which dams disrupt gene flow among platypus populations is investigated by using four rivers regulated by dams and three unregulated rivers. It was found that: genetic differentiation is significantly correlated with the number of generations since the dams were built; populations and individuals separated by dams are genetically more different than otherwise; and areas of high genetic differentiation coincide with the location of dams. It is suggested that dams jeopardise the long-term viability of platypus populations.

🎓 Welcome to the Research Alumni Portal, Luis Mijangos Araujo!

You will be able to download the finalised version of all thesis submissions that were processed in GRIS here.

Please ensure to include the **completed declaration** (from the Declarations tab), your **completed Inclusion of Publications Statement** (from the Inclusion of Publications Statement tab) in the final version of your thesis that you submit to the Library.

Information on how to submit the final copies of your thesis to the Library is available in the completion email sent to you by the GRS.

Thesis submission for the degree of Doctor of Philosophy

Thesis Title and Abstract

Declarations

Inclusion of Publications Statement

Corrected Thesis and Responses

ORIGINALITY STATEMENT

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.

COPYRIGHT STATEMENT

I hereby grant the University of New South Wales or its agents a non-exclusive licence to archive and to make available (including to members of the public) my thesis or dissertation in whole or part in the University libraries in all forms of media, now or here after known. I acknowledge that I retain all intellectual property rights which subsist in my thesis or dissertation, such as copyright and patent rights, subject to applicable law. I also retain the right to use all or part of my thesis or dissertation in future works (such as articles or books).

For any substantial portions of copyright material used in this thesis, written permission for use has been obtained, or the copyright material is removed from the final public version of the thesis.

AUTHENTICITY STATEMENT

I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis.

🎓 Welcome to the Research Alumni Portal, Luis Mijangos Araujo!

You will be able to download the finalised version of all thesis submissions that were processed in GRIS here.

Please ensure to include the **completed declaration** (from the Declarations tab), your **completed Inclusion of Publications Statement** (from the Inclusion of Publications Statement tab) in the final version of your thesis that you submit to the Library.

Information on how to submit the final copies of your thesis to the Library is available in the completion email sent to you by the GRS.

Thesis submission for the degree of Doctor of Philosophy

Thesis Title and Abstract

Declarations

Inclusion of Publications Statement

Corrected Thesis and Responses

UNSW is supportive of candidates publishing their research results during their candidature as detailed in the UNSW Thesis Examination Procedure.

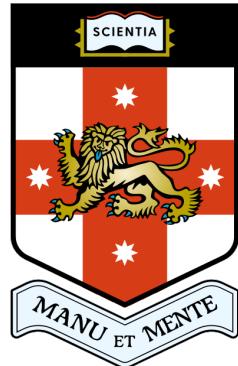
Publications can be used in the candidate's thesis in lieu of a Chapter provided:

- The candidate contributed **greater than 50%** of the content in the publication and are the "primary author", i.e. they were responsible primarily for the planning, execution and preparation of the work for publication.
- The candidate has obtained approval to include the publication in their thesis in lieu of a Chapter from their Supervisor and Postgraduate Coordinator.
- The publication is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in the thesis.

The candidate has declared that **their thesis contains no publications, either published or submitted for publication.**

Candidate's Declaration

I declare that I have complied with the Thesis Examination Procedure.



UNSW
CANBERRA

GENETIC DIFFERENTIATION: FROM THEORY TO PRACTICE

JOSE LUIS MIJANGOS ARAUJO

A thesis in fulfilment of the requirements for the degree of
Doctor of Philosophy

SCHOOL OF SCIENCES

JUNE 2021

THESIS ABSTRACT

Genetic differentiation is a vital aspect of population genetics and is a direct consequence of evolutionary forces acting on genetic diversity. By interpreting patterns of genetic differentiation, we can detect, infer and estimate the extent to which natural selection, genetic drift and gene flow affect genetic diversity. In this thesis, estimation of genetic differentiation is used as a tool to answer the following questions, three mainly theoretical, and the other an applied study on platypus conservation.

1. Can a form of linked selection termed associative overdominance (AOD) explain lower levels of genetic differentiation between populations (F_{ST}), and higher heterozygosity, than expected under neutrality in experimental populations (*Drosophila melanogaster*) and in a feral population (*Bos taurus*)?
2. Under which circumstances does AOD affect F_{ST} and heterozygosity?
3. Can AOD be detected in natural populations?
4. Do dams restrict gene flow among platypus groups?

AOD is triggered by the occurrence of recessive deleterious mutations that are physically linked and form haplotypes when recombination events are scarce, as in small populations. When haplotypes within an individual contain recessive deleterious mutations at different positions, a heterozygote for two different haplotypes is fitter than either one of the homozygotes. As a result, heterozygosity is higher, and F_{ST} lower than expected under neutrality. Here, using feral, experimental and computer-simulated populations, it is demonstrated how AOD might be prevalent in small populations, and a framework for predicting and detecting AOD is provided.

The extent to which dams disrupt gene flow among platypus populations is investigated by using four rivers regulated by dams and three unregulated rivers. It was found that: genetic differentiation is significantly correlated with the number of generations since the dams were built; populations and individuals separated by dams are genetically more different than otherwise; and areas of high genetic differentiation coincide with the location of dams. It is suggested that dams jeopardise the long-term viability of platypus populations.

**I dedicate this work to
my wife Cinthya,
to whom half of this thesis belongs.**

ACKNOWLEDGEMENTS

I consider as a fact that I would not have been able to finish this stage of my life without the invaluable support from my three principal supervisors: **Bill Sherwin**, **Harvi Sidhu** and **Clare Holleley**. Bill was immensely supportive from the first day when I arrived at the airport and then in each of the challenges I faced during these last years. The commitment that Bill took with me far exceeded all responsibilities from what could be expected from a supervisor. One of the many things that I learned from Bill is the academic rigour required in science. Somehow Harvi found the time to watch after me, at many different levels, despite his busy schedule, and encouraged me when I needed it the most. I learned from Harvi's leadership and from his attitude in solving problems. The amount of support that I received from Clare was unexpected, given that she had no responsibility towards me; despite this, Clare guided me intellectually and personally in an exemplary manner. I learned from Clare's intelligence and her creative and innovative thinking.

I am grateful for the advice, support and time that **Simon Watt** provided me. Simon helped me in many different aspects throughout these years and especially with the simulations. Simon was always happy to help me in solving questions and giving me feedback on my writing. I am also thankful for **Isaac Towers'** support in each of our meetings, particularly during the thesis submission. I want to thank **Zlatko Jovanoski** for his help and advice specially during the Shannon conference. I am grateful to **Aaron Adamack**, who introduced me to the simulations and for his detailed feedback. I want to express my gratitude to **Richard Nicholls**, who had the time to provide me with his wise advice despite being an extremely busy person.

I would like to thank **John Williams**, who allowed me to work with his Chillingham project data. I want to especially thank **Gilad Bino**, **Tahneal Hawke**, **Richard Kingsford** and all the **platypus volunteers** to allow me to analyse the platypus samples that they, with great effort, collected in the field. I thank **Thomas Grant** for sharing his invaluable experience with the platypus. I am grateful to **Jenna Day** and **Kimberly Noel Dias** for all their support in the lab for the platypus project.

I am thankful to **Jaime Gongora** for giving me his support as a mentor and as a friend. I want to thank all Bill's and Lee Ann Rollins' lab members, especially **Lee Ann Rollins**, **Alex Sentinella**, **John Sved**, **Juliet Byrnes** and **Jia Zhou**, who provided me with their feedback to improve my research. I am grateful for the great support of **Bernd Gruber** and **Arthur Georges** during the last stage of my PhD.

I am indebted to the **University of New South Wales Campus Canberra** for the essential economic support I received during my PhD. I am also grateful for the financial support from the **Taronga Conservation Society** and the **Australian Government's Environmental Water Holder** to carry out the platypus project, and the **Australian Research Council**, which funded most of the collection and genetics.

The economic and emotional support that I received from my parents, **Jose Luis** and **Mercedes**, my sister **Andrea** and my brother **Alexis** were without a doubt crucial for completing this stage. Lastly, I want to thank the persons I owe the most, my wife **Cinthya**, my son **Luca**, and my daughter **Mila** who gave me the motivation, encouragement, happiness, and time needed to finish this thesis.

"People have told me, directly and indirectly, that the neutral theory is not important biologically because neutral genes are not involved in adaptation. My own view is that what is important is to find the truth, and that if the neutral theory is a valid investigative hypothesis, then to establish the theory, test it against the data and defend it is a worthwhile scientific enterprise".

Motoo Kimura's closing statement of his article "The Neutral Theory of Molecular Evolution" published in the journal Scientific American in 1979.

SYMBOLS AND ABBREVIATIONS

Symbol/abbreviation	Meaning
ABM	Agent-based models
AFD	Allele frequency difference
AOD	Associative overdominance
BGS	Background selection
bp	Base pairs
c	Recombination rate
CDS	Coding sequences
cM	CentiMorgan
del/ind/cM	Deleterious alleles per individual per centiMorgan
e.g.	<i>Exempli gratia</i> ; for example
et al.	<i>Et alia</i> ; and others
F_{ST}	Fixation index
GTF	Gene transfer format
h	Dominance coefficient
He	Heterozygosity
He⁰	Heterozygosity at generation zero
HS	Mean heterozygosity across subpopulations
HT	Total heterozygosity of the subpopulations
i.e.	<i>Id est</i> ; that is
K	Number of equal sized populations
kbp	Thousand base pairs
LD	Linkage disequilibrium
In	Natural logarithm
m	Dispersal rate
Mbp	Million base pairs
n	Number
Nc	Census population size
Ne	Effective population size
NS	Non-synonymous mutations
p	Frequency of the alternative allele
ρ	Number of recombination events per generation occurring in a population; $\rho = 4 * Ne * r$
q	Frequency of the deleterious allele
r	Recombination rate in units of meiosis per generation
r²	Linkage disequilibrium statistic

R²	Coefficient of determination; proportion of the variance in the response variable that is predictable from the explanatory variable
s	Selection coefficient
S	Synonymous mutations
SNP	Single nucleotide polymorphism
t	Number of generations
T0	Beginning of the fly experiment
T1	Expected time for populations to reach half of the drift-dispersal F_{ST} equilibrium value
T2	Twice the expected time for populations to reach half of the drift-dispersal F_{ST} equilibrium value
w	Relative fitness of a genotype
W	Fitness of an individual

TABLE OF CONTENTS

THESIS ABSTRACT	II
ACKNOWLEDGEMENTS	IV
SYMBOLS AND ABBREVIATIONS	VII
1 Thesis introduction	1
1.1 Genetic differentiation as a tool to investigate evolutionary processes	1
1.2 Linked selection: an overlooked determinant of genetic diversity	3
1.3 Datasets to investigate linked selection	5
1.4 Computer simulations as a tool to investigate evolutionary processes	6
1.5 Research questions	9
2 Unexpected genetic resilience of small populations is produced by selection against deleterious alleles of linked genes	11
2.1 ABSTRACT	11
2.2 INTRODUCTION	12
2.3 METHODS	14
2.3.1 Evolutionary, demographic and genomic conditions triggering AOD	14
2.3.2 Study populations	16
2.3.3 Applying the AOD model to live populations	18
2.3.4 Investigating AOD hypotheses in live populations	18
2.4 RESULTS	22
2.4.1 Evolutionary, demographic and genomic conditions triggering AOD	22
2.4.2 Biases in genetic statistics are consistent with AOD in the fly experiment	25
2.4.3 Applying the AOD model to the fly experiment	26
2.4.1 AOD and genomic neighborhood in flies	27
2.4.2 Applying the AOD model to the Chillingham cattle	30
2.4.3 Patterns of polymorphism in the Chillingham cattle are consistent with linked selection	33
2.5 DISCUSSION	37
2.5.1 Evolutionary, demographic, and genomic conditions triggering AOD	37
2.5.2 AOD maintains genetic variation within and between populations	39
2.5.3 Detecting AOD	39
2.5.4 AOD implications for conservation and further research	41
2.6 CONCLUSION	43
2.7 AUTHORS CONTRIBUTIONS	44
2.8 EXTENDED DATA	45
2.9 SUPPLEMENTARY METHODS	74
2.9.1 Inference of N_e under high dispersal	74
2.9.2 Patterns of non-synonymous and synonymous mutations	75

2.9.2.1	Bioinformatic data processing pipeline to obtain the location of the NS in the Australian population	78
2.9.3	Methods to investigate associative overdominance	85
2.9.3.1	Genomic neighbourhood	85
2.9.3.2	Haplotype identification	88
3	Computer simulations to investigate linked selection mechanisms	90
3.1	INTRODUCTION	90
3.2	GENERAL SIMULATIONS	94
3.2.1	Model description	94
3.2.1.1	Purpose and patterns	94
3.2.1.2	Entities, state variables, and scales	94
3.2.1.3	Process overview and scheduling	95
3.2.1.4	Design concepts	98
3.2.1.5	Initialization	106
3.2.1.6	Input data	112
3.2.1.7	Submodels	112
3.2.2	Calibration of the effective size to census size ratio	126
3.2.3	Validation	128
3.2.3.1	The effective size to census size ratio	128
3.2.3.2	Heterozygote advantage	130
3.2.3.3	Fixation time for a neutral allele	130
3.2.3.4	Directional selection	131
3.2.3.5	Recombination	131
3.2.4	Simulation experiment	132
3.2.4.1	Effective population size	134
3.2.4.2	Dominance coefficient	134
3.2.4.3	Selection coefficient	135
3.2.4.4	Number of deleterious alleles per individual per centiMorgan	136
3.2.4.5	Number of offspring per pair of parents	141
3.2.4.6	Variance in the number of offspring per mating	141
3.2.4.7	Dispersal rate	143
3.2.4.8	Number of generations	147
3.2.4.9	Number of neutral loci and number of neutral alleles	147
3.2.4.10	Recombination	148
3.2.4.11	Fitness model	148
3.2.4.12	Selection model	148
3.2.4.13	Number of iterations	150
3.3	FLY SIMULATIONS	153
3.3.1	Model description	153
3.3.1.1	Purpose	153
3.3.1.2	Process overview and scheduling	153
3.3.1.3	Design concepts	156
3.3.1.4	Initialization	157
3.3.1.5	Input data	161
3.3.1.6	Submodels	163
3.3.2	Simulation experiment	165
3.3.2.1	Dominance	168

3.3.2.2	Selection coefficient	169
3.3.2.3	Initial allele frequency of the deleterious allele	169
3.3.2.4	Number of loci under selection to be simulated	169
3.3.2.5	Fitness model	169
3.3.2.6	Selection model	170
3.3.2.7	Effective population size	170
3.3.2.8	Census population size	170
3.3.2.9	Dispersal rate	170
3.3.2.10	Number of generations	171
3.3.2.11	Chromosome simulated	171
3.3.2.12	Number of neutral loci and number of neutral alleles	171
3.3.2.13	Recombination	173
3.3.2.14	Pre-adaptation phase	173
3.4	CHILLINGHAM SIMULATIONS	176
3.4.1	Model description	176
3.4.1.1	Purpose	176
3.4.1.2	Process overview and scheduling	176
3.4.1.3	Initialization	177
3.4.1.4	Input data	177
3.4.1.5	Submodels	179
3.4.2	Simulation experiment	180
3.4.2.1	Dominance	183
3.4.2.2	Selection coefficient	183
3.4.2.3	Initial allele frequency of the deleterious allele	184
3.4.2.4	Number of loci under selection to be simulated	184
3.4.2.5	Fitness model	184
3.4.2.6	Selection model	184
3.4.2.7	Effective population size	185
3.4.2.8	Census population size	185
3.4.2.9	Dispersal rate	185
3.4.2.10	Number of generations	185
3.4.2.11	Chromosome simulated	185
3.4.2.12	Number of neutral loci and number of neutral alleles	185
3.4.2.13	Recombination	186
3.4.2.14	Pre-adaptation phase	186
3.5	CONCLUSION	187
3.6	AUTHORS CONTRIBUTIONS	188
4	<i>Fragmentation by large dams and implications for the future viability of small platypus populations</i>	189
4.1	ABSTRACT	189
4.2	INTRODUCTION	190
4.3	METHODS	195
4.3.1	Study areas and field work	195
4.3.2	Laboratory work	196
4.3.3	SNP filtering	196
4.3.4	Data analyses	200

4.3.4.1	Genetic variation within groups	200
4.3.4.2	Investigating whether dams affect connectivity between platypus groups	200
4.4	RESULTS	203
4.4.1	Genetic variation within groups	203
4.4.2	Connectivity between platypus groups – effects of dams	205
4.5	DISCUSSION	213
4.6	CONCLUSION	217
4.7	AUTHORS CONTRIBUTIONS	217
4.8	EXTENDED DATA	219
5	<i>Thesis discussion</i>	231
5.1	Introduction	231
5.2	Answers to research questions	232
5.3	Discussion of results	235
5.4	Research gaps, future directions and challenges	237
5.5	Concluding remarks	239
6	<i>Appendix 1: Investigating AOD in each Chillingham chromosome</i>	240
7	<i>References</i>	270

1 Thesis introduction

In the PDF version of this thesis, each reference to a Figure, Table, Equation or Section contains a hyperlink that, when clicked, will take the reader to that element. For hyperlinks, typing **Alt+left arrow** (or **⌘+left** in a Mac) and **Alt+right arrow** (or **⌘+right** in a Mac) act like the “go back” and “go forward” buttons in a web browser.

1.1 Genetic differentiation as a tool to investigate evolutionary processes

In population genetics, genetic differentiation is a core concept that describes the dissimilarity between two or more populations in terms of their allele composition. The allele composition of populations can be characterised by estimating the allele frequencies of their genotyped loci. As an example, maximum genetic differentiation between two populations at a locus is reached if allele loss leads to a situation where the two populations share no alleles. In contrast, two populations are completely similar if alleles are found in the same frequency in both populations.

Given that genetic differentiation reflects the action of evolutionary forces such as gene flow, genetic drift and natural selection, the examination of its patterns is key to understand the ecological and evolutionary processes that are influenced by these forces (Charlesworth & Charlesworth, 2010). For instance, genetic differentiation is counteracted by gene flow, the exchange of genetic material between populations when a dispersing individual reproduce (Halliburton, 2004). The effect of genetic drift, the random change in allele frequencies that occurs from generation to generation, on genetic differentiation is proportional to population size in such a way that small populations become differentiated more rapidly than large populations (Frankham *et al.*, 2017). Natural selection can both increase genetic differentiation when selection eliminates the least fit alleles (directional selection) or decrease genetic differentiation when selection maintains different alleles in the populations (balancing selection; Allendorf *et al.*, 2013).

Being a field rich in theoretical models, population genetics provides us with a vast number of approaches to interpret genetic differentiation to infer, estimate and detect the action of evolutionary forces on genetic diversity. For instance, Wright's fixation index (1951), known as F_{ST} , was among the first models developed to characterise and measure genetic differentiation between populations. Wright envisaged F_{ST} as an index to measure the degree of approach towards fixation of the alleles within a subpopulation relative to the total population. Subsequently, other authors extended Wright's work to estimate F_{ST} using loci with more than two alleles (G_{ST} ; Nei, 1973), correct for population size and sampling effects (θ ; Weir & Cockerham, 1984), incorporate information about the mutation process of microsatellites (R_{ST} ; Slatkin, 1995) and haplotype diversity (AMOVA; Excoffier *et al.*, 1992). Even though F_{ST} and its relatives can predict evolutionary processes (Holsinger & Weir, 2009), they are not true measures of genetic differentiation in the sense that they are dependent on the diversity within populations (Meirmans & Hedrick, 2011), the number of populations analysed (Alcalá & Rosenberg, 2017) and are not monotonic (Sherwin *et al.*, 2017). Recent approaches have been developed to accommodate these mathematical restrictions (G'_{ST} and Jost's D; Hedrick, 2005; Jost, 2008). More recently, novel approaches based on information theory (Mutual Information; Sherwin *et al.*, 2017) and allele frequencies (Allele Frequency Difference; Berner, 2019) have distinct properties that make them valuable resources to interpret genetic differentiation. Note that each measure of genetic differentiation has advantages and drawbacks, and the decision of using a particular measure is usually based on the research question.

In particular, models predicting the behaviour of genetic diversity in the absence of natural selection (*i.e.*, neutral models) have played a predominant role to understand evolutionary processes. Neutral models can be used as a null hypothesis, that when tested and rejected, allows the inference of the effects that evolutionary forces and other factors have on genetic diversity. For instance, the Hardy-Weinberg principle (Hardy, 1908; Weinberg, 1908) predicts that in the simplest scenario (an infinitely large population of diploid organisms reproducing sexually at random in non-overlapping generations in which there is no migration, no mutation and no selection), allele frequencies will remain unchanged within a population between generations and

genotype proportions will be multinomial. The Hardy-Weinberg principle is used in a wide range of applications, from identifying genotyping errors (Chen, Cole, *et al.*, 2017) to inferring genetic structure (Pritchard *et al.*, 2000). Further neutral models show that F_{ST} can be predicted by population size and the dispersal rate between the populations (Wright, 1931) and that it reaches an equilibrium over time (Whitlock, 1992). Even though these models rely on assumptions that are difficult to meet in natural populations, such as constant and equal population sizes and continuous and symmetrical dispersal, they can be used as a null hypothesis to detect the action of natural selection under controlled conditions as in a laboratory setting or in computer simulations. This approach is used in Chapters 2 and 3 to investigate the effects of linked selection (see linked selection section below) using experimental populations of *Drosophila melanogaster* and computer simulations.

Genetic differentiation between populations is particularly suitable to investigate gene flow because, in the absence of selection, it is positively correlated with the geographic distance separating populations (Ramachandran *et al.*, 2005). This relationship can be used, for example, to identify a barrier separating two populations by testing whether genetic differentiation is higher than expected by the distance separating them. This approach is used in Chapter 4 to investigate the degree to which dams impede gene flow between platypus (*Ornithorhynchus anatinus*) populations.

1.2 Linked selection: an overlooked determinant of genetic diversity

By the time the field of population genetics was formally established, mainly through the foundation works of Wright (1931), Haldane (1932) and Fisher (1930), it was already clear that natural selection is the predominant force driving evolutionary change at the phenotypic level, as first proposed by Darwin (1859). However, little was known about the kind and amount of variation at the gene level due to a lack of a suitable technique to determine it unambiguously. In line with the assertion that natural selection is the primary evolutionary force at the phenotypic level, researchers at the time considered that natural selection played an equally important role in determining variation at the gene level. Two main genetic variation hypotheses existed. The classical hypothesis proposed that variation should be low because most mutations would be deleterious

and eliminated rapidly by natural selection (Muller, 1950). The balance hypothesis proposed that variation should be high because natural selection maintains variation by favouring heterozygote genotypes (Dobzhansky, 1955). Researchers using the first molecular techniques based on the electrophoretic mobility of proteins, including enzymes (Harris, 1966; Hubby & Lewontin, 1966; Johnson *et al.*, 1966; Lewontin & Hubby, 1966) reported unexpectedly high genetic variation, data which supported the balance hypothesis. However, soon after, Kimura (1968) proposed that most genetic variation between individuals and between species is selectively almost neutral (*i.e.*, it has little or no negative or positive effects on fitness for survival or reproduction), and thus the fate of neutral genes, whether lost or becoming fixed, is dictated by genetic drift rather than by natural selection. Kimura's quasi-neutral theory of molecular evolution (1983) is now an established model in population genetics to describe patterns of genetic diversity.

Patterns of neutral genetic variation are often consistent with the predictions of neutral models. However, this consistency breaks down if there is strong or persistent selection coupled with low recombination. It is now widely accepted that natural selection can affect not only genetic variation with direct consequences on fitness, but can also affect adjacent neutral genetic variation due to genetic linkage (Smith & Haigh, 1974). This effect has been increasingly known as "linked selection".

Linked selection can increase the loss of neutral genetic variation under two different scenarios. Firstly, an advantageous allele will spread through the population along with alleles at other loci, including neutral, that are linked to it. This scenario was first described by Smith and Haigh (1974) and termed "selective sweeps". Secondly, a deleterious allele will be eliminated from the population along with alleles at other loci linked to it. This scenario was first described by Charlesworth (1994) and termed "background selection".

Paradoxically, neutral genetic variation linked to deleterious alleles can be, in some circumstances, maintained instead of being lost (Rumball *et al.*, 1994; Gilligan *et al.*, 2005). This third scenario was first proposed by Frydenberg (1963) and termed "associative overdominance" (AOD). Associative overdominance (AOD) is a type of

linked selection which was first proposed by Frydenberg (1963). AOD is thought to arise if a genome location has two or more haplotypes (groups of alleles at different loci in linkage disequilibrium that are inherited together) and each one of these haplotypes have various deleterious recessive alleles, and these mutations are located in different positions in each haplotype (*i.e.*, haplotypes have opposite combinations of deleterious mutations). This arrangement of deleterious mutations in the haplotypes will result in the expression of the deleterious recessive mutations (*i.e.*, reduction in fitness) in an individual carrying the same haplotype in both chromosomes (homozygote). In contrast, in individuals carrying two different haplotypes (heterozygote), deleterious mutations will not be expressed due to the opposite or “repulsion” arrangement of deleterious recessive mutations in the haplotypes. This mechanism produces an apparent overdominance (*i.e.*, heterozygous individuals have a higher fitness than homozygous individuals) that results in the maintenance of genetic diversity within populations (*e.g.*, heterozygosity; H_e) and the depression of genetic differentiation between populations (*e.g.*, F_{ST}). A detailed description of the genetic mechanisms of AOD and the elements and factors involved in the development of AOD can be found in Section 3.2.1.4.1 “Basic principles”.

Given the accumulating evidence suggesting that the effects of linked selection are more pervasive than previously appreciated (Castellano *et al.*, 2016; Elyashiv *et al.*, 2016), it is likely that linked selection has the potential to disrupt the predictions of neutral models and therefore bias the many genetic resources that are used to guide the conservation and restoration of biodiversity.

1.3 Datasets to investigate linked selection

In Chapters 2 and 3, I used three distinct types of datasets, each with particular attributes that allow me to investigate AOD from different perspectives:

- 1) **Experimental populations (*Drosophila melanogaster*; Holleley, 2009).** The experimental design of these populations aimed to recreate *in vivo* theoretical island models (Wright, 1943), in which three dispersal regimes were imposed upon population pairs. Important attributes of this dataset are: each dispersal regime was replicated in

twelve populations, population pairs reached an F_{ST} equilibrium and more importantly this experimental design provided me with a priori expectations for H_e and F_{ST} , which I could use as null hypotheses to compare the observed results. Additionally, *D. melanogaster* has characteristics that are particularly suitable to investigate AOD. These characteristics are: a large effective population size (N_e) that allows for the accumulation of a great number of deleterious alleles compared to other species; a small genome that allows for a higher density of targets of selection; and a fewer number of recombination events compared to other species, because recombination only occurs in females. Finally, because *D. melanogaster* is a model species, several genomic resources, such as recombination maps and genome annotations, are available (Thurmond *et al.*, 2019).

2) Chillingham cattle (*Bos taurus*; Williams *et al.*, 2016). This is a feral population in northern England that has not been subjected to selection for commercial traits and remained small ($N_e = 8$) and isolated for at least 350 years (~67 generations). In this dataset more than 777,000 SNPs were genotyped in 20 individuals. Several genomic resources are also available in *B. taurus* (Ma *et al.*, 2015; Eusebi *et al.*, 2020).

3) Computer simulations (agent-based model, Chapter 3). I developed a simulation model in the programming language R (R Core Team, 2021) that allowed me to model with great detail the main evolutionary processes driving linked selection, namely recombination, natural selection and genetic drift. Additionally, my model allowed me to simulate the main genomic, life history and demographic characteristics of my study species. The main advantage of computer simulations is that they allow us to determine explicitly several parameters regarding loci with deleterious alleles, which are unknown in empirical datasets, such as their location, dominance coefficients and selection coefficients and the number of recombination events occurring between them.

1.4 Computer simulations as a tool to investigate evolutionary processes

Hypothesis testing can be of various forms, including empirical, theoretical or a combination (Haller, 2014). Empirical approaches, such as the fly and cattle investigations, test verbal hypotheses by gathering and analysing data and confronting

the results with the hypothesis predictions. The testing of a hypothesis through a theoretical approach usually does not involve data gathering, but instead, these approaches make use of mathematical principles to translate verbal hypotheses into abstract models. One of the aims of theoretical models is to identify the overarching elements, principles, mechanisms, and interactions that govern a particular system or process. To achieve this goal, models simplify the process or system under study by making assumptions. Assumptions are statements describing the conditions in which a specific model is situated. Theoretical models can be seen as "proof-of-concept" models because they test unambiguously and precisely whether certain assumptions lead to specific predictions (Servedio *et al.*, 2014).

The development of theoretical models has been the primary approach used in population genetics to explain and describe how genetic diversity is affected by evolutionary forces, such as mutation, genetic drift, natural selection, gene flow and recombination. For instance, under certain assumptions, we can predict with great precision the rate at which heterozygosity is lost over generations (Crow & Kimura, 1970) or the amount of genetic differentiation existing between populations connected by constant gene flow (Wright, 1943). More sophisticated theoretical models have been developed using complex mathematical approaches such as gas diffusion theory (Kimura, 1964) or coalescent theory (Kingman, 1982). These approaches allow us to predict how genetic diversity is affected by evolutionary forces in more complex scenarios.

A central assumption in many population genetic models is that loci segregate independently from each other within a population (Kimura, 1965; Nagylaki, 1974) and therefore these models ignore linkage disequilibrium (LD; the non-random association between alleles at different loci). This assumption holds for the great majority of loci and many demographic and evolutionary scenarios, because loci are reassorted within a population, by several processes: random mating, sexual reproduction, the independent assortment of chromosomes and recombination, which includes cross over and gene conversion. This assumption is especially convenient in theoretical models because it makes possible to isolate the effects of evolutionary forces on individual loci.

Moreover, the modelling of the interaction between natural selection, genetic drift and recombination at multiple linked loci becomes too complex to be solved analytically and therefore are intractable for theoretical models, so that computer modelling is required.

Computer simulations allow us to explore scenarios that cannot be solved analytically by theoretical models. Furthermore, computer simulations can be viewed as formal experiments because the same variables, factors and manipulations can be performed within a simulated system (Peck, 2004). Advantages of computer simulations over empirical research include easier manipulation, lower costs and shorter timescales (Peck, 2004). Particularly, agent-based models (ABM) offer a convenient way to extend theoretical models. ABM simulate a system using a collection of autonomous individual entities called agents, whose interactions are dictated by a set of rules (Bonabeau, 2002). A particular, and one of the more critical, attribute of ABM is that of emergence. Emergence can be defined as that system behaviour that can emerge from the agents behaviour and from their environment instead of being imposed by equations or rules (Grimm *et al.*, 2020).

1.5 Conservation genetics

The field of conservation genetics commonly investigates the negative consequences of small population size and isolation of once large and connected populations (Frankham *et al.*, 2009). Populations experiencing these circumstances have a reduced ability to respond and adapt to environmental change due to the loss of genetic variation resulting from the increase of the effects of genetic drift and a lack of gene flow. As a result, both phenotypic variation and effectiveness of natural selection (Franklin & Frankham, 1998) are reduced, and in consequence, evolutionary potential is compromised. Furthermore, in small and isolated populations, the probability of mating between relatives increases, leading to the exposure and accumulation of deleterious variants and reduced heterozygosity in overdominant loci, an effect known as inbreeding depression. Ultimately, individual fitness and population persistence will be affected negatively (Frankham, 2005; O'Grady *et al.*, 2006).

1.6 Research questions

This thesis uses genetic differentiation and computer simulations to investigate relevant questions related to gene flow, genetic drift and natural selection.

In Chapters 2 and 3, general computer simulations are used to answer the question:

What are the evolutionary, demographic and genomic conditions in which linked selection via deleterious alleles impacts genetic variation within and between populations?

In these chapters, computer simulations are designed firstly to reproduce the predictions of neutral models for genetic variation within populations (heterozygosity) and between populations (F_{ST}). Then this design is used to observe how neutral genetic variation behaves when recombination and natural selection are added to the simulation model.

In the second part of Chapters 2 and 3, more specific computer simulations are used to answer the question:

Can computer simulations parameterised with realistic values explain patterns of genetic diversity observed in live populations?

Two different live populations datasets were used. The first dataset was an experiment conducted using the vinegar fly (*Drosophila melanogaster*) and performed by Holleley *et al.* (2014). This experiment replicated *in vivo* the neutral island model developed by Wright (1943). The second dataset was from the Chillingham white cattle population of northern England (*Bos taurus*; Williams *et al.*, 2016). These cattle have not been subjected to selection for commercial traits and have remained as a small population closed to immigrants population for at least 350 years (~67 generations).

In the final part of Chapter 2, two different approaches are used to answer the question:

Can linked selection be detected in natural populations?

The first method involves sets of multiple regressions to identify the likely distance at which surrounding deleterious alleles and recombination influence each locus. The second method is an approach that identifies haplotypes to measure the amount of recombination and the number of potentially deleterious alleles within the haplotypes.

In Chapter 4, the platypus (*Ornithorhynchus anatinus*) is used as a study species to answer the question:

Do dams affect dispersal and gene flow between platypus populations?

In this chapter, platypus individuals were sampled from four rivers regulated by dams and three unregulated rivers, to serve as alternative and null hypotheses, respectively.

2 Unexpected genetic resilience of small populations is produced by selection against deleterious alleles of linked genes

Note that this chapter is formatted for submission to Nature Ecology and Evolution.

2.1 ABSTRACT

Low genetic diversity restricts the capacity of populations to adapt to environmental change, elevating extinction risk (Willi *et al.*, 2006; Jones *et al.*, 2012). Perplexingly, some small populations retain disproportionately high levels of genetic diversity, in some cases more than 100% of the expected diversity under neutrality (Groombridge *et al.*, 2000; Lawrence *et al.*, 2008; Schou *et al.*, 2017; Saccheri *et al.*, 2020). The processes responsible for this unexpected resilience to the depletion of genetic diversity are still unresolved. Here I show in wild, experimental and computer simulated populations how a form of selection on linked deleterious mutations, associative overdominance (AOD), is prevalent in small populations. AOD occurs when two or more deleterious alleles are segregating at linked locations in the genome. With an appropriate combination of selection, population size and low recombination (Zhao & Charlesworth, 2016; Becher *et al.*, 2020; Gilbert, Pouyet, *et al.*, 2020) the selection to remove the deleterious alleles is impeded by the advantage to double heterozygotes with the deleterious alleles in repulsion (*i.e.*, on different haplotypes), in which the deleterious effects of the mutations are obscured. Genetic diversity is then maintained at other sites in the genomic regions surrounding the deleterious loci, with alternative alleles being maintained in the haplotypes linked to each deleterious mutation. Nevertheless, this outcome is not inevitable – with other realistic combinations of parameters, the effect of selection on deleterious loci will be to remove genetic diversity (Charlesworth, 1994). Hence the outcome depends crucially on several genetic particularities of the species and its population size. I provide a framework to predict and detect the prevalence and intensity of AOD. This widespread but largely neglected evolutionary process could be leveraged to preserve heterozygosity, slow allelic fixation and improve conservation outcomes. I highlight the risk of adverse effects during genetic rescue if AOD is ignored. This pervasive linked selection disrupts classical populations genetic theory, which assumes selective neutrality and unlinked loci. Importantly, the resulting statistical

biases may disproportionately impact threatened species, where analytical accuracy is paramount to species survival.

2.2 INTRODUCTION

Associative overdominance (AOD) is a type of selection that can reduce the rate of loss of genetic diversity within a population, which is paradoxically driven by the presence of deleterious alleles (Zhao & Charlesworth, 2016; Becher *et al.*, 2020; Gilbert, Pouyet, *et al.*, 2020). It could arise when, in a particular part of the genome, there are two or more haplotypes that bear recessive deleterious alleles at different positions. This arrangement means that heterozygous individuals for these haplotypes will be at an advantage because dominant advantageous alleles mask the detrimental effects of the deleterious alleles on the alternative haplotype (see schematic representation and further details in Chapter 3, Figure 3.2 and Section 3.2.1.4.1). This advantage to heterozygotes can maintain the deleterious alleles in the population for much longer than predicted for the same selection on unlinked loci. More broadly, polymorphism will also be retained at other sites in the haplotypes since their fate is bound together with that of the selected sites by genetic linkage. Analytical, numerical and simulation studies (Latter, 1998; Zhao & Charlesworth, 2016; Becher *et al.*, 2020; Gilbert, Pouyet, *et al.*, 2020) suggest that the occurrence of AOD becomes a significant factor influencing genetic diversity only in some circumstances. These conditions occur for certain combinations of the parameters describing selection, effective population size and recombination (Gilbert, Pouyet, *et al.*, 2020). In particular, AOD becomes important when the physical linkage between deleterious alleles is broken infrequently by recombination events. Because the effect of recombination scales with effective population size (N_e ; Stumpf & McVean, 2003), the critical parameter is the product of recombination rate and N_e .

For other parameter combinations, including larger scaled recombination rates, theoretical models suggest that selection on the deleterious alleles has the opposite effect to AOD: the loss of genetic diversity at linked loci would be accelerated, as the deleterious alleles are removed from the population – an effect called background selection (Charlesworth, 1994). Recent studies (Zhao & Charlesworth, 2016; Becher *et*

al., 2020) suggest that both effects can be detected in samples from *Drosophila* populations.

The values of critical parameters in natural populations, especially Ne and dominance of deleterious alleles, are challenging to estimate. In particular, it has been difficult to predict whether either process, AOD or background selection, is relevant in species of conservation concern with small Ne . As well as uncertainty about the real values of the critical parameters, some of the transient dynamics of linkage, after populations are founded, may not be captured by existing models.

To explore the above issues, I investigated the relative importance of AOD in small experimental populations of vinegar flies (*Drosophila melanogaster*) that were comparable in size to some threatened populations. In addition, I searched for signatures of AOD in a feral cattle population (*Bos taurus*; Williams *et al.*, 2016). I conducted simulations to establish whether realistic parameter values can explain deviations from neutral expectations that could be attributed to AOD, which include higher genetic diversity within populations (measured as heterozygosity, He) and lower differentiation between populations (F_{ST}) than expected under neutrality, each expected to be more prevalent in regions of the genome with lower recombination rate and a higher density of deleterious alleles.

2.3 METHODS

2.3.1 Evolutionary, demographic and genomic conditions triggering AOD

I constructed an agent-based model (see Bonabeau, 2002 for a description of agent based approaches) in the programming language R v4 (R Core Team, 2021) to investigate the evolutionary, demographic and genomic conditions under which AOD can occur, including scenarios that are intractable using equation-based approaches (DeAngelis & Rose, 1992). In brief, my model simulates a pair of populations connected by gene flow, made up of diploid organisms that reproduce in non-overlapping generations. Each individual has a pair of homologous chromosomes that contains interspersed selected and neutral loci. For the initial generation, the genotype for each individual's chromosomes was randomly drawn from distributions at linkage equilibrium and in Hardy-Weinberg equilibrium. See Chapter 3 for a complete description of the model.

I explored a range of values for four main variables that I identified as likely drivers of AOD based on prior research (Zhao & Charlesworth, 2016; Becher *et al.*, 2020; Gilbert, Pouyet, *et al.*, 2020) and my preliminary simulations (Chapter 3, Section 3.2.4). The variables were: number of deleterious alleles per individual per centiMorgan (del/ind/cM; 1, 10, 20 and 40), selection coefficient (s ; 0.0001, 0.0005, 0.001 and 0.005), dominance coefficient (h ; 0, 0.1, 0.3 and 0.5, where $h = 0.5$ is semidominant and $h = 0$ completely recessive), and effective population size (N_e ; 10, 50, 100, 150 and 200). The variable del/ind/cM depends on three variables: the initial frequency of the deleterious allele (q), the total number of loci with deleterious alleles to be simulated and the length of the chromosome in centiMorgans. Two of these three variables should be maintained constant in the simulations to isolate the effect that the del/ind/cM might have on AOD. In these simulations q ($q = 0.15$) and the total number of loci with deleterious alleles ($n = 2,000$) that were simulated were kept constant across all the combinations. Note that a q value = 0.15 is within the range of the frequency of deleterious alleles per gene that would be normally found in *Drosophila* and human individuals (Supplementary information Table 3.13, column del/ind/cM). Further justification for the choice of the parameters and values used in the simulations is provided in Chapter 3 section 3.2.4. For each combination of variable values, I simulated two populations connected by a constant dispersal rate ($m = 0.01$).

This dispersal rate was chosen to accommodate my computational resources with the number of generations required to reach drift-dispersal equilibrium (Whitlock, 1992) calculated as:

$$t_{1/2} = \frac{\ln(1/2)}{\ln\left[(1-m)^2\left(1-\frac{1}{2Ne}\right)\right]}, \quad \text{Equation 2.1}$$

where $t_{1/2}$ is the expected number of generations to reach the half drift-dispersal equilibrium F_{ST} value and m is the dispersal rate.

For all simulation scenarios, populations were allowed to reproduce for 62 generations, which is the estimated number of generations required to reach drift-dispersal equilibrium for the largest Ne that was tested ($Ne = 200$). Based on an initial power analysis (Chapter 3, Section 3.2.4.13), each simulation scenario was replicated 250 times. To quantify the impact of AOD on the neutral loci, I used two parameters: heterozygosity (He) and the genetic differentiation between populations (F_{ST}). These were standardised as the deviation from the expected value predicted by neutral models, *i.e.*, bias of an estimator calculated as: (observed - expected) / expected.

The expected value of He under neutrality (Crow & Kimura, 1970) was calculated as:

$$He_t = He_0 \left(1 - \frac{1}{2Ne_{dispersal}}\right)^t, \quad \text{Equation 2.2}$$

where He_0 is heterozygosity at generation zero, t is the number of generations and $Ne_{dispersal}$ is the effective population size (Ne) of each subpopulation corrected for dispersal. This correction is needed to adapt equation 2.2 to use for the heterozygosity of a system of two populations connected by dispersal, rather than the single isolate for which it was designed, because dispersal from other populations reduces the rate of loss of heterozygosity (Wang & Whitlock, 2003). Further details of the calculation of $Ne_{dispersal}$ are in Supplementary Methods 2.9.1. For example, in my neutral simulations, the decline in heterozygosity was substantially slower with dispersal ($m = 0.01$) than without ($m = 0$); the slower rate being equivalent to a 50% increase in Ne , as estimated by fitting Equation 2.2 to the decline in heterozygosity.

The expected equilibrium F_{ST} under neutrality (Takahata, 1983) was calculated as:

$$F_{ST} = \frac{1}{4N_e m \left(\frac{n}{n-1} \right)^2 + 1}, \quad \text{Equation 2.3}$$

where N_e is the effective population size for each subpopulation, m is dispersal rate and n the number of subpopulations (two).

2.3.2 Study populations

The first biological dataset was from a study of laboratory populations of *Drosophila melanogaster* (Holleley *et al.*, 2014) created to investigate *in vivo* the predictions of theoretical island models, initially developed by Wright (1943) and extended by many others (Takahata, 1983). Importantly, this experimental design provided me with a priori expectations for equilibrium H_e and F_{ST} , which I could use as null hypothesis to compare the observed results.

The experimental design comprised three groups of twelve population-pairs, each pair connected by dispersal. The populations contained 25 males and 25 females, with a known effective population size of $N_e = 14.3$ (England, 1997; Gilligan, 2001; Gunn, 2003; Holleley, 2009). Populations were established from four genetically isolated stock lines, each founded by 200 individuals (equal sex ratio) and kept in separate stock cages for ~60 generations. Founding individuals were collected from Tyrell's Winery, Hunter Valley, New South Wales (Australia). In each of the three groups of experimental populations, six population pairs were initiated with populations from different stock lines. Hence they were expected to be genetically differentiated at the start of the experiment (due to the divergence among stock lines under culture). The other six population pairs were initiated with both populations from the same stock line, and thus these experiments started with genetically similar populations (Extended Data Figure 2.1a). This difference in starting conditions allowed to track the progress of these population pairs towards drift-dispersal equilibrium by observing the convergence of initially disparate F_{ST} values towards a single intermediate equilibrium value (Extended

Data Figure 2.1b and Extended Data Figure 2.1c). Within each group a range of dispersal rates was imposed: high ($m = 0.04$; exchanging two individuals every generation; with a predicted neutral equilibrium $F_{ST} = 0.099$), moderate ($m = 0.01$; one individual every two generations; equilibrium $F_{ST} = 0.304$) and low ($m = 0.0025$; one individual every eight generations; equilibrium $F_{ST} = 0.636$). From each subpopulation, 24 individuals (equal sex ratio) were sampled at the beginning of the experiment (T0), at the expected time for populations to reach half of the drift-dispersal equilibrium value (T1; Equation 2.1; Whitlock, 1992) and then again after twice this period (T2). Specifically, the populations under the high dispersal regime were bred for $T2 = 12$ non-overlapping generations, moderate dispersal for $T2 = 26$ generations and low dispersal for $T2 = 34$ generations. Initially, 73 single nucleotide polymorphism (SNP) loci were genotyped in 2,304 individuals and twelve microsatellites were genotyped in 4,224 individuals. My final dataset comprised ten microsatellites and 32 SNPs after applying the following filters to both datasets: minimum minor allele frequency > 0.05 , missing values per locus $< 5\%$ and Hardy-Weinberg equilibrium exact test p-value $< 10^{-20}$.

Homogeneity of reproductive success between immigrant and resident individuals was monitored throughout to detect any differential fitness between dispersers and residents (Holleley *et al.*, 2011). Populations were cytogenetically screened for the presence of inversions that could alter linkage between loci (Lindsley & Grell, 1968; Krimbas & Powell, 1992).

The second biological dataset was from the Chillingham white cattle population of northern England (*Bos taurus*; Williams *et al.*, 2016). These cattle have not been subjected to any artificial selection, *e.g.*, for commercial traits, and the population has been closed to immigrants for at least 350 years (~67 generations). At the time of sampling, the cattle had a small census population size (approximately 50 males and 50 females), had experienced significant inbreeding and had an estimated effective population size of $Ne = 8$ (calculated from genetic data, and therefore compensating for any overlap of generations; Visscher *et al.*, 2001). We confirmed this Ne by using the linkage disequilibrium method in NeEstimator V2.1 (Do *et al.*, 2014). Initially, more than 777,000 SNPs were genotyped in 20 individuals. My final dataset comprised 38,589 SNPs

and 16 individuals after applying the following filters: minimum minor allele frequency > 0.05 , missing values per locus $< 5\%$, Hardy-Weinberg equilibrium exact test p-value $< 10^{-20}$ and missing values per individual $< 5\%$.

2.3.3 Applying the AOD model to live populations

To investigate whether the genetic patterns observed in the fly experiment and the Chillingham cattle were consistent with AOD, I parameterised two additional versions of my AOD model with the particular genetic and demographic characteristics for the two species. Specifically, my AOD model simulated: the number, location, frequencies, selection coefficients (s) and dominance coefficients (h) of deleterious alleles. The number and location of recombination events were based on high-resolution recombination maps of *D. melanogaster* (Comeron *et al.*, 2012) and *B. taurus* (Ma *et al.*, 2015). The demographic history and population size were based on the experimental design (for the *Drosophila* experiment) and historical records (for the Chillingham cattle; Williams *et al.*, 2016). To simulate the *Drosophila* data, neutral markers were placed in the same genomic location as those that were empirically genotyped. Fly simulations were based on the chromosome arm 2L because this chromosome had the greatest number of loci genotyped in the fly experiment. Chillingham simulations were based on the chromosome 18 because this chromosome had signatures of AOD according to my analyses (see Appendix 1: Investigating AOD in each Chillingham chromosome).

2.3.4 Investigating AOD hypotheses in live populations

I wished to assess whether selection and recombination affected H_e and F_{ST} as predicted by my AOD hypotheses in my study populations, *i.e.*, higher H_e and lower F_{ST} than expected under neutrality, each more prevalent in genomic regions with lower recombination and higher density of deleterious alleles. To this end, I developed two methods for evaluating this. The first method consisted of performing sets of multiple linear regression analyses. For the fly experiment, I used as response variables: the F_{ST} of each locus averaged across the twelve replicates of each dispersal regime; and for H_e the amount of H_e lost from the beginning (T0) to the end (T2) of the experiment for each locus, averaged across the twelve replicates of each dispersal regime. For this analysis, I used as explanatory variables a proxy for the number of putative targets of selection

(*i.e.*, deleterious alleles, as described below); and the recombination rate surrounding each locus. Studies modelling linked selection (Comeron, 2014; Elyashiv *et al.*, 2016; Casillas & Barbadilla, 2017) suggest that for any one locus, there is likely to be a specific but unknown limit to the distance at which each locus is influenced by surrounding loci under selection and recombination. I call this unknown distance the “genomic neighbourhood”. To infer the likely size of the genomic neighbourhood, I first calculated the explanatory variables surrounding each locus within windows of different physical distance (*i.e.*, numbers of base pairs). Then I performed a separate regression analysis for each window (see schematic representation in Supplementary Methods Figure 2.4 a). I also replicated all the regression analyses using genetic distance (*i.e.*, recombination in centiMorgans) instead of physical distance to infer the size of the genomic neighbourhood (Supplementary Methods Figure 2.4 b).

For the first explanatory variable for the flies and as a proxy for the number of targets of selection within a gene, I used the difference between the number of non-synonymous mutations (mutations resulting in substitution of a different amino acid sequence; NS) and synonymous mutations (mutations resulting in the same amino acid sequence; S ; *i.e.*, $NS - S$). The location of S and NS sites were obtained from the Ensembl’s (www.ensembl.org; Yates *et al.*, 2020) *D. melanogaster* database, which covers two divergent populations in the U.S.A. (Jordan *et al.*, 2007) and Africa (Begin & Lindfors, 2005). Greater number of synonymous mutations (*i.e.*, $S > NS$) can be taken as evidence that a gene is more selectively constrained (Kryazhimskiy & Plotkin, 2008; Chu & Wei, 2019), particularly when two divergent populations are compared (Kryazhimskiy & Plotkin, 2008), as is the case with the Ensembl’s *D. melanogaster* database. For this analysis I assumed that, on average, the more selectively constrained a gene is, the less targets of selection it has, because highly deleterious alleles are eliminated by selection rapidly (Charlesworth, 2015). Therefore, a negative result in my proxy would indicate a lower number of targets of selection, and a positive result would indicate a higher number of targets of selection in a given gene.

To investigate whether the density patterns of NS and S across the genome were similar in the Ensembl database and in my fly populations, I analysed publicly available raw

sequence data from an Australian fly population collected ~1,000 km apart away from the source population of the experimental fly populations (Griffin *et al.*, 2017; Supplementary Methods 2.9.2). I found good concordance between the patterns of density of non-synonymous mutations observed in the Australian population and the Ensembl populations (Supplementary Methods Figure 2.2 and Supplementary Methods Figure 2.3). As the second explanatory variable for the flies, I used the recombination rate measured in centiMorgans (cM) surrounding each locus (for further details see Supplementary Methods 2.9.3.2). To calculate recombination rates, I relied on published recombination maps for *D. melanogaster* (Comeron *et al.*, 2012).

In the Chillingham cattle for the response variable, I used He calculated in bins of 1 Mbp, rather than per locus, to reduce the statistical noise of the dataset arising from the large number of SNPs genotyped. For the first explanatory variable for the cattle, I noted that in mammals, the difference between NS and S is smaller than in insects, suggesting that selection is less efficient in these species owing probably to a lower Ne , because purging of deleterious mutations is more efficient in large populations than in small ones (Galtier, 2016; Chen, Glémén, *et al.*, 2017). Both the smaller Ne and smaller difference between NS and S , add to the error of my proxy ($NS - S$). Hence for the Chillingham cattle, I used the number of NS as the proxy for the number of targets of selection. The data were obtained from Ensembl's *B. taurus* database. For the second explanatory variable in the cattle, to calculate recombination rates, I relied on a recently published recombination map of *B. taurus* (Ma *et al.*, 2015). The large number of loci genotyped allowed me to use my regression method for each chromosome individually.

To further assess the suitability of my regression method, I applied it to the output of the fly and Chillingham simulations as well as their corresponding cases in neutral simulations.

The large number of SNPs genotyped in the Chillingham cattle allowed me to identify the haplotypes segregating in this population and investigate whether my AOD explanatory variables, *i.e.*, recombination (cM/Mbp) and my proxy of targets of selection (NS), could explain haplotype polymorphism. I developed a script in R to identify and visualise haplotypes (Supplementary Methods 2.9.3.2). I considered a

haplotype to be those contiguous SNPs in complete linkage disequilibrium (*i.e.*, with an r^2 statistic (Hill & Robertson, 1968) = 1) and containing more than ten polymorphic loci. I calculated the r^2 statistic using the R package *snpStats* (Clayton, 2015). For each haplotype, I measured the number of centiMorgans/Mbp, number of polymorphic loci/Mbp and the number targets of selection (as measured by the proxy)/Mbp. To test non-linear relationships between the variables measured, I used Generalised Additive Models (GAM; Wood *et al.*, 2015) and to visualise the results, I used the R package *mgcViz* (Fasiolo *et al.*, 2020). I used as response variable (the log of the number of polymorphic loci/Mbp in each of the haplotypes identified), and as explanatory variables (the log of the number of proxies for targets of selection/Mbp) and the number of centiMorgans/Mbp in each of the haplotypes identified.

2.4 RESULTS

2.4.1 Evolutionary, demographic and genomic conditions triggering AOD

My general modelling scenarios showed that the occurrence and strength of AOD was promoted by high selection coefficients (s), low dominance coefficients (h), low effective population sizes (N_e) and high densities of deleterious alleles per individual per centiMorgan (del/ind/cM). These trends are summarized in Figure 2.1, Extended Data Figure 2.2 , and Extended Data Figure 2.3 . My simulations demonstrate that when deleterious alleles are completely recessive ($h = 0$), AOD has the effect of biasing neutral genetic diversity within populations (H_e) upwards up to 114% (Figure 2.1 and Extended Data Figure 2.3a) and biasing neutral genetic differentiation between populations (F_{ST}) as low as -64% (Extended Data Figure 2.2 and Extended Data Figure 2.3b). In contrast, scenarios where dominance coefficients (h) were high showed a reduction in the impact of AOD, this is especially evident in scenarios where high dominance coefficients ($h = 0.5$) are coupled with high selection coefficient values ($s = 0.005$). In general, the impact of AOD on F_{ST} was weaker but more prevalent than the impact of AOD on H_e across the different scenarios.

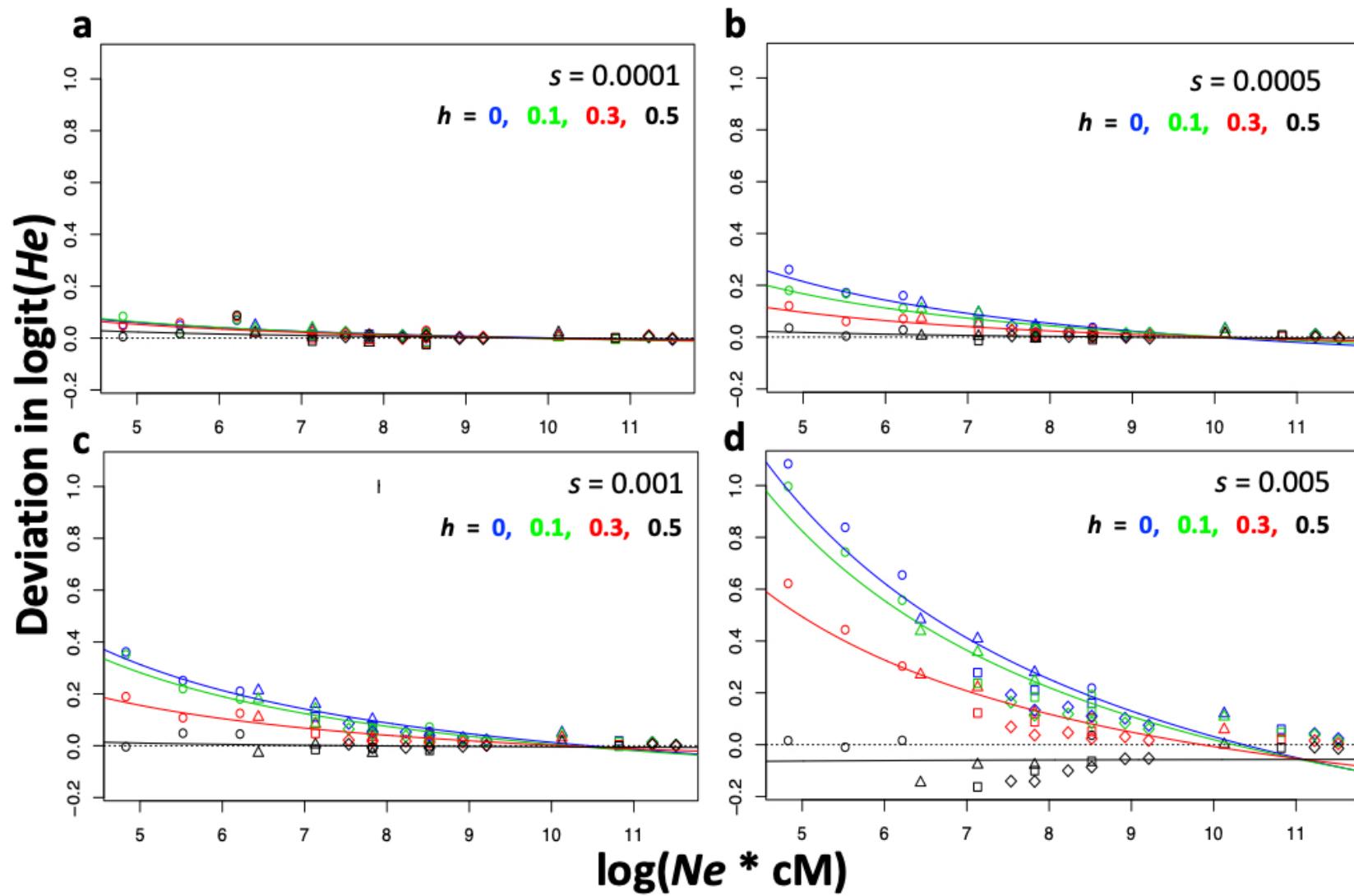


Figure 2.1 | The effects of AOD on heterozygosity in the general simulation model. Plots show heterozygosity (He) bias as a function of the product of Ne and the total genetic map length of the chromosome (cM). Each subplot shows a different level of selection, going from weak selection (top left; **a**) to strong selection (bottom right; **d**). Each symbol point in each subplot represents a separate simulation scenario. Symbol color shows the level of dominance (h ; blue = 0, green = 0.1, red = 0.25 and black = 0.5, where $h = 0.5$ is semidominant and $h = 0$ completely recessive). Symbol shape indicates length of the chromosome in centiMorgans (cM; circles = 12.5, triangles = 25, squares = 50 and rhombus = 500). Note that longer chromosomes have less del/ind/cM because other relevant variables are being held constant. In all scenarios a total of 2,000 loci with deleterious alleles evenly distributed across the genome were simulated. The initial allele frequency of each deleterious allele (q) was 0.15, resulting in that each individual carried, on average, 555 deleterious alleles (see Chapter 3 Section 3.2.4.4 for further details). In all scenarios, each individual had 50 neutral loci evenly distributed across the genome. Each neutral locus had initially nine alleles with a frequency of 1/9 or 0.111 (see Chapter 3 Section 3.2.4.9 for further details).

2.4.2 Biases in genetic statistics are consistent with AOD in the fly experiment

At the end of the fly experiment (T2), the same amount of genetic differentiation (F_{ST}) was attained by population pairs that were formed from different stock lines and population pairs that were created from the same stock lines (Extended Data Figure 2.1b and Extended Data Figure 2.1c). This convergence of F_{ST} from high and low starting values strongly suggests that population pairs had reached drift-dispersal equilibrium. Mild hybrid advantage or disadvantage were equally likely, and each only occurred in one replicate (Holleley *et al.*, 2011). Three very small inversions were detected in two of the four stock lines and occurred in 0 - 3.3% of individuals examined (Holleley, 2009), indicating that inversions did not have a major influence on the patterns observed in the data.

As predicted by my hypothesis, H_e was biased upwards relative to neutral expectations (Equation 2.2) and neutral simulations (Extended Data Figure 2.4a and Extended Data Figure 2.4b), while F_{ST} was biased downwards relative to neutral expectations (Equation 2.3) and neutral simulations (Extended Data Figure 2.4c and Extended Data Figure 2.4d). The same patterns were observed in the three dispersal regimes of the fly experiment. Mean H_e in the high dispersal regime was 61% (coefficient of variation 30%) higher than expected; in the moderate dispersal regime, it was 69% (cv 12%) higher than expected, and in the low dispersal regime, it was 70% (cv 10%) higher than expected. Mean F_{ST} in the high dispersal regime was 61% (cv 30%) lower than expected, while in the moderate dispersal and low regimes, it was 69% (cv 12%) to 70% (cv 10%) lower than expected, respectively. H_e and F_{ST} showed low variability across populations at the end of the experiment (T2), despite the populations starting from different initial levels of genetic diversity and differentiation (Extended Data Figure 2.5-2.7).

2.4.3 Applying the AOD model to the fly experiment

My AOD simulation model, customised for the fly experiment, replicated the observed patterns of both F_{ST} and He . As shown by a linear regression of the empirical results on the simulation results, my simulations replicated the F_{ST} patterns ($R^2 = 0.50$, p-value = 0.02; Figure 2.2b) more accurately than the He patterns ($R^2 = 0.43$, p-value = 0.04; Figure 2.2a). No significant regression was observed in neutral simulations for either F_{ST} or He (Data not shown).

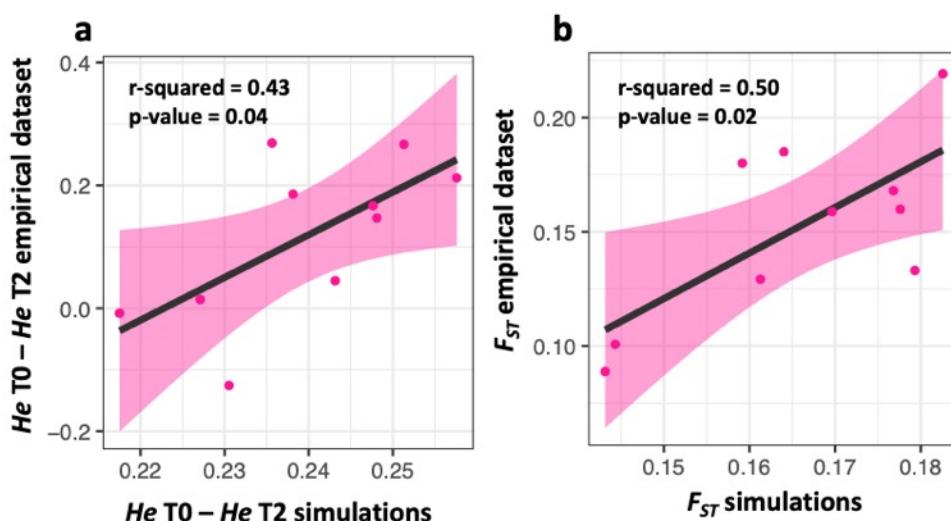
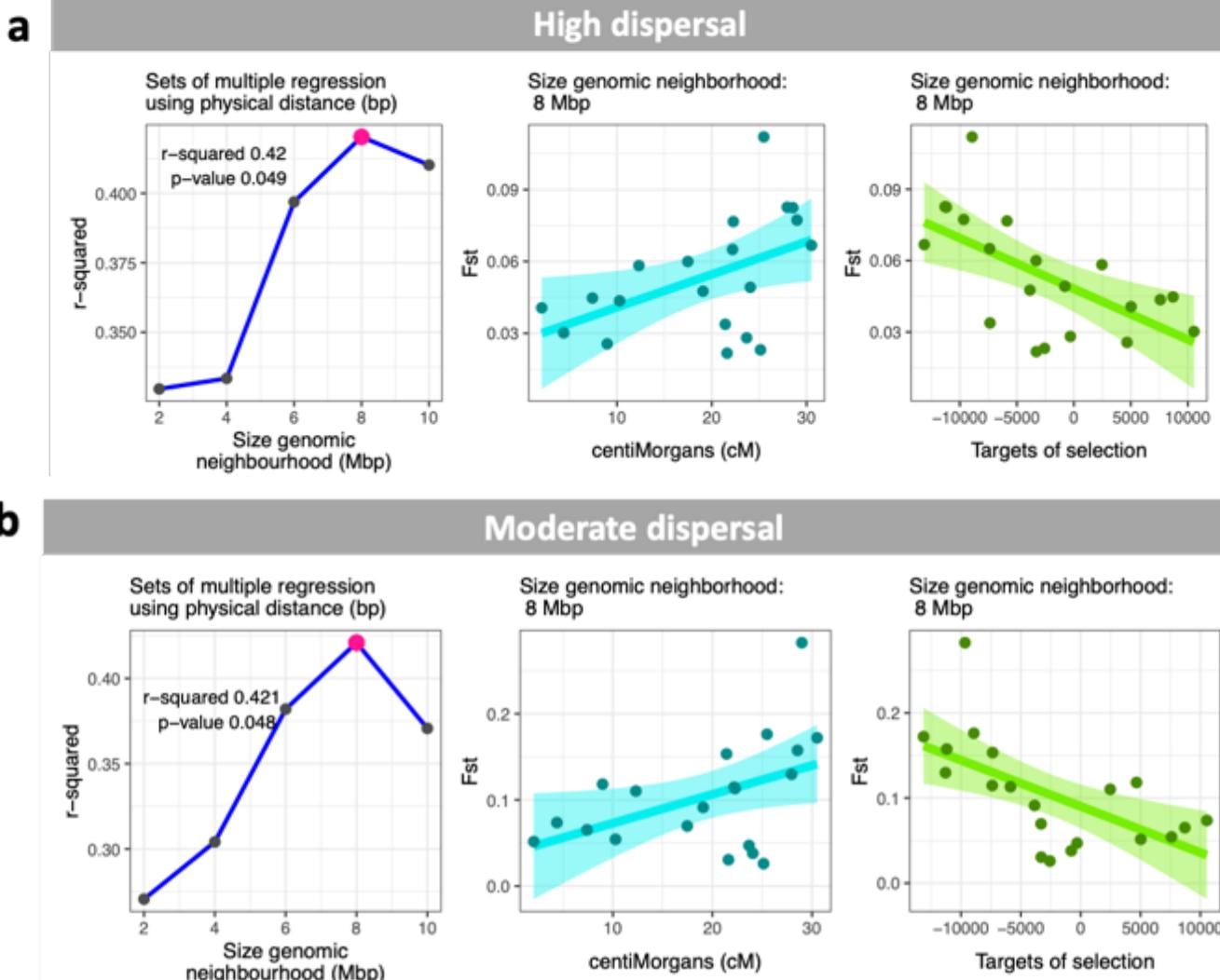


Figure 2.2 | Fly empirical dataset vs AOD simulations of loci located in the chromosome arm 2L. **a)** Regression of the amount of heterozygosity lost from T0 to T2 in the fly empirical dataset vs values in the simulations of loci that were in the same genome location as the neutral loci genotyped in the fly experiment. **b)** Regression of F_{ST} values in the fly empirical dataset vs values in the simulations of loci that were in the same genome location as the neutral loci genotyped in the fly experiment. Each dot represents one locus from the chromosome arm 2L at the end of the experiment (T2) in populations under low dispersal. In the case of the fly empirical dataset, values were averaged across the twelve population replicates. In the case of the simulations, values were averaged across 250 replicates. Regression of the fly empirical values against those produced by neutral simulations showed no significant association (Data not shown).

2.4.1 AOD and genomic neighborhood in flies

Results from my multiple regression method based on physical distance (base pairs) and using recombination and proxies of targets of selection as explanatory variables, and F_{ST} as response variable suggest that the size of the genomic neighbourhood (the likely distance at which each locus is influenced by surrounding deleterious alleles and recombination) was between 8 and 10 Mbp (the size with the highest R^2 value) and it was of similar size across the three dispersal regimes (Figure 2.3). F_{ST} was positively associated with recombination rate measured in centiMorgans and negatively associated with my proxy for targets of selection. These associations were significant after Bonferroni correction in the high ($R^2 = 0.42$, p-value < 0.05) and moderate ($R^2 = 0.42$, p-value < 0.05) dispersal regimes, but not in the low dispersal regime ($R^2 = 0.10$, p-value > 0.05). Similar patterns to those mentioned above were observed in my regression method using genetic distance (centiMorgans; Extended Data Figure 2.10).



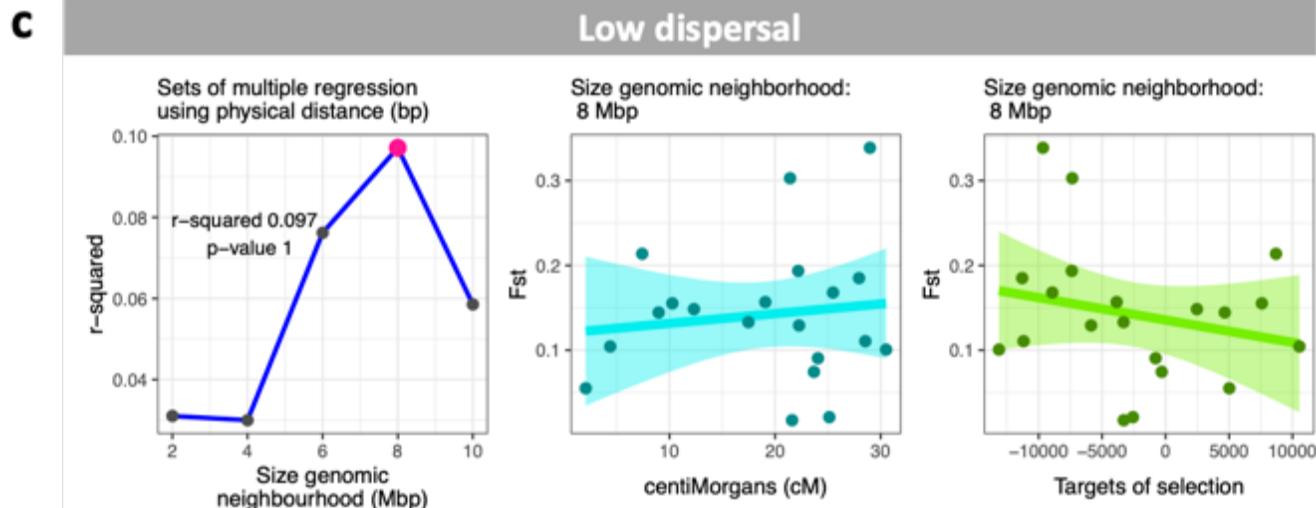


Figure 2.3 | Relationship between F_{ST} and recombination and proxies for targets of selection in the fly experiment using physical distance (bp). Each row presents the results for each dispersal regime: **a**) high dispersal; **b**) moderate dispersal; and, **c**) low dispersal. The first column is the inference of the size of the genomic neighbourhood using physical distance (bp), where each point is the R^2 of each multiple regression (F_{ST} on recombination [cM] and proxies for targets of selection) after Bonferroni correction, and the pink point shows the regression with the highest R^2 value. The second column is the regression line between F_{ST} and the recombination rate within the inferred genomic neighborhood (centiMorgans per 8 Mbp). The third column is the regression line between F_{ST} and proxies for targets of selection within the inferred genomic neighbourhood (8 Mbp). Each point in the second and third columns is the F_{ST} of one locus averaged across the twelve population pairs of each dispersal regime. Shaded areas are 95% confidence intervals of the regression line. The units of proxies for targets of selection is the difference between the number of non-synonymous (NS) and synonymous (S) mutations (*i.e.*, NS - S) in a given gene; a negative value indicates a lower number of targets of selection, and a positive result indicates a higher number of targets of selection.

Results from my regression method using physical distance and the loss of He as explanatory variable suggest that the size of the genomic neighbourhood for He was smaller than for F_{ST} , being between 2 and 4 Mbp, and it was of similar size across the three dispersal regimes (Extended Data Figure 2.11). A slower loss of He was associated with a higher number of proxies for targets of selection and a lower recombination rate. These associations were not significant after Bonferroni correction in any of the dispersal regimes (high dispersal $R^2 = 0.100$; moderate dispersal $R^2 = 0.006$; low dispersal $R^2 = 0.140$). Similar patterns to those given above were observed in the regression method using genetic distance (centiMorgans; Extended Data Figure 2.12).

Results from my regression method applied to the fly simulations show the same observed patterns as in the fly experiment in F_{ST} (Extended Data Figure 2.13) and in the loss of He (Extended Data Figure 2.14). Consistent with the results in the fly experiment dataset, the size of the inferred genomic neighbourhood was shorter using the loss of He as explanatory variable than when F_{ST} was used. Furthermore, I observed that the regression of the size of the genomic neighbourhood at which the regression signal reaches its peak agreed with the distance at which pairwise LD (as measured by the statistic r^2) decayed below 0.2, a threshold that is commonly used to imply that two loci are unlinked (Delourme *et al.*, 2013; Li *et al.*, 2014). I also observed that the regression signal reaches a peak in significance and then declines. No regression signal was observed when the regression method was applied to neutral simulations (Data not shown).

2.4.2 Applying the AOD model to the Chillingham cattle

In the Chillingham cattle, the proportion of polymorphic loci (9.1%; 70,150 polymorphic of 772,488 SNPs) was higher than predicted by neutral models (2.4%; Williams *et al.*, 2016), and in contrast to other cattle breeds (Williams *et al.*, 2016), polymorphic loci are clustered in blocks in specific chromosome regions (*e.g.*, Figure 2.4b). Using my method to visualize LD patterns of each chromosome, I found that blocks of polymorphic loci were usually in complete LD and were located in some of the chromosome regions where recombination is lower than 1 cM/Mbp (*e.g.*, Figure 2.4a).

AOD simulations of the Chillingham cattle replicated the genomic patterns observed in the empirical dataset, *i.e.*, blocks of polymorphic loci with high genetic diversity were located, in general, in the same chromosomal position in the simulations (Figure 2.4c) and in the empirical dataset (Figure 2.4b). Neutral simulations (Figure 2.4d) show a depletion of H_e compared to the AOD simulations, and the pattern of blocks of polymorphic loci observed in the empirical dataset was absent in this neutral simulations.

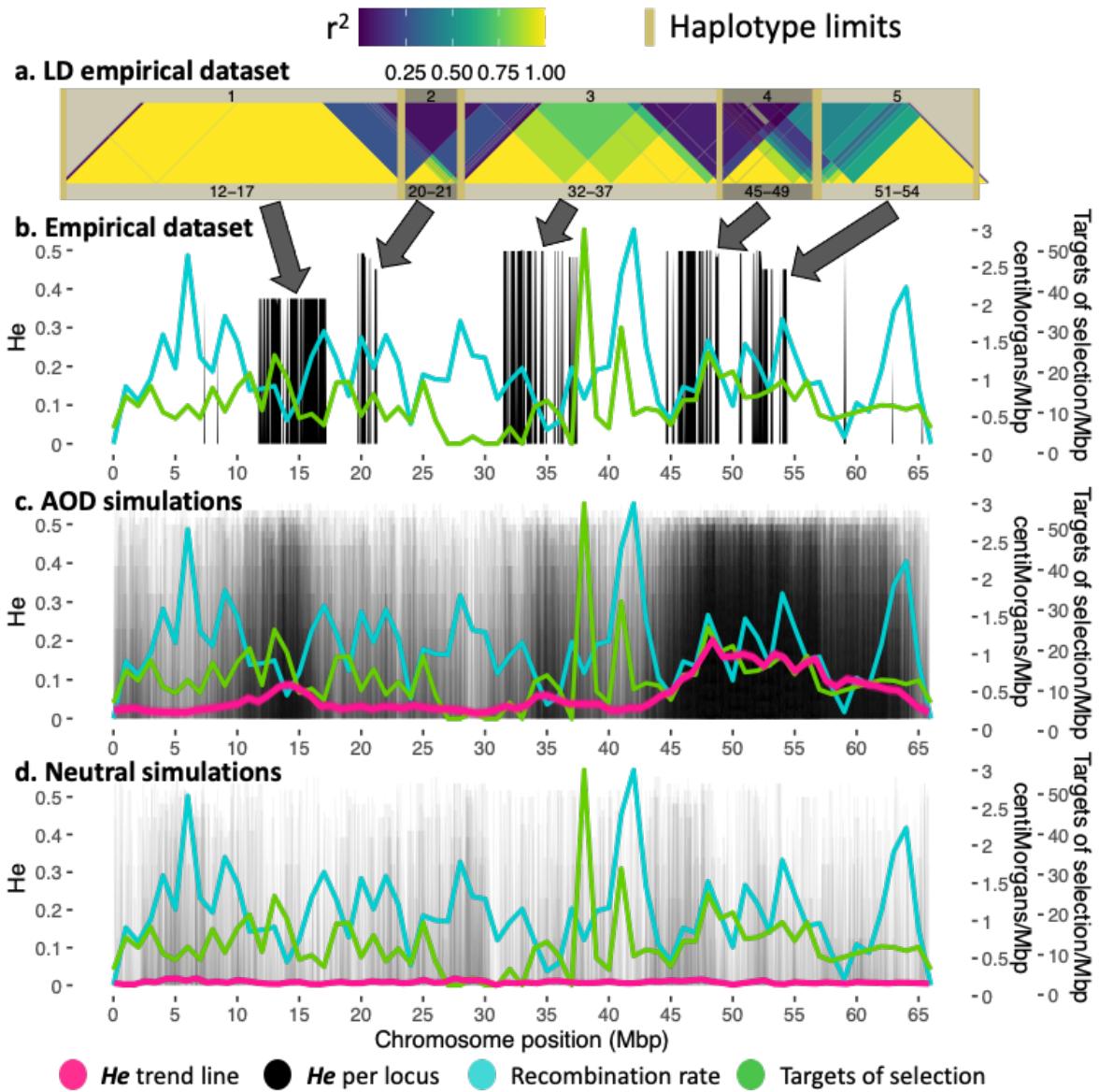


Figure 2.4 | Simulations replicate patterns of polymorphism in the Chillingham cattle.

a) Linkage disequilibrium (LD) heatmap of the chromosome 18 where the colour shading indicates the strength of LD (r^2). Genomic locations are shown on the horizontal axis in Mbp, and LD blocks (haplotypes) appear as bright yellow triangles and delimited by dark yellow vertical lines. Arrows show the location of five different haplotypes in **b**. **b)** Patterns of He (black columns), recombination (blue line) and proxies for targets of selection (green line) across the chromosome 18. **c)** 200 independent AOD simulation replicates (*i.e.*, with selection) showing that blocks of polymorphism are located approximately in the same chromosome position as in the empirical dataset in **b**. The trend line (pink) is a regression line using LOESS (locally estimated scatterplot smoothing). **d)** 200 independent neutral simulation replicates (*i.e.*, without selection). The trend line (pink) is a regression line using LOESS.

2.4.3 Patterns of polymorphism in the Chillingham cattle are consistent with linked selection

Applying my multiple regression method in the Chillingham cattle, I observed a negative association between H_e and recombination and a positive association between H_e and my proxy for targets of selection in 43% of the chromosomes I tested (12 of 28 chromosomes; Appendix 1: Investigating AOD in each Chillingham chromosome). I observed that the inferred size of the genomic neighbourhood (distance at which the correlation signal reaches its peak) generally agreed with the physical distance at which pairwise LD (r^2) decayed below 0.2. Results from applying my regression method to the Chillingham simulations show the same observed patterns in the empirical dataset (Extended Data Figure 2.15). No correlation between response variables and explanatory variables was observed when the regression method was applied to neutral simulations (Data not shown).

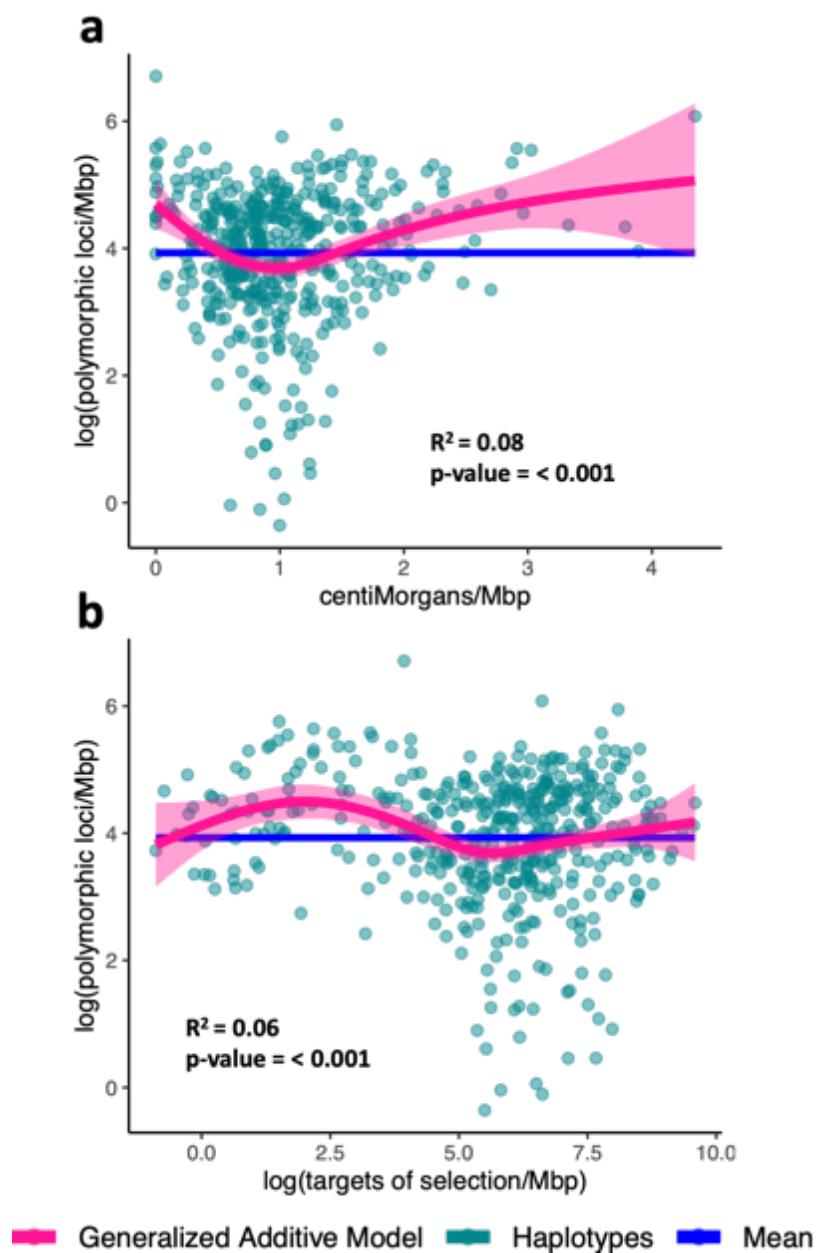
My method to identify haplotypes detected 454 haplotypes in the 28 chromosomes of the Chillingham cattle that were analysed (Appendix 1: Investigating AOD in each Chillingham chromosome). Chromosome X was not analysed because the recombination map for this chromosome was not available (Ma *et al.*, 2015). In addition, chromosome 28 had a level of polymorphism that was too low to identify haplotypes and thus was not included in the analyses.

My analyses using generalised additive models (GAM) revealed a significant, non-linear and bimodal relationship between the log of the number of polymorphic loci/Mbp and the number of centiMorgans/Mbp in each haplotype (Figure 2.5a; $R^2 = 0.08$, p-value < 0.001). The bimodal relationship shows an increment in the polymorphism of haplotypes with low recombination rate (< ~0.75 cM/Mbp) and in haplotypes with high recombination rate (> ~1.5 cM/Mbp). In contrast, haplotypes with a moderate recombination rate (~1 cM/Mbp) show, on average, a lower polymorphism.

Further results from my GAM analyses show a significant non-linear relationship between the log of the number of polymorphic loci/Mbp and the log of the number of proxies for targets of selection/Mbp (Figure 2.5b; $R^2 = 0.06$, p-value < 0.001]. An increment in polymorphism is observed in those haplotypes with < log(5 targets of

selection/Mbp) and a decrease in those haplotypes with $> \log(5)$ targets of selection/Mbp). The observed distribution of haplotypes polymorphism across the range of the number of targets of selection may be explained by linked selection predictions. Most haplotypes with low polymorphism are concentrated in the right-hand side of the distribution in Figure 2.5b (*i.e.*, higher number of targets of selection), as predicted by background selection models. In contrast, haplotypes with high polymorphism are distributed more evenly across the horizontal axis.

Finally, I found a significant association when the level of recombination and the number of proxies for targets of selection in each haplotype were used to explain haplotype polymorphism (Figure 2.5c; $R^2 = 0.11$, p-value < 0.001). The combined effects of recombination and targets of selection seem to converge in a specific area in which polymorphism of haplotypes decreases markedly (dark area), identifying the area in which background selection might be operating.



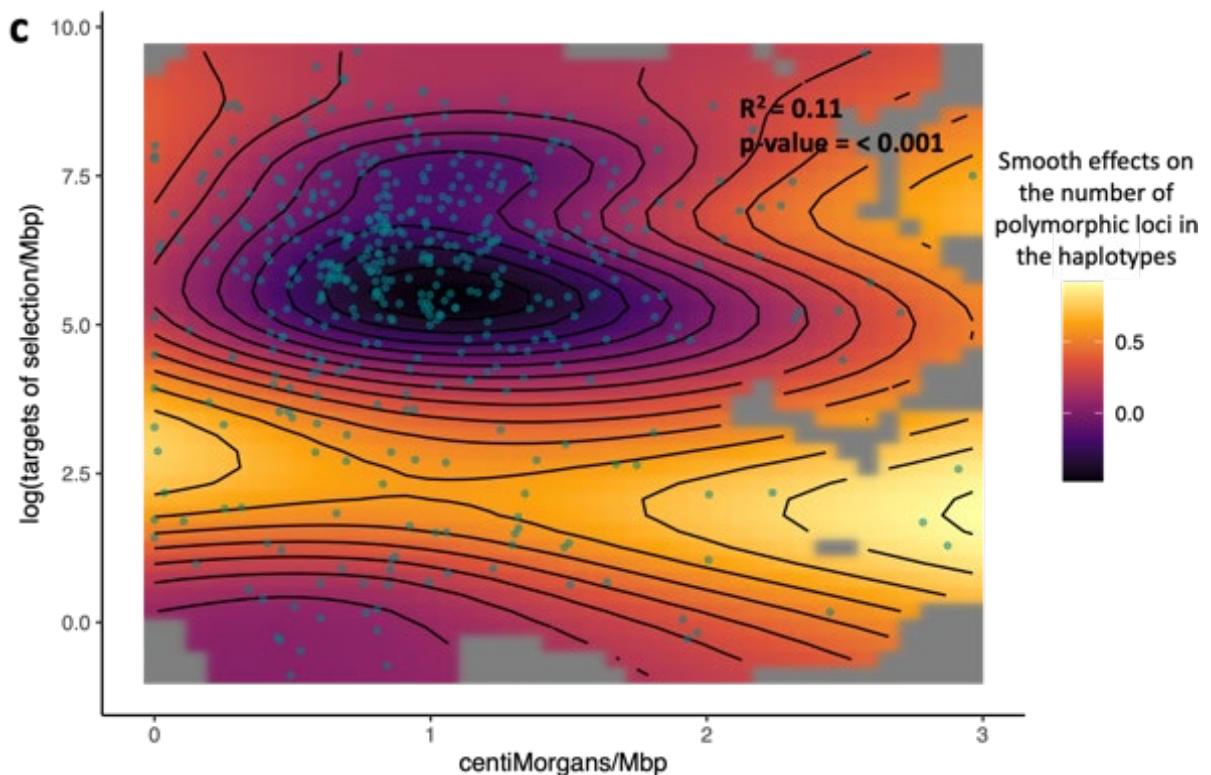


Figure 2.5 | Patterns of polymorphism within haplotypes can be explained by recombination and proxies for targets of selection in the Chillingham cattle. **a)** Effect of recombination (centiMorgans/Mbp) on the log of the number of polymorphic loci/Mbp per haplotype of the 454 haplotypes identified on the 28 chromosomes of the Chillingham cattle using generalised additive models (GAM; pink line; $R^2 = 0.08$, $p\text{-value} < 0.001$). The mean of the log of the number of polymorphic loci/Mbp per haplotype is indicated by a blue line. The pink shaded area is the 95% confidence intervals of the GAM. **b)** Effect of the log of the number of proxies for targets of selection/Mbp in each haplotype on the log of the number of polymorphic loci/Mbp in each haplotype using GAM (pink line; $R^2 = 0.06$, $p\text{-value} < 0.001$). The mean of the log of the number of polymorphic loci/Mbp per haplotype is indicated by a blue line. The pink shaded area is the 95% confidence intervals of the GAM. **c)** Two-dimensional plot of the smoothed effects of recombination and proxies for targets of selection in each haplotype on the log of the number of polymorphic loci in each haplotype using GAM ($R^2 = 0.11$, $p\text{-value} < 0.001$). Each point in **a**, **b** and **c** is a haplotype. Grey shades in **c** are areas in which the smooth effects could not be calculated due to absence of data points. The units of proxies for targets of selection is the number of non-synonymous mutations (NS).

2.5 DISCUSSION

I provide empirical evidence for the occurrence of associative overdominance (AOD) in living populations. AOD is a poorly recognised evolutionary force that can be a major driver of genetic diversity in small populations. Counterintuitively, by triggering AOD, deleterious alleles can be an unexpected source of resilience to the loss of genetic diversity in small populations. My AOD simulations recapitulated empirical results in experimental fly and feral cattle populations, where small populations retained more genetic diversity than expected under selective neutrality and in the case of the flies resisted the development of genetic differentiation between populations. These findings were true for both replicated, controlled, experimental fly populations (Figure 2.2) and in an unmanaged, feral, inbred cattle population (Figure 2.4). My findings suggest that at the chromosomal level AOD is a function of the local density of deleterious alleles per individual per centiMorgan (del/ind/cM). At the population level, AOD is a function of the number of recombination events in the population per generation, which can be approximated by $\rho = 4 * Ne * r$, where r is the recombination rate in units of meiosis per generation (Peñalba & Wolf, 2020; Figure 2.1 and Extended Data Figure 2.2 and Extended Data Figure 2.3).

Given my observations from two highly divergent species, which are confirmed by my simulations, and results from others (Latter, 1998; Schou *et al.*, 2017; Becher *et al.*, 2020; Gilbert, Pouyet, *et al.*, 2020), it is probable that AOD may be a stronger and more prevalent phenomenon in natural populations and across the genome than previously thought.

2.5.1 Evolutionary, demographic, and genomic conditions triggering AOD

My generalised AOD model (Extended Data Figure 2.3a) suggests that AOD becomes an important driver of genetic diversity (H_e) for populations with an effective population size (Ne) < 75 , a value seen for many populations which are of conservation concern (Frankham, 1995). This trend holds even for scenarios with the lowest density of deleterious alleles per individual per centiMorgan that I tested (1 del/ind/cM). Conversely, in populations with an $Ne = 300$, the effects of AOD on H_e were limited even with very high densities of deleterious alleles (40 del/ind/cM). I showed that the positive

or negative impacts of linked selection for population management are of significant concern in small populations and should be actively managed to improve conservation outcomes.

The experimental fly populations might exemplify the conditions in which AOD strongly impacts genetic diversity: a species which due to its large effective population size accumulates abundant deleterious alleles and has a small genome resulting in high density of deleterious alleles. A further factor that could have increased the impact of AOD in the experimental populations is that recombination is absent in males in *D. melanogaster*. Moreover, before commencing the dispersal experiment, stock lines were bred for several generations at relatively large census population sizes (~2,000). This factor might have prevented the elimination of deleterious alleles via founder effects (Santos *et al.*, 2012). However, the general agreement between the fly and cattle results indicates that AOD may be general.

In agreement with earlier theoretical research (Zhao & Charlesworth, 2016; Gilbert, Pouyet, *et al.*, 2020), my simulations showed that the impact of AOD on H_e and F_{ST} becomes more noticeable as selection coefficients increase and dominance coefficients decrease, *i.e.*, for more recessive deleterious alleles (Figure 2.1, Extended Data Figure 2.2 and Extended Data Figure 2.3). In contrast, the occurrence of AOD was not seen in most of the scenarios where dominance was the highest ($h = 0.5$). In reality, it is likely that the dominance coefficient of most deleterious alleles is < 0.25 (Manna *et al.*, 2011). Evidence suggests that deleterious alleles in complex organisms, such as humans, have higher selection coefficient values (Huber *et al.*, 2017) and deleterious alleles with higher selection coefficients are more recessive (Huber *et al.*, 2018). This supports my findings that AOD is expected to be an important driver of genetic diversity, not just in small populations of species with a small genome and a high density of deleterious alleles per individual per centiMorgan, as in the fly populations, but also in small populations of species with complex genomic features, as in the Chillingham cattle.

2.5.2 AOD maintains genetic variation within (H_e) and between populations (F_{ST})

My simulations and analyses of flies and cattle (Figure 2.1-2.4) show that selection favouring heterozygous haplotypes with recessive deleterious alleles at different positions (*i.e.*, AOD) resulted in the maintenance of genetic diversity (H_e) of neutral loci across generations. H_e was higher than expected by neutral models in all of the datasets (up to 114% higher in general simulations, up to 70% higher in the fly experiment and 280% higher in the Chillingham cattle). These values are similar to those reported by previous studies in *D. melanogaster* (Rumball *et al.*, 1994; Schou *et al.*, 2017). In contrast, for the Chillingham cattle, the number of polymorphic loci in the X chromosome (2.9%) is closer to the neutral expectation (2.4%), suggesting that AOD has not occurred on that chromosome. This observation is in line with findings indicating that deleterious alleles in the X chromosome might be eliminated more efficiently by selection (Vicoso & Charlesworth, 2006) because the detrimental consequences of deleterious alleles cannot be masked in males, which carry only one X chromosome.

My simulations (Extended Data Figure 2.8 and Extended Data Figure 2.9) show that if in two populations connected by dispersal H_e is maintained by AOD, genetic differentiation will be depressed because F_{ST} depends upon allele frequencies becoming more and more different between populations. Therefore, when allele loss is suppressed (*i.e.*, H_e maintained) by a mechanism such as AOD, F_{ST} will be lower than neutral expectations. By replicating Wright's island model (1943) in the fly experiment, it was shown that F_{ST} was biased downwards from neutral expectations (Extended Data Figure 2.4). This deviation could not be explained by differential reproductive success of immigrants or inversion polymorphisms (Holleley, 2009; Holleley *et al.*, 2011).

2.5.3 Detecting AOD

To properly account for AOD and allow conservation managers to make data-driven decisions about the likelihood of AOD, we must have a method to detect its occurrence and predict the severity of its impact. Until now, there has been no way to do so, here, I provide the first practical approach to detect AOD.

By investigating different sizes of the “genomic neighbourhood” (the likely distance at which each locus is influenced by surrounding deleterious alleles and recombination), I showed that at an optimal neighbourhood size, AOD results in a positive association of H_e and a negative association of F_{ST} relative to the local density of targets of selection (based on non-synonymous and synonymous mutations). These associations were not observed in neutral simulations (Data not shown). In simulations and empirical data, I observed that the regression signal reaches a peak in significance and then declines after the pairwise LD, as measured by r^2 , decayed below 0.2, a threshold that is commonly used to imply that two loci are unlinked (Delourme *et al.*, 2013; Li *et al.*, 2014).

I found a stronger correlation when using F_{ST} than when using H_e . This may be because in small populations, H_e has a high stochastic variance (Avery & Hill, 1977). In contrast, F_{ST} progresses towards an equilibrium (Whitlock, 1992) as observed empirically in the fly experiment in which high and low F_{ST} values at the start of the experiment converged at the end of the experiment (Extended Data Figure 2.1). Additionally, F_{ST} has a smaller variance than H_e , attributes that have been exploited by outlier tests to identify loci under selection (Bierne *et al.*, 2013). These differences between F_{ST} and H_e might also explain the fact that the genomic neighbourhood for H_e was smaller than for F_{ST} .

Generalised additive models (GAM) using as explanatory variables proxies for targets of selection and recombination, and as response variable haplotype polymorphism (Figure 2.5), revealed polymorphism patterns in agreement with predictions of linked selection. First, haplotypes in regions with very low recombination (< 0.5 cM/Mbp) show high polymorphism consistent with AOD; second, haplotypes located in regions of higher recombination (0.5 to 2 cM/Mbp) show low polymorphism possibly due to ongoing effects of background selection. Finally, haplotypes in regions with higher recombination (> 2 cM/Mbp) appear able to escape from the effects of background selection as predicted by linked selection models (Cutter & Payseur, 2013). I did not observe a clear increase in the level of polymorphism with a higher number of targets of selection (Figure 2.5b), possibly because background selection depletes all the deleterious alleles in genomic regions with a high density of deleterious alleles. This is also consistent with linked selection predictions (Cutter & Payseur, 2013). I did observe

a peak in the level of polymorphism of haplotypes with a moderate number of proxies for targets of selection, *i.e.*, $\log(2.5 \text{ targets of selection}/\text{Mbp})$. My results for the Chillingham cattle support previous work suggesting that both background selection and AOD can occur simultaneously, possibly in different parts of the genome (Zhao & Charlesworth, 2016; Becher *et al.*, 2020). I observed a more pronounced and more prevalent effect of AOD in the fly experiment than in the Chillingham cattle, which might reflect the higher density of deleterious alleles and lower recombination due to the absence of recombination in male flies.

2.5.4 AOD implications for conservation and further research

I modelled AOD in the context of its effects on linked neutral genetic diversity. However, it can be argued that the effects of AOD may have a similar impact on alleles of biological and ecological importance, as long as they are contained within the region under AOD. Therefore, a better understanding of the factors triggering AOD and determining the prevalence of AOD in natural populations could provide new avenues of research in various fields.

Conservation and restoration efforts carried out to protect threatened species often monitor and set goals for maintaining genetic variation to assess the impact of these programs (Mijangos *et al.*, 2015). My simulations show that there is not a simple relationship between N_e and H_e and that in small populations, linked selection may, in fact, sometimes be the primary driver of the preservation of genetic variation. Thus to ignore this effect could have a significant and unexpected impact on conservation outcomes.

Therefore AOD could be a critical factor to consider in conservation management. AOD shows that small populations retain higher genetic diversity than predicted by neutral models in some circumstances. For captive breeding, AOD could be leveraged as a mechanism to slow the rate of loss of genetic diversity and decrease the rate of genetic divergence from remnant wild populations. However, if the intensity of AOD is not properly assessed, there is a likelihood that it could have adverse effects on vulnerable managed species. I hypothesise that introducing new individuals from other populations

in an attempt to restore genetic diversity could perversely cause heterozygosity to crash if haplotypes maintained by AOD are broken. A first step to test this hypothesis could be to perform simulations as in this study. I speculate that studies reporting a decline in fitness, genetic diversity and population growth after translocations or dispersal events could be examples of this effect (Marshall & Spalton, 2000; Marr *et al.*, 2002; Hedrick *et al.*, 2014). However, these cases might be potential exceptions rather than the norm.

My findings add to the accumulating evidence that suggests that population genetic methodologies that rely on unlinked, selectively neutral loci may be inadequate (Ellegren & Galtier, 2016; Charlesworth & Charlesworth, 2018). This impacts the full gamut of population genetic applications and could affect our ability to estimate dispersal, genetic structure, population differentiation and detection of loci under selection (Whitlock & Lotterhos, 2015). In many cases, these problems might be due to unrecognised effects of different types of linked selection, including AOD, background selection, selective sweeps (Elyashiv *et al.*, 2016) and the Hill–Robertson interference (Castellano *et al.*, 2016). It is becoming clear that there is an increasing need for the field to transition from models where neutral genetic drift is assumed to be the force driving most evolutionary processes (*i.e.*, the neutral theory of molecular evolution; Kimura, 1983) to models that incorporate and take into account the pervasive effects of selection on linked loci. This transition might be crucial to develop empirical approaches to test for and predict the combined effects of multiple evolutionary processes on genetic variation, an important challenge in population genetics (Semenov *et al.*, 2019).

Steps towards this objective are already being made, such as the development of models that explicitly use linked selection as the null model to explain genetic diversity (Gillespie, 2000; Johri *et al.*, 2020), the generation of genomic maps of linked selection (Comeron, 2014; Elyashiv *et al.*, 2016; Salvador-Martínez *et al.*, 2017), the development of statistics that allow for estimating the distance at which selection might influence other loci (Castellano *et al.*, 2016; Siewert & Voight, 2017) or the use of computer simulations to predict the possible effects of linked selection (Messer & Petrov, 2013).

While my results suggest that AOD is more widespread than anticipated, further research is needed to resolve why AOD has not been more obvious in previous studies.

Also, further work is warranted to determine the generality of my results at the species, population, chromosomal and gene levels. We could then assess the degree to which AOD is responsible for puzzling observations such as the lack of a strong correlation between levels of genetic diversity and either measure of population size: N_e and N_c (*i.e.*, Lewontin's paradox; Leffler *et al.*, 2012; Corbett-Detig *et al.*, 2015), the correlation between H_e and fitness in small populations (Szulkin *et al.*, 2010) or the high genetic variability in fitness observed in *Drosophila* that cannot be explained by current population genetics theory (Charlesworth, 2015).

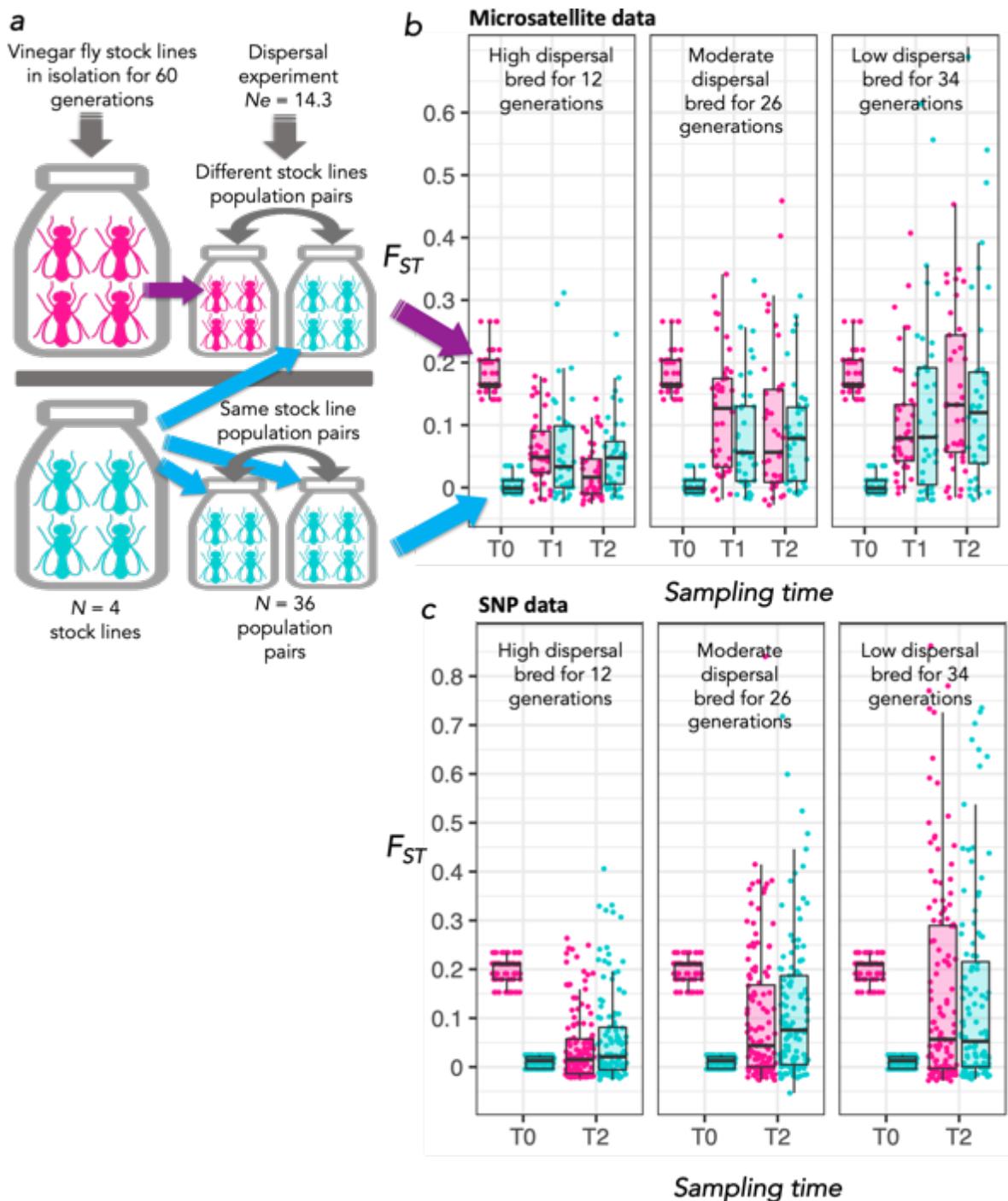
2.6 CONCLUSION

I have presented the first controlled replicated experimental results demonstrating that AOD is a phenomenon that occurs more widely in the genome of small populations than previously anticipated. This interpretation is borne out in two widely divergent organisms (cattle and flies) and in simulations that demonstrate generality. My work highlights the fact that the generation of genomic resources, such as reference genomes (Brandies *et al.*, 2019; Eldridge *et al.*, 2020), genome annotation (Yandell & Ence, 2012) and recombination maps (Adrion *et al.*, 2020; Peñalba & Wolf, 2020) are essential to conservation efforts. Particularly noteworthy is the development of novel and promising methods for constructing recombination maps that will facilitate this challenging process (Dréau *et al.*, 2019; Sun *et al.*, 2019). Coupled with my framework for predicting the occurrence and intensity of AOD, these resources will inform conservation strategies and lead to a better understanding of genomic evolution in small populations. My study turns the neutral assumptions of population genetics on their head for small populations. Compared to neutral dynamics, pervasive genome-wide effects of AOD have a significant influence on the maintenance of genetic diversity and oppose the effects of neutral genetic drift, which tend to reduce genetic diversity and increase allelic fixation over time. The impact of AOD is likely to be of considerable importance for small populations that are frequently studied in a range of fields, such as evolutionary biology, ecology, ecological restoration and biodiversity conservation.

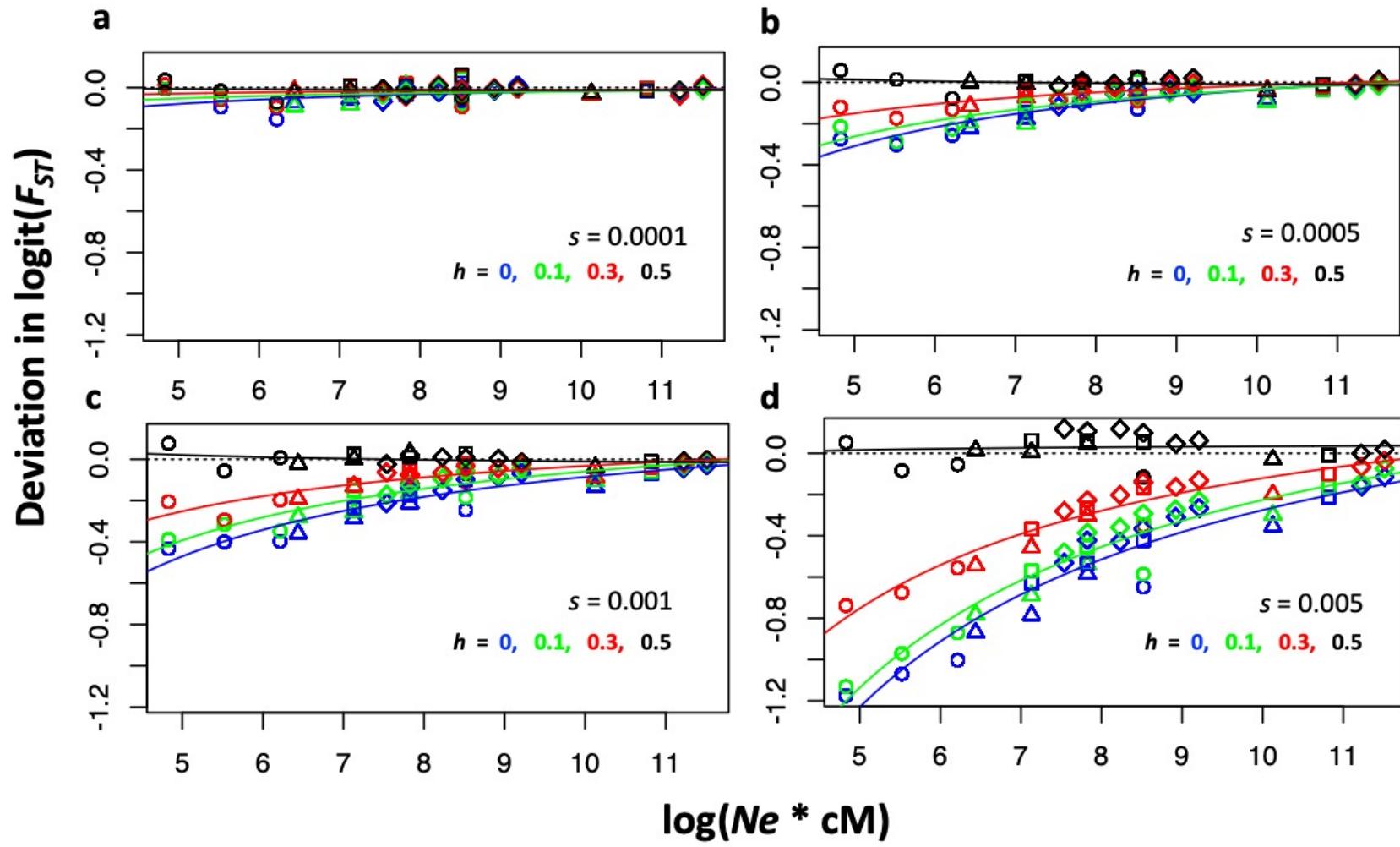
2.7 AUTHORS CONTRIBUTIONS

Bill Sherwin and Richard Nichols conceived the project and acquired the research funds for the fly experiment. Clare Holleley performed the fly experiment with support from Bill Sherwin. John Williams provided the Chillingham dataset. Luis Mijangos developed the simulation model; Simon Watt run the simulations in the National Computational Infrastructure (NCI); Luis Mijangos analysed the data; Luis Mijangos wrote the manuscript with support from Bill Sherwin, Harvinder Sidhu, Clare Holleley and Richard Nichols. Zlatko Jovanoski, Isaac Towers, Simon Watt, Bill Sherwin, Harvinder Sidhu and Clare Holleley supervised Luis Mijangos. Richard Nichols provided critical feedback in the development of the model. Aaron Adamack introduced Luis Mijangos to the development of simulations. All authors discussed the results and contributed to the final manuscript.

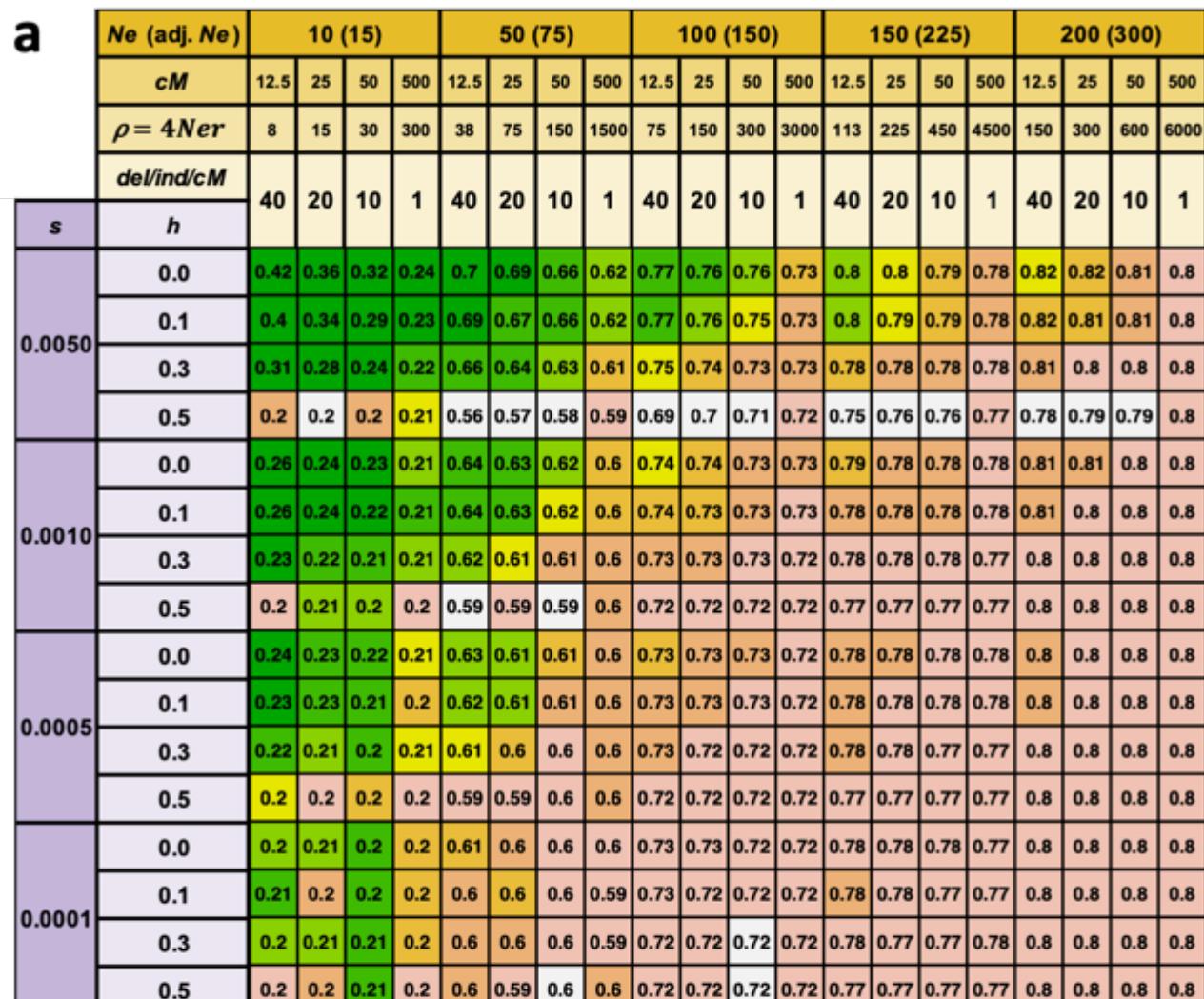
2.8 EXTENDED DATA



Extended Data Figure 2.1 | Population pairs of different demographic origin reached drift-dispersal equilibrium. **a)** Experimental design. Four stock lines of vinegar fly served as source populations for the fly experiment. There were two types of pairs of populations with different demographic origin (six population pairs formed from the same stock line and six population pairs formed from different stock lines) and three dispersal regimes (low, moderate, and high). Therefore, every combination of demographic origin and dispersal regime had six replicates, making a total of 36 population pairs or 72 populations. **b)** F_{ST} from microsatellite data. Box plots indicate median, first and third quartiles and 5th–95th percentiles of F_{ST} values at the beginning of the experiment (T0), at the expected time for populations to reach half of the drift-dispersal equilibrium (T1; Whitlock, 1992) and twice this period (T2). Each dot represents the F_{ST} value of one locus (microsatellite). **c)** F_{ST} from SNP data. Box plots indicate median, first and third quartiles and 5th–95th percentiles of F_{ST} values at the beginning of the experiment (T0) and twice the expected time for populations to reach half of the drift-dispersal equilibrium (T2). Each dot represents the F_{ST} value of one locus (single nucleotide polymorphism; SNP).

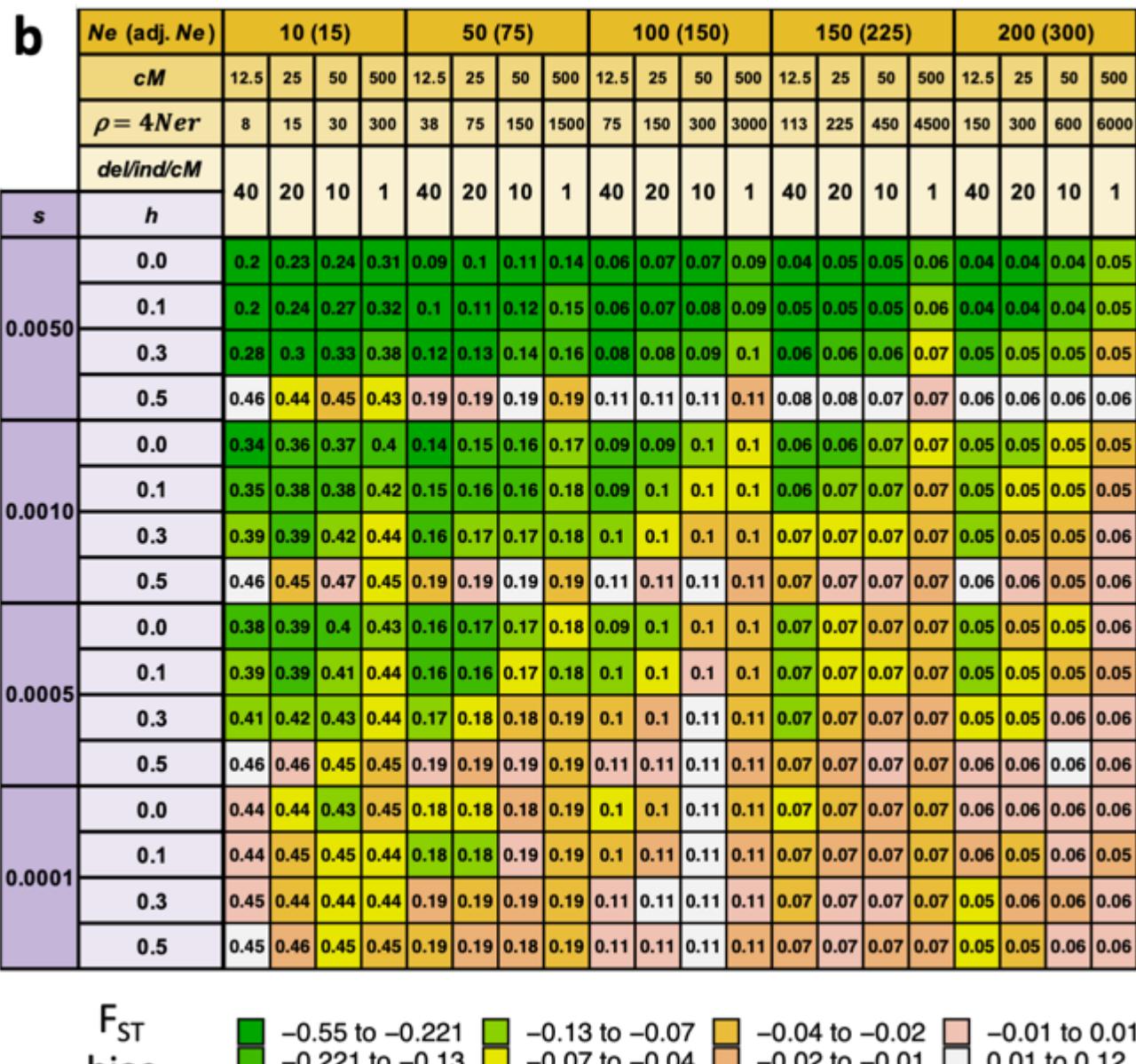


Extended Data Figure 2.2 | The effects of AOD on F_{ST} in the general simulation model. Plots show F_{ST} bias as a function of the joint effects of Ne and the total genetic distance of the chromosome (cM). Each subplot shows a different level of selection going from weak selection (top left; **a**) to strong selection (bottom right; **d**). Each symbol point in each subplot represents a separate simulation scenario. Symbol color shows the level of dominance (h ; blue = 0, green = 0.1, red = 0.25 and black = 0.5, where $h = 0.5$ is semidominant and $h = 0$ completely recessive). Symbol shape indicates length of the chromosome in centiMorgans (cM; circles = 12.5, triangles = 25, squares = 50 and rhombus = 500). Note that longer chromosomes have less del/ind/cM because other relevant variables are being held constant. In all scenarios a total of 2,000 loci with deleterious alleles evenly distributed across the genome were simulated. The initial allele frequency of each deleterious allele (q) was 0.15, resulting in that each individual carried, on average, 555 deleterious alleles (see Chapter 3 Section 3.2.4.4 for further details). In all scenarios, each individual had 50 neutral loci evenly distributed across the genome. Each neutral locus had initially nine alleles with a frequency of 1/9 or 0.111 (see Chapter 3 Section 3.2.4.9 for further details).

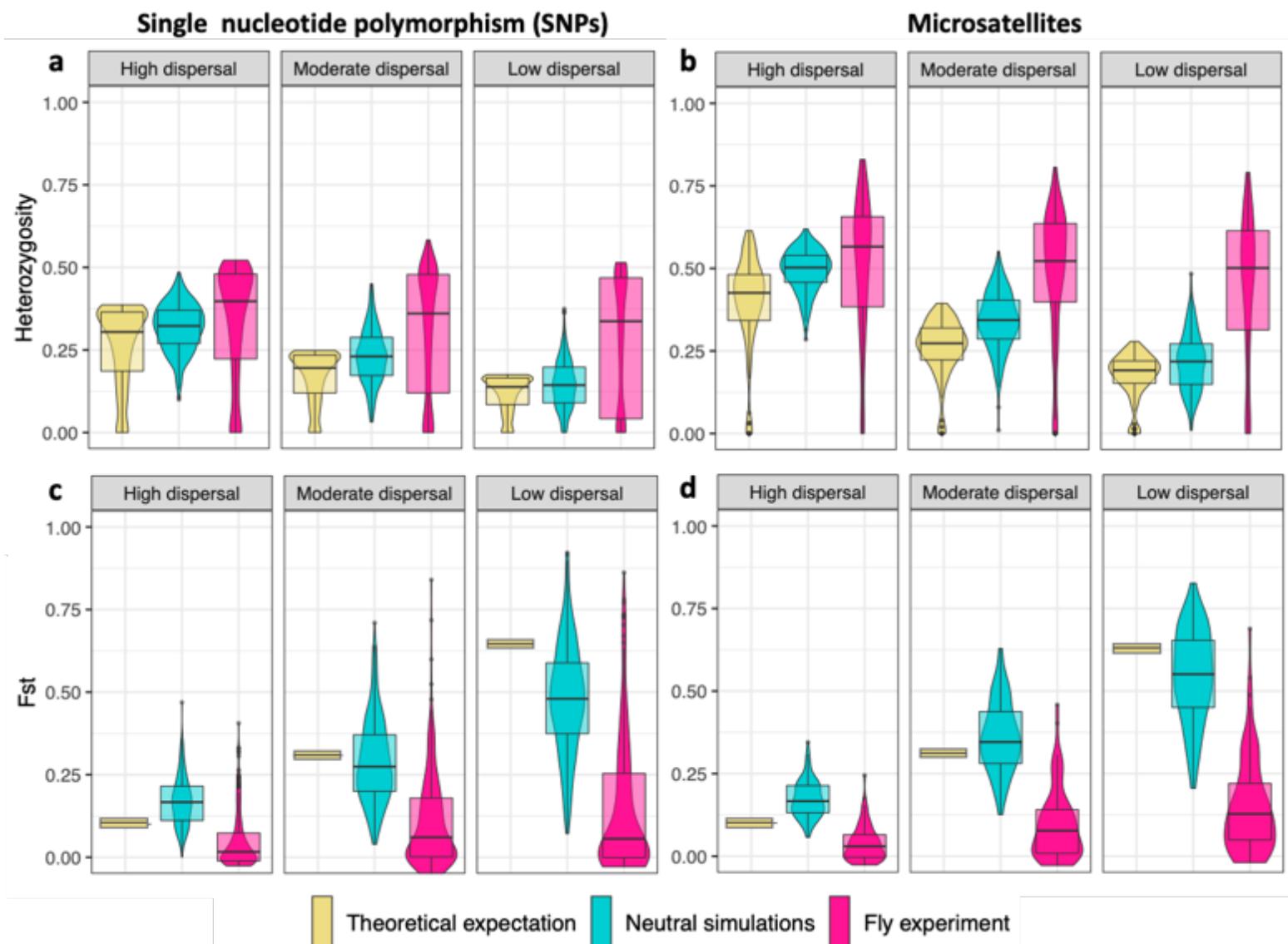


Heterozygosity bias

-0.06 to 0 0.01 0.02 to 0.02 0.03 to 0.05 0.05 to 0.14 0.14 to 1.14



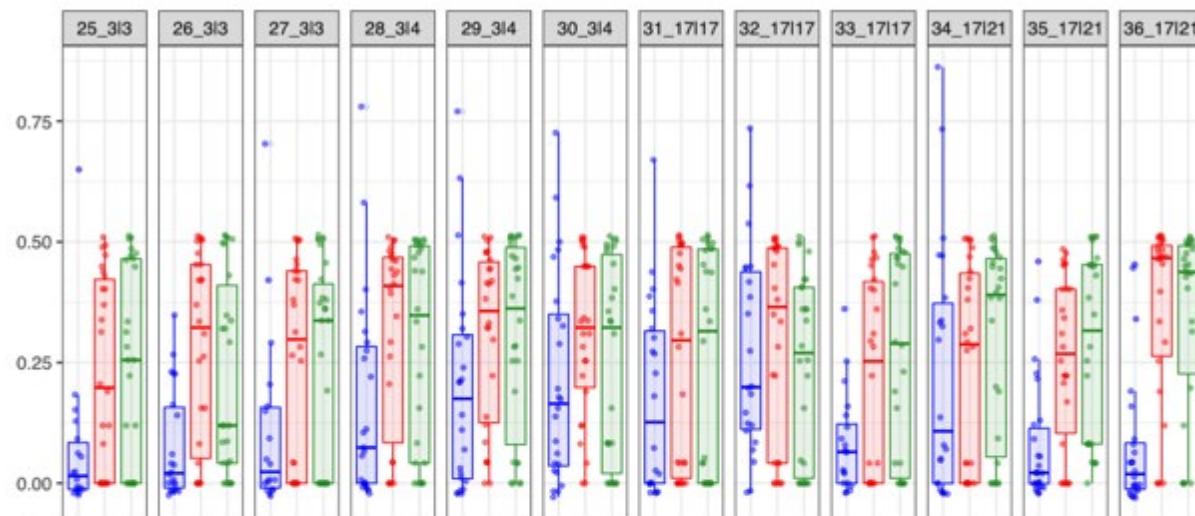
Extended Data Figure 2.3 | Simulation scenarios. Matrix where each cell is a unique combination of the main four variables that I identified as driving the strength of AOD: number of deleterious alleles per individual per centiMorgan (del/ind/cM), selection coefficient (s), dominance coefficient (h) and effective population size (Ne). The color of each cell represents the mean bias relative to neutral simulations validated against theoretical equations for **a)** heterozygosity and **b)** F_{ST} . Simulation validation for He see Chapter 3 Section 3.2.3.1 and for F_{ST} see Chapter 3 Section 3.2.4.7. The bias was calculated as: (observed value - expected value) / expected value. Numbers inside each cell are the mean values of the genetic metric (heterozygosity or F_{ST}) for 50 neutral loci tested at the end of the simulations. The first row is the Ne and corrected Ne in brackets (for each subpopulation connected by a dispersal rate of 0.01, the Ne is approximately 50% more than its Ne estimated individually; Supplementary Methods 2.9.1). The second row is the total length of the chromosome in centiMorgans (cM). The third row is an approximation of the number of recombination events per generation occurring in a population (Peñalba & Wolf, 2020) using the corrected Ne ($\rho = 4 * Ne * r$; where r is the recombination rate in units of meiosis per generation). Results are based on simulations of two populations connected by constant dispersal rate of $m = 0.01$ that reproduced for 62 generations, using 2,000 loci under selection with an initial frequency of the deleterious allele of 0.15, which results in each individual carrying 555 deleterious alleles. 50 neutral loci with nine initial alleles in equal frequencies were simulated. Simulations were run for 250 replicates for each scenario. The variable del/cM/ind depends on three variables: the initial frequency of the deleterious allele (q), the total number of loci with deleterious alleles to be simulated and the length of the chromosome in centiMorgans. Therefore, two of these three variables should be maintained constant in the simulations to isolate the effect that the del/ind/cM might have on AOD. In these simulations q ($q = 0.15$) and the total number of loci with deleterious alleles ($n = 2,000$) that were simulated were kept constant across all the combinations. Four chromosome lengths in centiMorgans were simulated: 12.5, 25, 50 and 500.



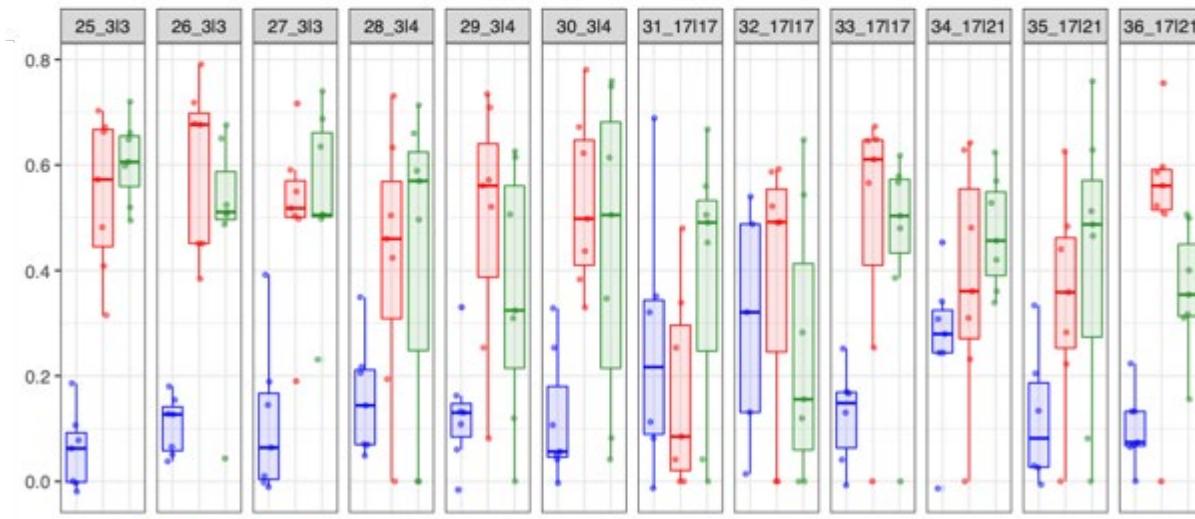
Extended Data Figure 2.4 | Genetic statistics of the fly experiment at T2 (Data from Holleley, 2009). **a)** Heterozygosity in SNPs and in **b)** microsatellites and **c)** F_{ST} in SNPs and in **d)** microsatellites values at generation T2, obtained from theoretical equations (yellow), neutral simulations (cyan) and the fly experiment (pink). Theoretical expectations for heterozygosity were calculated for each locus individually using Equation 2.2, whereas just one value of F_{ST} is calculated for each dispersal regime using Equation 2.3, which assumes drift-dispersal equilibrium. Violin plots and box plots indicate median, first and third quartiles and 5th–95th percentiles of He and F_{ST} values. Methodology used in simulations are described in Chapter 3 Section 3.3. All comparisons within dispersal regimes were significantly different by two-sided Mann–Whitney U-test.

Low dispersal regime

a Single nucleotide polymorphism (SNPs)



b Microsatellites



● F_{ST}

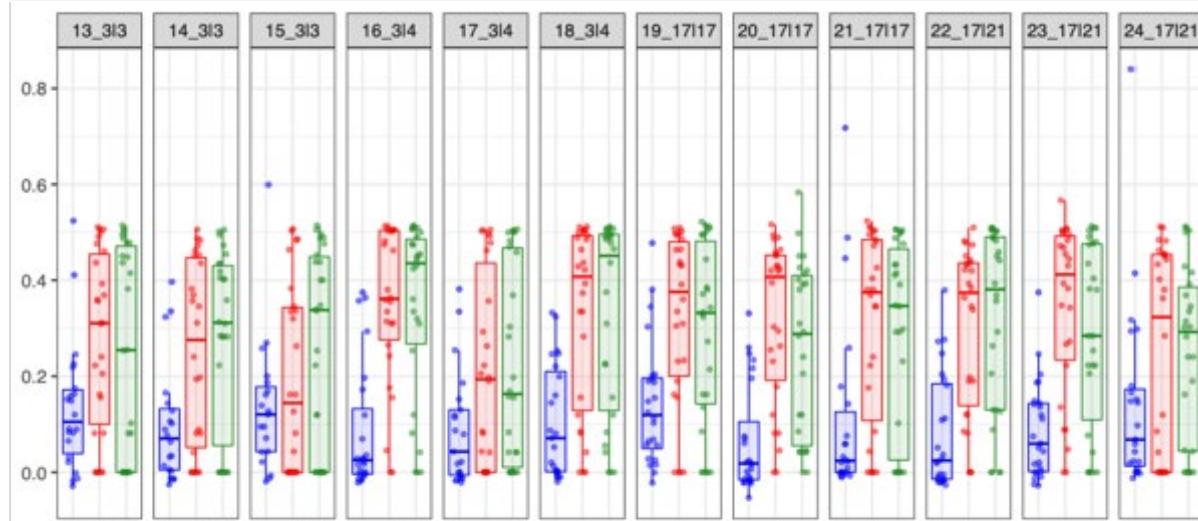
● Heterozygosity population one

● Heterozygosity population two

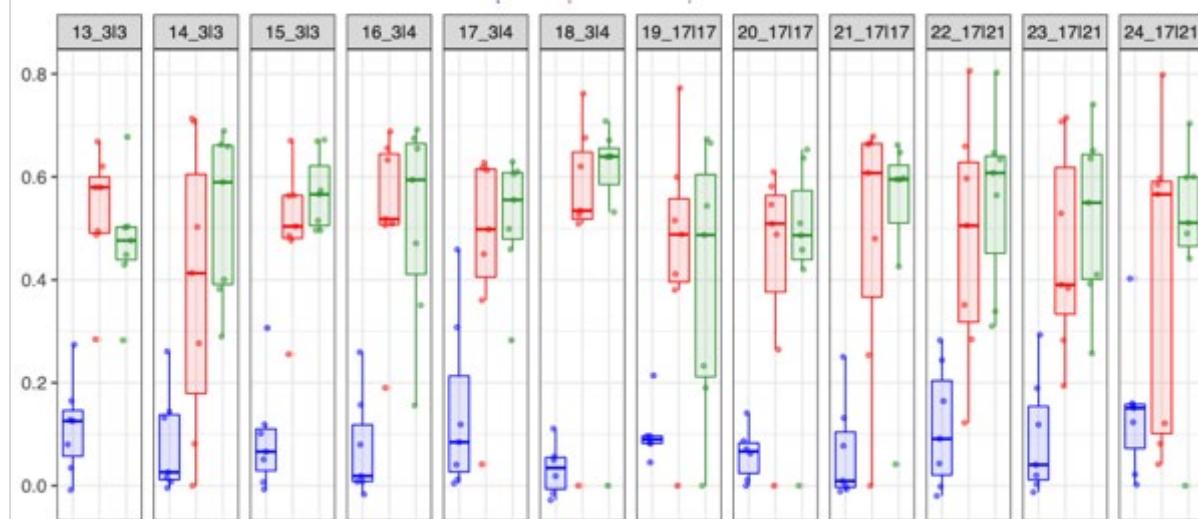
Extended Data Figure 2.5 | Low variability in genetic statistics across populations at the end of the experiment in the low dispersal regime (Data from Holleley, 2009). Each subplot shows genetic statistics in **a**) SNPs and **b**) microsatellites for each population pair ($n = 12$) of the low dispersal regime at T2 (34 generations). Header in each subplot indicates first the identification number of each population pair and then the identification number of the stock line where each subpopulation was originated from. Heterozygosity (He) in population one is shown in red, He in population two in green and F_{ST} in blue. Each point is one locus. Box plots indicate median, first and third quartiles and 5th–95th percentiles of He and F_{ST} values.

Moderate dispersal regime

a Single nucleotide polymorphism (SNPs)



b Microsatellites



● F_{ST}

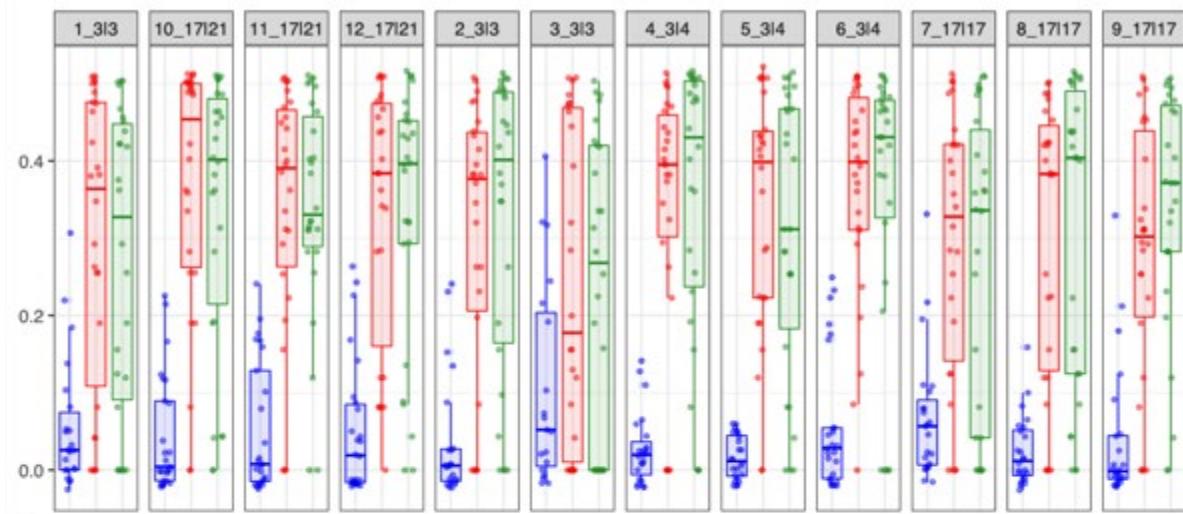
● Heterozygosity population one

● Heterozygosity population two

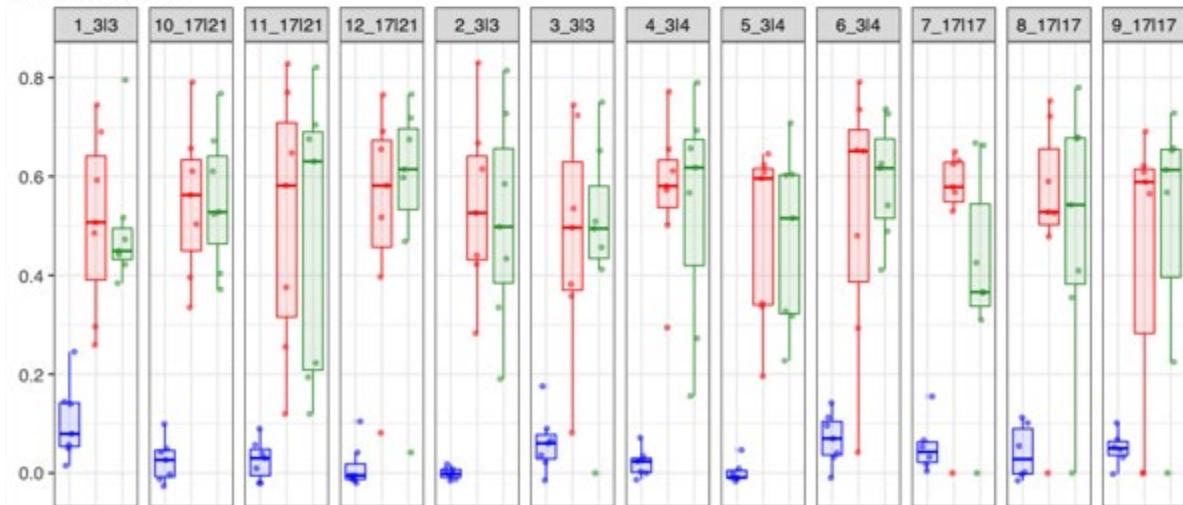
Extended Data Figure 2.6 | Low variability in genetic statistics across populations at the end of the experiment in the moderate dispersal regime (Data from Holleley, 2009). Each subplot shows genetic statistics in **a**) SNPs and **b**) microsatellites for each population pair ($n = 12$) of the moderate dispersal regime at T2 (26 generations). Header in each subplot indicates first the identification number of each population pair and then the identification number of the stock line where each subpopulation was originated from. Heterozygosity (He) in population one is shown in red, He in population two in green and F_{ST} in blue. Each point is one locus. Box plots indicate median, first and third quartiles and 5th–95th percentiles of He and F_{ST} values.

High dispersal regime

a Single nucleotide polymorphism (SNPs)



b Microsatellites

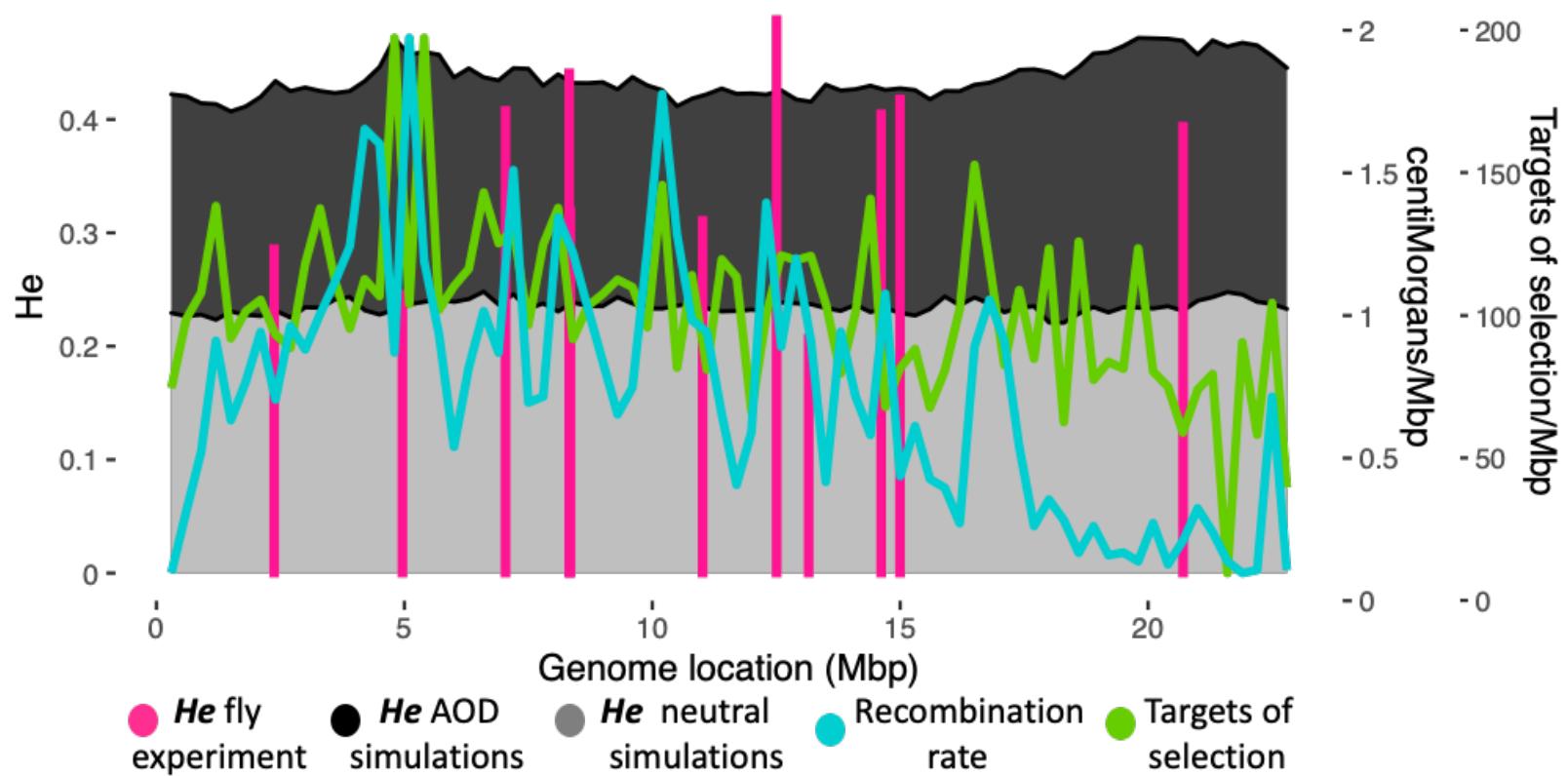


● F_{ST}

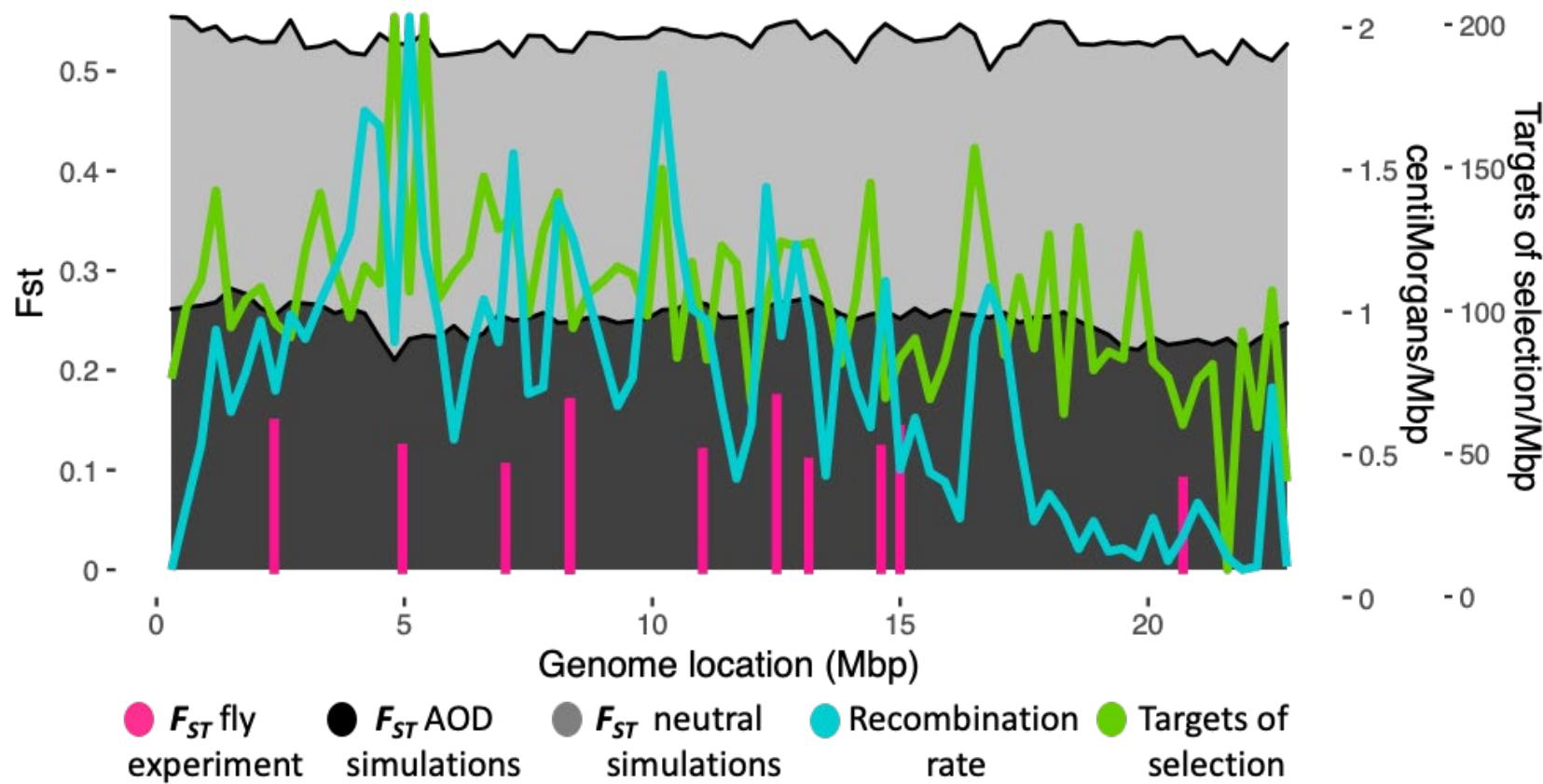
● Heterozygosity population one

● Heterozygosity population two

Extended Data Figure 2.7 | Low variability in genetic statistics across populations at the end of the experiment in the high dispersal regime (Data from Holleley, 2009). Each subplot shows genetic statistics in **a**) SNPs and **b**) microsatellites for each population pair ($n = 12$) of the high dispersal regime at T2 (12 generations). Header in each subplot indicates first the identification number of each population pair and then the identification number of the stock line where each subpopulation was originated from. Heterozygosity (He) in population one is shown in red, He in population two in green and F_{ST} in blue. Each point is one locus. Box plots indicate median, first and third quartiles and 5th–95th percentiles of He and F_{ST} values.



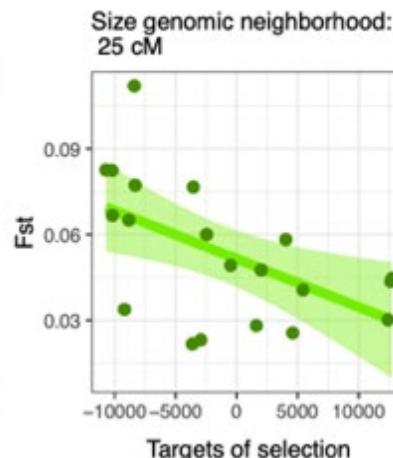
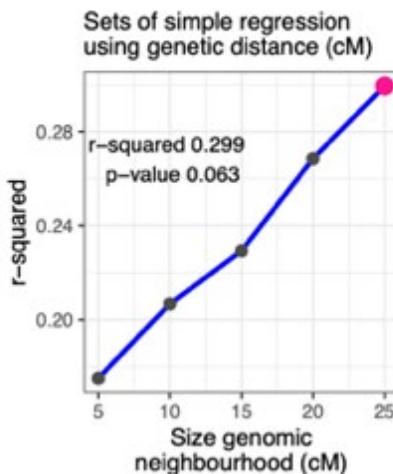
Extended Data Figure 2.8 | Fly simulations vs fly dataset: heterozygosity in loci of chromosome 2L in the low dispersal regime populations. Shaded areas show Heterozygosity (He) values across the chromosome from neutral simulations (*i.e.*, without selection; grey) and AOD simulations (*i.e.*, with selection; black). Columns in pink represent He of genotyped loci in the fly experiment. Lines represent recombination rate (cyan) and targets of selection (green) across the chromosome. Methodology used in simulations is described in Chapter 3 Section 3.3.



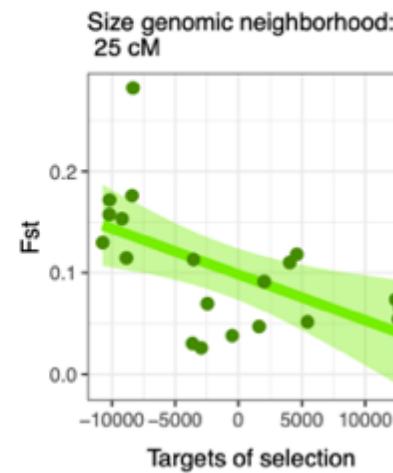
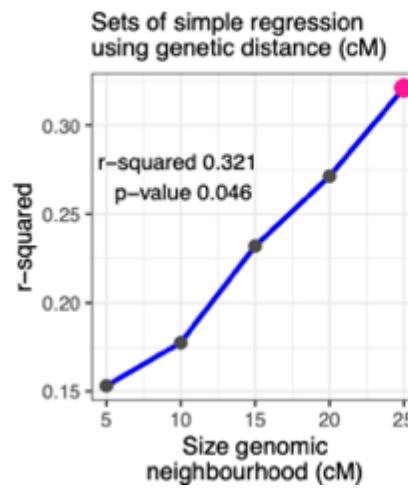
Extended Data Figure 2.9 | Fly simulations vs fly dataset: F_{ST} in loci of chromosome 2L in the low dispersal regime populations. Shaded areas show F_{ST} values across the chromosome from neutral simulations (*i.e.*, without selection; grey) and AOD simulations (*i.e.*, with selection; black). Columns in pink represent genotyped loci in the fly experiment. Lines represent recombination rate (cyan) and targets of selection (green) across the chromosome. Methodology used in simulations is described in Chapter 3 section 3.3.

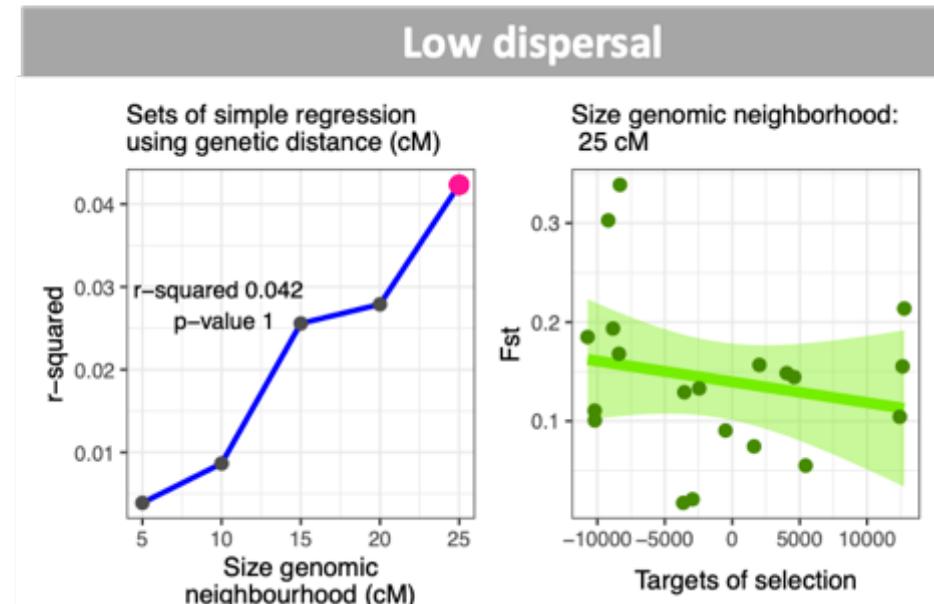
a

High dispersal

**b**

Moderate dispersal

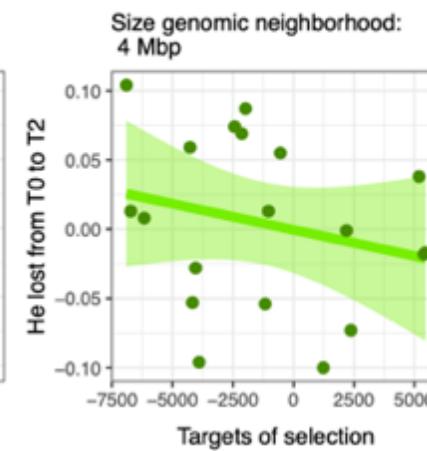
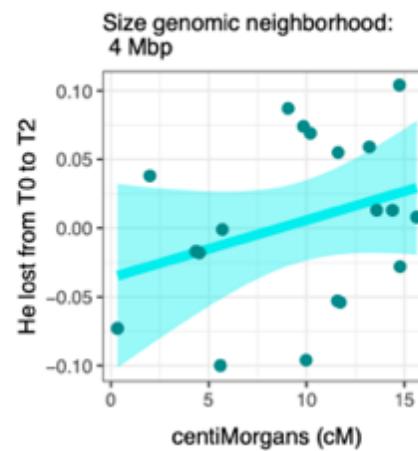
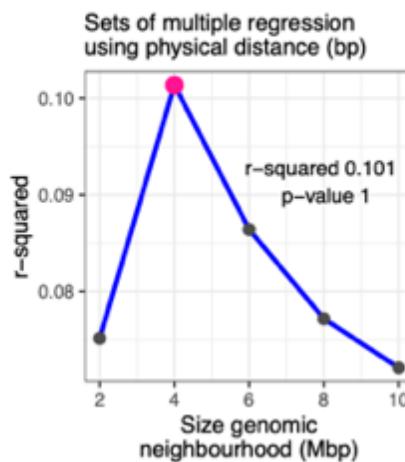


C

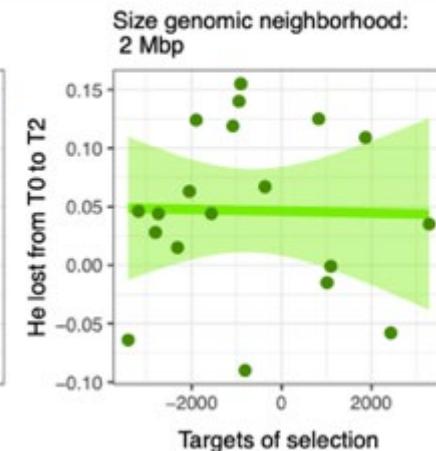
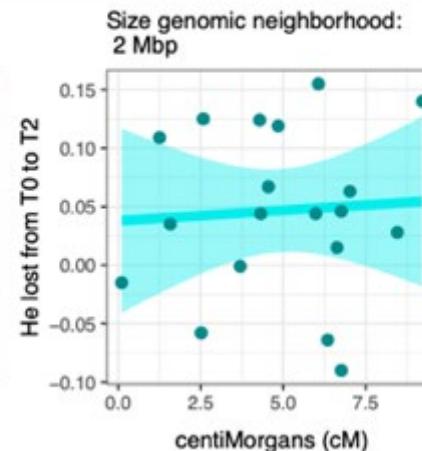
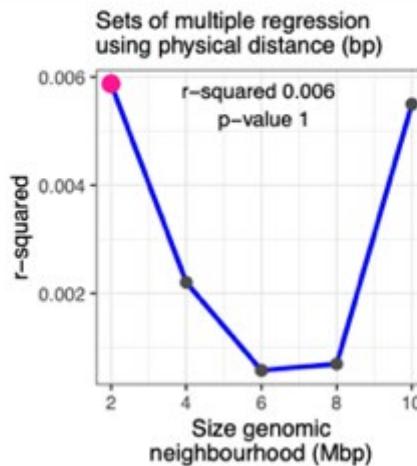
Extended Data Figure 2.10 | Relationship between F_{ST} and proxies for targets of selection in the fly experiment using genetic distance (i.e., centiMorgans; cM). Each row presents the results for each dispersal regime: **a)** high dispersal; **b)** moderate dispersal, and; **c)** low dispersal. The first column is the inference of the size of the genomic neighbourhood using genetic distance, where each point is the R^2 of each simple regression (F_{ST} on proxies for targets of selection), and the pink point shows the regression with the highest R^2 value. The second column is the regression line between F_{ST} and proxies for targets of selection within the inferred genomic neighbourhood. Each point in the second column is the F_{ST} of one locus averaged across the twelve population pairs of each dispersal regime. Shaded areas are 95% confidence intervals of the regression line. The units of targets of selection is the difference between the number of non-synonymous (NS) and synonymous (S) mutations (i.e., NS - S) in a given gene. A negative result indicates a lower number of targets of selection, and a positive result indicates a higher number of targets of selection, as explained in the methods section. A genomic neighbourhood > 25 cM was not trialed because the physical distance (cM) between the extreme of the chromosome and the first and last loci was > 25 cM.

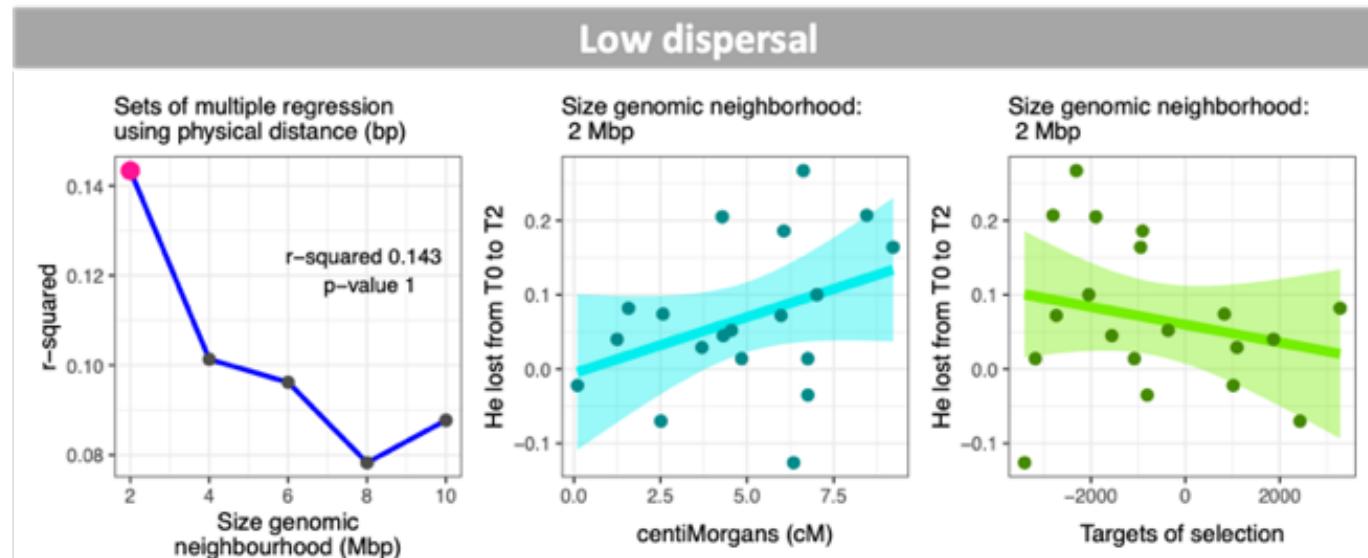
a

High dispersal

**b**

Moderate dispersal

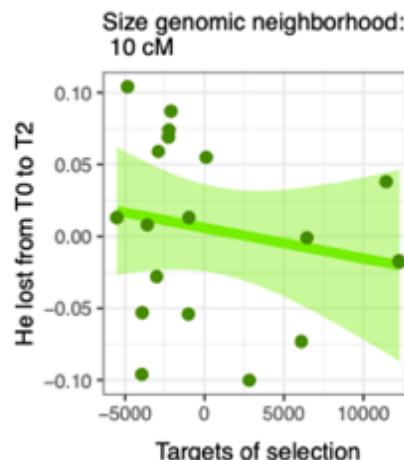
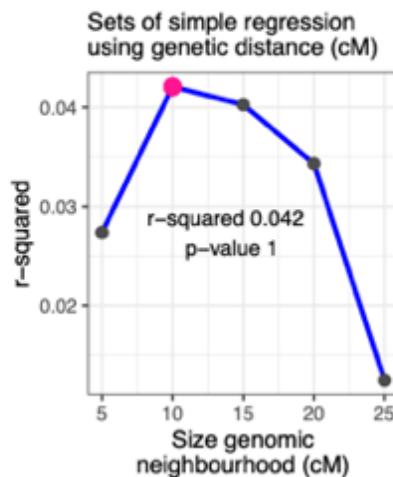


C

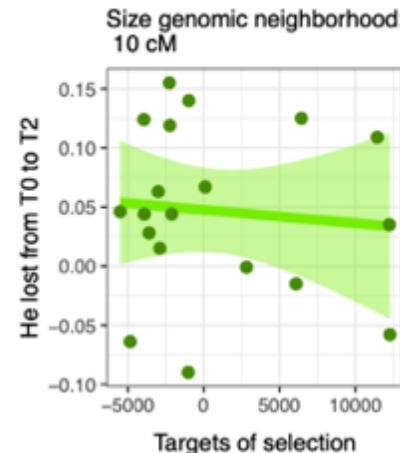
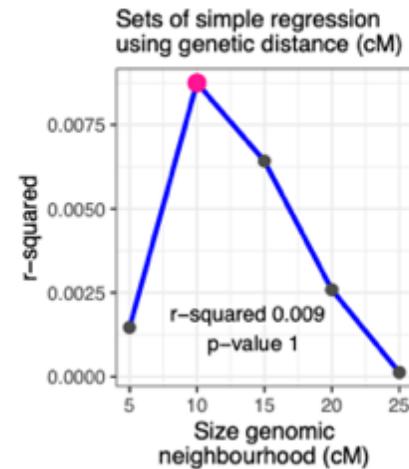
Extended Data Figure 2.11 | Relationship between the loss of He and recombination and proxies for targets of selection in the fly experiment using physical distance (bp). Each row presents the results for each dispersal regime: **a**) high dispersal; **b**) moderate dispersal, and; **c**) low dispersal. The first column is the inference of the size of the genomic neighbourhood using physical distance, where each point is the R^2 of each multiple regression (F_{ST} on recombination [cM] and proxies for targets of selection), and the pink point shows the regression with the highest R^2 value. The second column is the regression line between the loss of He and the recombination rate within the inferred genomic neighborhood. The third column is the regression line between the loss of He and proxies for targets of selection within the inferred genomic neighbourhood. Each point in the second and third columns is the loss of He of one locus averaged across the twelve population pairs of each dispersal regime. Shaded areas are 95% confidence intervals of the regression line. The units of targets of selection is the difference between the number of non-synonymous (NS) and synonymous (S) mutations (*i.e.*, NS - S) in a given gene. A negative result indicates a lower number of targets of selection, and a positive result indicates a higher number of targets of selection, as explained in the methods section.

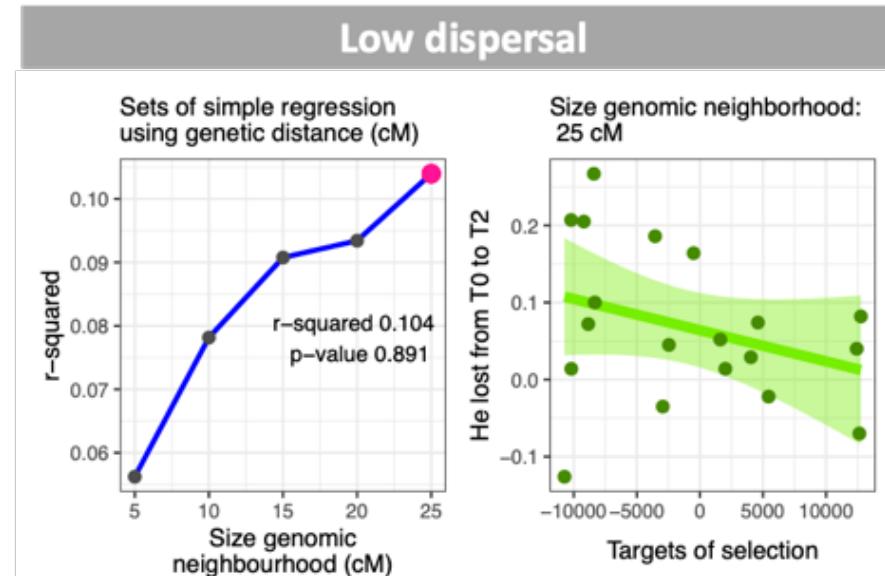
a

High dispersal

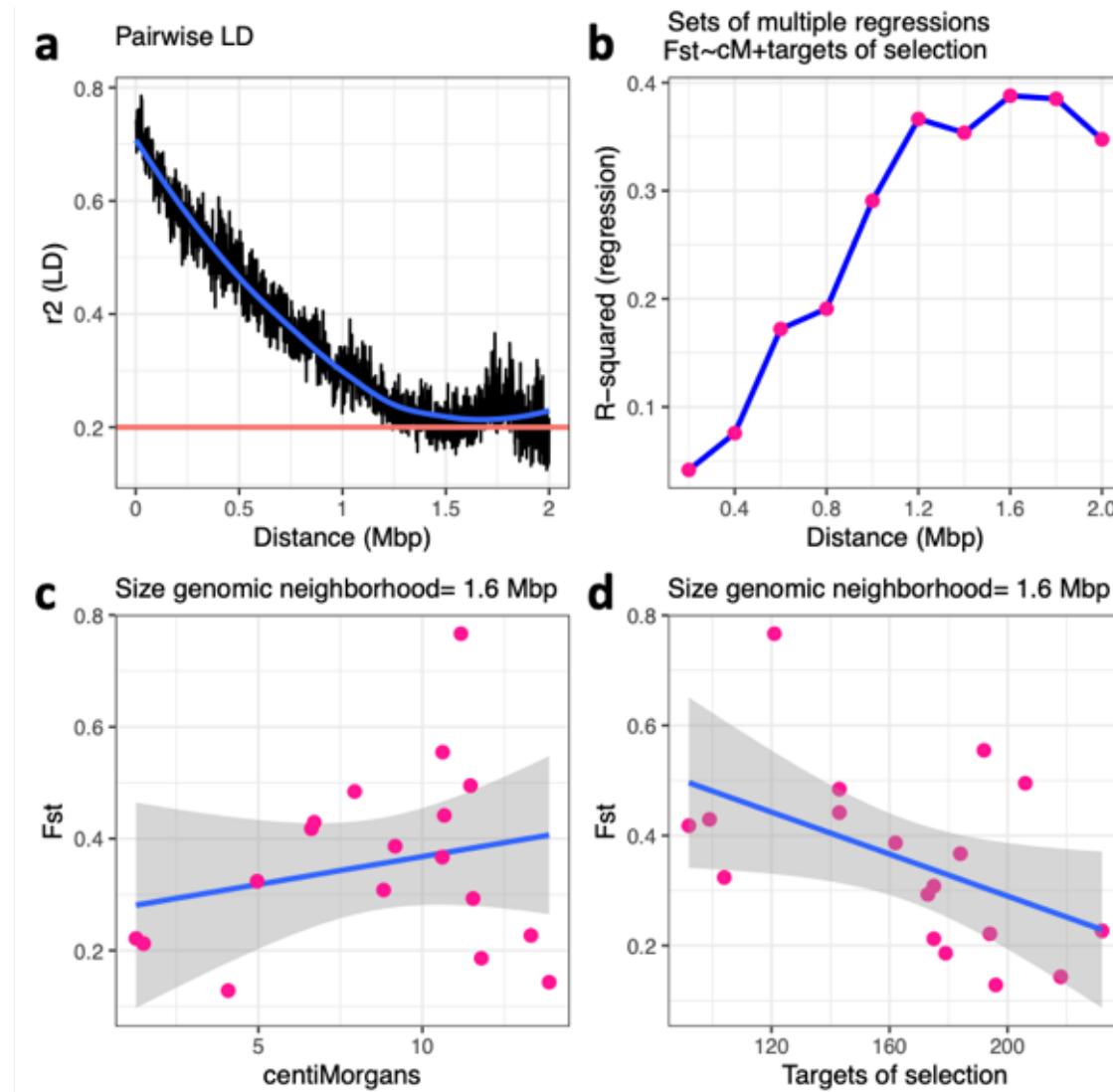
**b**

Moderate dispersal

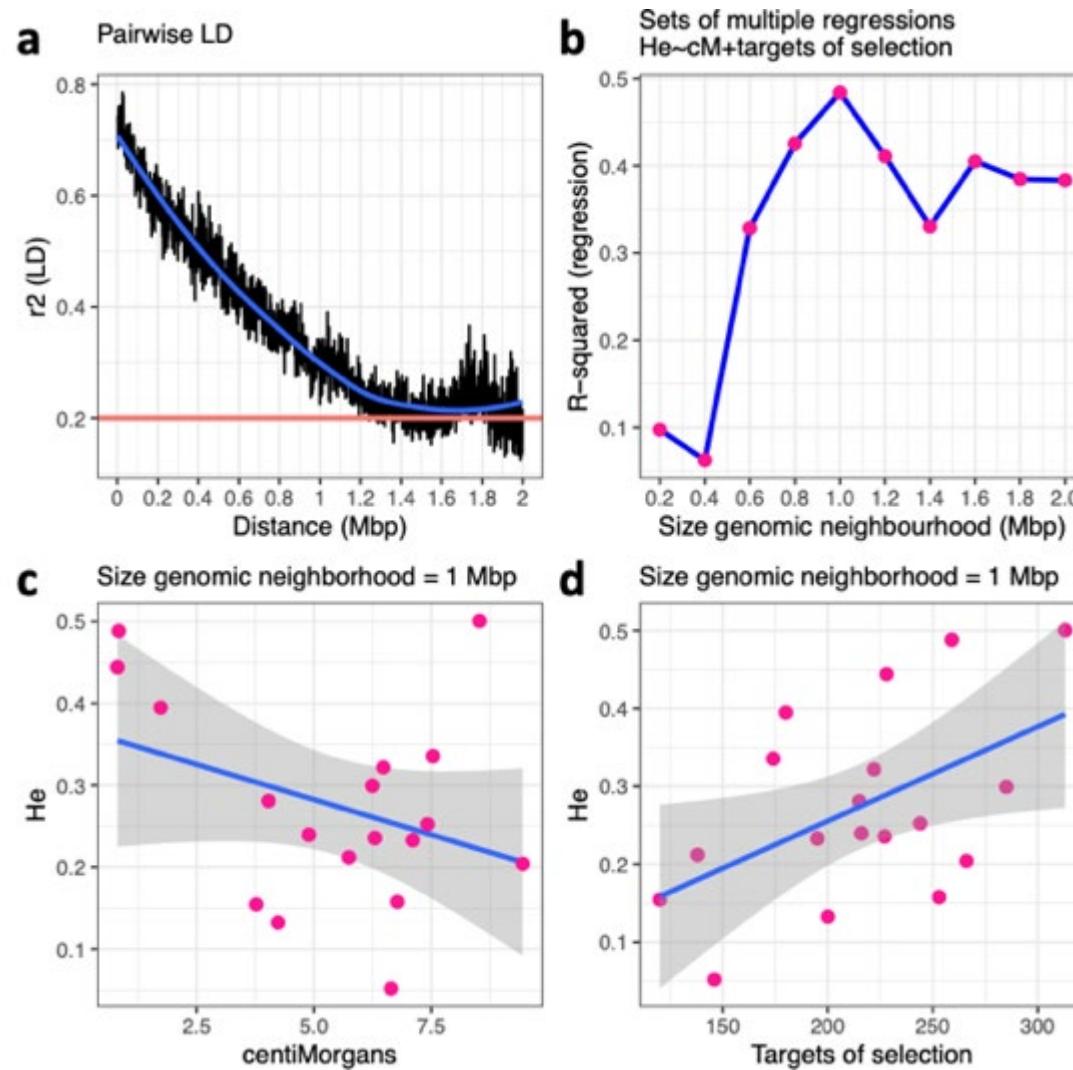


C

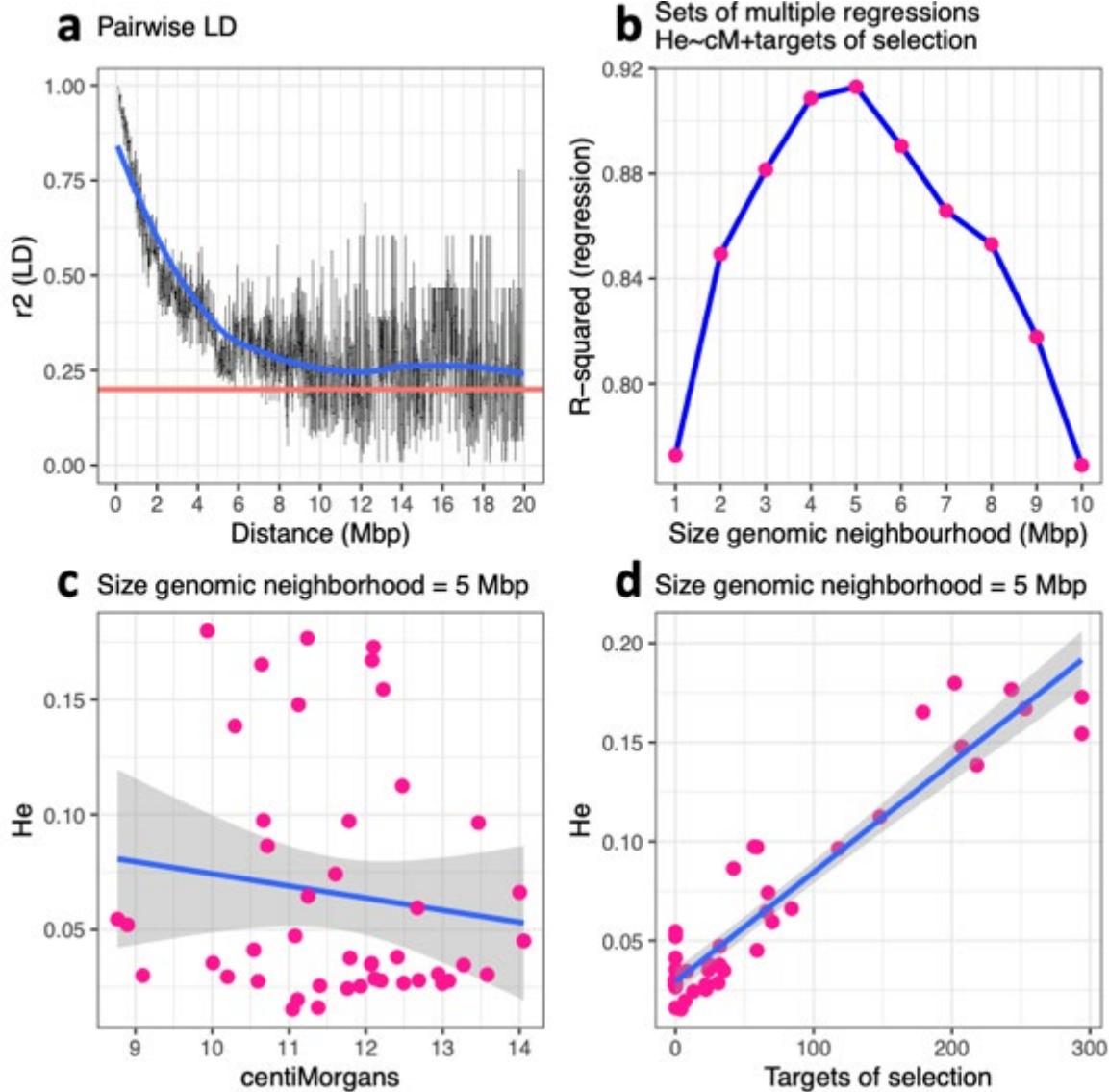
Extended Data Figure 2.12 | Relationship between the loss of H_e and proxies for targets of selection in the fly experiment using genetic distance (i.e., centiMorgans; cM). Each row presents the results for each dispersal regime: **a)** high dispersal; **b)** moderate dispersal, and; **c)** low dispersal. The first column is the inference of the size of the genomic neighbourhood using genetic distance, where each point is the R^2 of each simple regression (the loss of H_e on proxies for targets of selection), and the pink point shows the regression with the highest R^2 value. The second column is the regression line between the loss of H_e and proxies for targets of selection within the inferred genomic neighbourhood. Each point in the second column is the loss of H_e of one locus averaged across the twelve population pairs of each dispersal regime. Shaded areas are 95% confidence intervals of the regression line. The units of targets of selection is the difference between the number of non-synonymous (NS) and synonymous (S) mutations (i.e., NS - S) in a given gene. A negative result indicates a lower number of targets of selection, and a positive result indicates a higher number of targets of selection, as explained in the methods section.



Extended Data Figure 2.13 | Relationship between F_{ST} and recombination and targets of selection (i.e., deleterious alleles) in the fly simulations using physical distance (bp). **a)** Decay of pairwise linkage disequilibrium as a function of physical distance, using the statistic r^2 (Hill & Robertson, 1966). Horizontal red line indicates the LD threshold ($r^2 = 0.2$) which is commonly used to imply that two loci are unlinked (Delourme *et al.*, 2013; Li *et al.*, 2014). **b)** Inference of the size of the genomic neighbourhood using physical distance, where each point is the R^2 of each multiple regression (F_{ST} on recombination [cM] and targets of selection *i.e.*, deleterious alleles). The distance at which the regression signal reaches its peak of significance is considered the size of the genomic neighborhood. Note that this distance coincides with the distance at which pairwise LD (r^2) is < 0.2 . **c)** Regression line between F_{ST} and the recombination rate within the inferred genomic neighborhood. **d)** Regression line between F_{ST} and targets of selection *i.e.*, deleterious alleles, within the inferred genomic neighbourhood. Each point in **c** and **d** is the F_{ST} of one neutral locus of one simulation replicate. Shaded areas are 95% confidence intervals of the regression line. The units of targets of selection are the actual number of simulated deleterious alleles. Methodology used in simulations are described in Chapter 3 Section 3.3.



Extended Data Figure 2.14 | Relationship between heterozygosity (H_e) and recombination and targets of selection (i.e., deleterious alleles) in the fly simulations using physical distance (bp). **a)** Decay of pairwise linkage disequilibrium as a function of physical distance, using the statistic r^2 (Hill & Robertson, 1966). Horizontal red line indicates the LD threshold ($r^2 = 0.2$) that is commonly used to imply that two loci are unlinked (Delourme *et al.*, 2013; Li *et al.*, 2014). **b)** Inference of the size of the genomic neighbourhood using physical distance, where each point is the R^2 of each multiple regression (H_e on recombination [cM] and targets of selection *i.e.*, deleterious alleles). The distance at which the regression signal reaches its peak of significance is considered the size of the genomic neighborhood. Note that this distance coincides with the distance at which pairwise LD (r^2) is < 0.2 . **c)** Regression line between H_e and the recombination rate within the inferred genomic neighborhood. **d)** Regression line between H_e and targets of selection *i.e.*, deleterious alleles, within the inferred genomic neighbourhood. Each point in **c** and **d** is the H_e of one neutral locus of one simulation replicate. Shaded areas are 95% confidence intervals of the regression line. The units of targets of selection are the actual number of simulated deleterious alleles. Methodology used in simulations are described in Chapter 3 Section 3.3.



Extended Data Figure 2.15 | Relationship between heterozygosity (H_e) and recombination and targets of selection (i.e., deleterious alleles) in the Chillingham simulations using physical distance (bp). **a)** Decay of pairwise linkage disequilibrium as a function of physical distance, using the statistic r^2 (Hill & Robertson, 1966). Horizontal red line indicates the LD threshold ($r^2 = 0.2$) that is commonly used to imply that two loci are unlinked (Delourme *et al.*, 2013; Li *et al.*, 2014). **b)** Inference of the size of the genomic neighbourhood using physical distance, where each point is the R^2 of each multiple regression (H_e on recombination [cM] and targets of selection i.e., deleterious alleles). The distance at which the regression signal reaches its peak of significance is considered the size of the genomic neighborhood. Note that this distance coincides with the distance at which pairwise LD (r^2) is < 0.2 . **c)** Regression line between H_e and the recombination rate within the inferred genomic neighborhood. **d)** Regression line between H_e and targets of selection i.e., deleterious alleles, within the inferred genomic neighbourhood. Each point in **c** and **d** is the H_e of one neutral locus of one simulation replicate. Shaded areas are 95% confidence intervals of the regression line. The units of targets of selection are the actual number of simulated deleterious alleles. Methodology used in simulations are described in Chapter 3 Section 3.4.

2.9 SUPPLEMENTARY METHODS

2.9.1 Inference of Ne under high dispersal

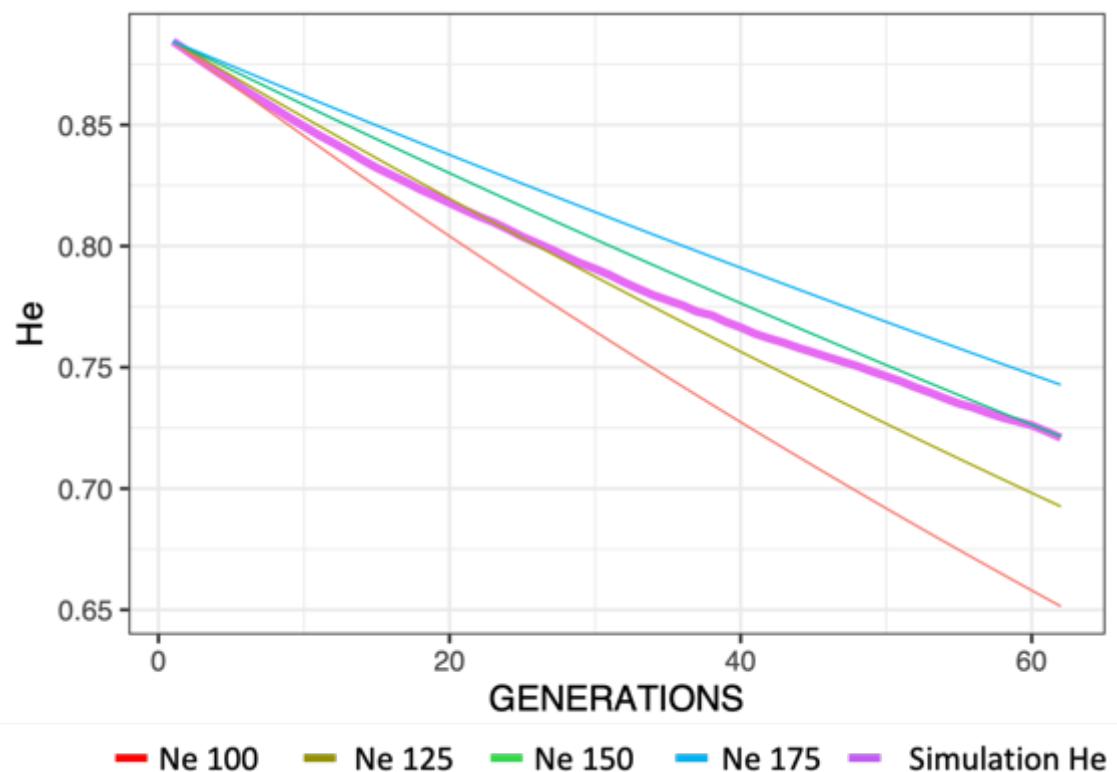
High dispersal rates (m) between populations can substantially affect effective population size (Ne) estimates (Wang & Whitlock, 2003). Therefore, to infer the corrected Ne for the population pairs that were simulated in the general simulations, I first ran neutral simulations (250 replicates) using the parameters that were used in the general simulations: dispersal rate ($m = 0.01$), populations reproduced for 62 generations and a ratio $Ne/Nc = 1$ (see Chapter 3, Section 3.2.3.1 for details about how this ratio is implemented in the simulations). Then, to infer the corrected Ne from populations connected by dispersal, I used the equation of the expected rate of loss of heterozygosity (He) under neutrality (Crow & Kimura, 1970, p. 329) as baseline to investigate the amount of overestimation due to dispersal calculated as:

$$He_t = He_0 \left(1 - \frac{1}{2Ne}\right)^t, \quad \text{Supplementary Methods Equation 2.1}$$

where He_0 is heterozygosity at generation 0 and t is the number of generations. This equation was then solved for each locus of the simulations and using various values of Ne . Finally, I plotted:

- The mean heterozygosity across all the loci and all replicates in each generation of the neutral simulations, and;
- The mean results of Supplementary Methods Equation 2.1 using different Ne values in each generation.

I considered as the corrected Ne , the Ne value used in the equation that approximated the closest the trajectory of heterozygosity across generations of the neutral simulations (Supplementary Methods Figure 2.1).



Supplementary Methods Figure 2.1 | Inference of the corrected effective population size (Ne) under dispersal. Plot showing the trajectory of the mean heterozygosity (He) over generations in neutral simulations using a dispersal rate of 0.01 and an $Ne = 100$ (bold purple line). Thin lines show the expected rate of loss of heterozygosity under neutrality (using Supplementary Methods Equation 2.1) using as input different values of Ne . The Ne that approximates most closely the trajectory of the neutral simulations was considered to be the corrected Ne ; in the above case the corrected Ne is 150 (green line).

2.9.2 Patterns of non-synonymous and synonymous mutations

The results below confirm that patterns of the density of non-synonymous mutations (NS) in *D. melanogaster* across the genome are unlikely to significantly differ between the fly experimental populations and the Ensembl database.

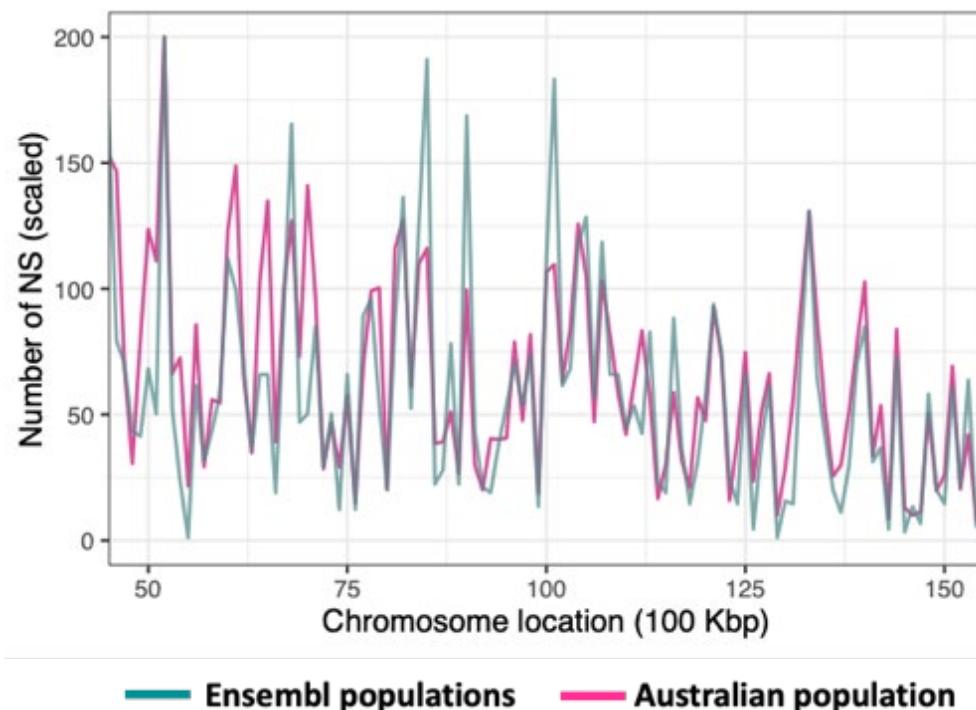
Throughout my analyses and simulations, I used the number of non-synonymous mutations (NS) as a proxy for the selection intensity that any given genomic region might experience (see main text). To obtain the number of NS , I accessed the Ensembl Fruitfly *Drosophila melanogaster* variation database (Zerbino *et al.*, 2017), using assembly

BDGP6, and extracted the location of all *NS* contained in this database. This database is sourced from *FlyBase* (Thurmond *et al.*, 2018) which contains over 6.7 million SNPs from two populations from North Carolina (Jordan *et al.*, 2007) and Malawi (Begun & Lindfors, 2005).

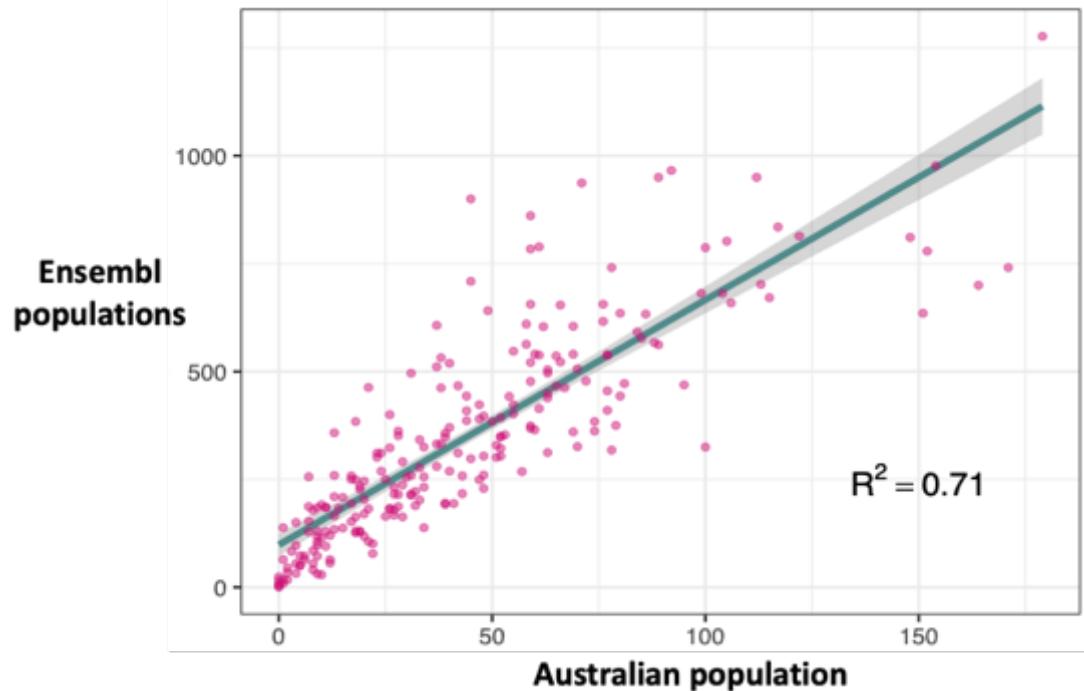
To investigate whether the pattern of the number of *NS* across the genome did not significantly differ between the experimental populations and the USA and African populations from the Ensembl database, I analysed raw sequence data from an Australian *D. melanogaster* population collected at Templestowe, Melbourne, Australia (Griffin *et al.*, 2017). This population is ~1,000 Km apart from the source population used in this study.

For this analysis, I used the chromosome arm 2L as the reference model. I determined the location of *NS* from the sub-sample 4, called MB4 in Griffin *et al.* (2017), which contains sequence reads from a pool of 25 female and 25 male fly individuals. The pipeline used for the bioinformatic data processing (using macOS) to obtain the location of the *NS* from this sample is described in the following section (2.9.2.1).

I identified 10,100 *NS* within the chromosome arm 2L of the Australian population, while the Ensembl database contains 80,046 *NS* in the same chromosome. To visually examine the pattern of the number of *NS* across the chromosome, I first calculated the number of *NS* in windows of 100 Kbp in both populations. I then transformed to the same scale (from 0 to 200) the number of *NS* in each window of 100 Kbp in both populations (Supplementary Methods Figure 2.2). I also performed a simple regression analysis using the number of *NS* in windows of 100 Kbp of both populations ($R^2 = 0.71$, $p\text{-value} < 0.05$; Supplementary Methods Figure 2.3). I found that the pattern of the density of the *NS* across the chromosome from the Australian population and the FlyBase populations is largely concordant. This agreement between three continents (North America, Africa and Australia) makes me confident that the same pattern occurs in the experimental fly populations.



Supplementary Methods Figure 2.2 | Visualisation of the (scaled) number of non-synonymous mutations in windows of 100 Kbp in the chromosome arm 2L of the Australian population (pink line) and the Ensembl populations (green line). This figure shows that the pattern of the number of non-synonymous mutations in both populations is largely concordant.



Supplementary Methods Figure 2.3 | Scatter plot of the number of non-synonymous mutations in the chromosome 2L in windows of 100 Kbp of the Australian population against the Ensembl populations. The regression line ($R^2 = 0.71$) is shown in green and its 95% confident interval in grey. This figure shows that the pattern of the number of non-synonymous mutations in both populations is largely concordant.

2.9.2.1 Bioinformatic data processing pipeline to obtain the location of the *NS* in the Australian population

1. Download and install the following programs:
 - 1.1. Trimmomatic (Bolger *et al.*, 2014)
 - 1.2. bwa - Burrows-Wheeler Alignment Tool (Li & Durbin, 2009)
 - 1.3. Samtools (Li *et al.*, 2009)
 - 1.4. Picard Toolkit (Broad Institute, 2019)
 - 1.5. RepeatMasker (Smit *et al.*, 2015)
 - 1.6. Genome Analysis Toolkit (McKenna *et al.*, 2010)
 - 1.7. Htslib (Li *et al.*, 2009)
 - 1.8. Variant Effect Predictor (McLaren *et al.*, 2016)
2. Download the *Drosophila melanogaster* reference genome (dmel-all-chromosome-r6.20.fasta.gz) from FlyBase (Thurmond *et al.*, 2018).

- 2.1. Duplicate file and unzipped one (both files are used in the pipeline).
- 3. Download from The European Nucleotide Archive (ENA) the two files of the paired reads (forward and reverse reads) of the MB4 experiment (run SRR3048953) used in Griffin *et al.*, (2017) using the Galaxy platform.
 - 3.1. Rename files as SRR3048953_1.fastq and SRR3048953_2.fastq
 - 3.2. Compress files using terminal
 - 3.2.1. gzip ~/SRR3048953_1.fastq
 - 3.2.2. gzip ~/SRR3048953_2.fastq
 - 3.2.3. output: SRR3048953_1.fastq.gz and SRR3048953_2.fastq.gz
- 4. Trimming Illumina (FASTQ) data and removing adapters using Trimmomatic.
 - 4.1. cd trimmomatic
 - 4.2. java -jar trimmomatic-0.39.jar PE -phred33 ~/SRR3048953_1.fastq.gz
 ~/SRR3048953_2.fastq.gz ~/output_forward_paired.fq.gz
 ~/output_forward_unpaired.fq.gz ~/output_reverse_paired.fq.gz
 output_reverse_unpaired.fq.gz ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10
 LEADING:30 TRAILING:30 SLIDINGWINDOW:10:25 MINLEN:40
 - 4.3. output: output_forward_paired.fq.gz and output_reverse_paired.fq.gz
- 5. Indexing reference genome using bwa:
 - 5.1. cd bwa
 - 5.2. ./bwa index -a bwtsw ~/dmel-all-chromosome-r6.20.fasta
 - 5.3. Output: dmel-all-chromosome-r6.20.fasta.gz.sa, dmel-all-chromosome-r6.20.fasta.gz.bwt, dmel-all-chromosome-r6.20.fasta.gz.amb, dmel-all-chromosome-r6.20.fasta.gz.ann and dmel-all-chromosome-r6.20.fasta.gz.pac
- 6. Creating the FASTA index file using Samtools
 - 6.1. cd samtools
 - 6.2. ./samtools faidx ~/dmel-all-chromosome-r6.20.fasta
 - 6.3. Output: dmel-all-chromosome-r6.20.fasta.fai
- 7. Generate the sequence dictionary using Picard tools
 - 7.1. cd picard
 - 7.2. java -jar picard.jar CreateSequenceDictionary R= ~/dmel-all-chromosome-r6.20.fasta.gz O=dmel-all-chromosome-r6.20.dict
 - 7.3. ouput: dmel-all-chromosome-r6.20.dict

8. Align to reference genome using bwa
 - 8.1. cd bwa
 - 8.2. ./bwa mem -t 4 ~/dmel-all-chromosome-r6.20.fasta.gz
~/output_forward_paired.fq.gz ~/output_reverse_paired.fq.gz >
aligned_pairs.sam
 - 8.3. Output: aligned_pairs.sam
9. Convert SAM file to BAM file using Samtools
 - 9.1. cd samtools
 - 9.2. ./samtools view -b ~/aligned_pairs.sam > aligned_pairs.bam
 - 9.3. Output: aligned_pairs.bam
10. Sort BAM file by coordinate using Picard tools
 - 10.1. cd picard
 - 10.2. java -jar picard.jar SortSam VALIDATION_STRINGENCY=LENIENT
I=~/aligned_pairs.bam O= output_sort.bam SORT_ORDER=coordinate
 - 10.3. Output: output_sort.bam
11. Remove duplicated reads using Picard tools
 - 11.1. cd picard
 - 11.2. java -jar picard.jar MarkDuplicates REMOVE_DUPLICATES=true
VALIDATION_STRINGENCY=LENIENT AS= true I= ~/output_sort.bam O=
output_duplicates.bam M=marked_dup_metrics.txt
 - 11.3. Output: output_duplicates.bam
12. Replace all read groups with a single new read group and assign all reads to this read group with Picard tools
 - 12.1. cd picard
 - 12.2. java -jar picard.jar AddOrReplaceReadGroups
VALIDATION_STRINGENCY=LENIENT I= ~/output_duplicates.bam
O=addOrReplace_output.bam RGLB=LaneX RGPL=illumina RGPU=None RGSM=
SRR3048953
 - 12.3. Output: addOrReplace_output.bam
13. Index BAM file using Samtools
 - 13.1. cd samtools
 - 13.2. ./samtools index ~/addOrReplace_output.bam

- 13.3. Output: addOrReplace_output.bam.bai
14. Remove unmapped reads using Samtools
- 14.1. cd samtools
 - 14.2. ./samtools view -h -F 4 -b ~/addOrReplace_output.bam -o mapped.bam
 - 14.3. Output: mapped.bam
15. Index BAM file using Samtools
- 15.1. cd samtools
 - 15.2. ./samtools index ~/mapped.bam
 - 15.3. Output: mapped.bam.bai
16. Sort BAM file by coordinate using Picard tools
- 16.1. cd picard
 - 16.2. java -jar picard.jar SortSam I= ~/mapped.bam O= mapped_sorted.bam
SORT_ORDER=coordinate
 - 16.3. Output: sorted.bam
17. Index BAM file using Samtools
- 17.1. cd samtools
 - 17.2. ./samtools index ~/mapped_sorted.bam
 - 17.3. Output: mapped_sorted.bam.bai
18. Masking interspersed repeats and low complexity DNA sequences from the D. melanogaster reference genome using RepeatMasker.
- 18.1. cd RepeatMasker
 - 18.2. perl RepeatMasker -species drosophila melanogaster -s -no_is -nolow -norna -pa 4 -div 50 -e rmblast ~/dmel-all-chromosome-r6.20.fasta
 - 18.3. Output: dmel-all-chromosome-r6.20.fasta.masked, dmel-all-chromosome-r6.20.fasta.cat.gz, dmel-all-chromosome-r6.20.fasta.tbl and dmel-all-chromosome-r6.20.fasta.out
 - 18.4. Change output names to: dmel-all-chromosome-r6.20_masked.fasta, dmel-all-chromosome-r6.20_masked.fasta.cat.gz, dmel-all-chromosome-r6.20_masked.fasta.tbl and dmel-all-chromosome-r6.20_masked.fasta.out
19. Indexing masked reference genome using BWA:
- 19.1. cd bwa
 - 19.2. ./bwa index -a bwtsw ~/dmel-all-chromosome-r6.20_masked.fasta

19.3. Output: dmel-all-chromosome-r6.20.fasta.gz.masked.sa, dmel-all-chromosome-r6.20.fasta.gz.masked.bwt, dmel-all-chromosome-r6.20.fasta.gz.masked.amb, dmel-all-chromosome-r6.20.fasta.gz.masked.ann and dmel-all-chromosome-r6.20.fasta.gz.masked.pac

20. Create the FASTA index file using Samtools

20.1. cd samtools

20.2. ./samtools faidx ~/dmel-all-chromosome-r6.20_masked.fasta

20.3. Output: dmel-all-chromosome-r6.20_masked.fasta.fai

21. Generate the sequence dictionary using Picard tools

21.1. cd picard

21.2. java -jar picard.jar CreateSequenceDictionary R= ~/dmel-all-chromosome-r6.20.fasta.gz O=dmel-all-chromosome-r6.20.dict

21.3. ouput: dmel-all-chromosome-r6.20_masked.dict

22. Call SNP's and indels using Genome Analysis Toolkit

22.1. cd gatk

22.2. java -jar GenomeAnalysisTK.jar -R ~/dmel-all-chromosome-r6.20_masked.fasta -T HaplotypeCaller -I ~/mapped_sorted.bam -ERC GVCF -L 2L:1-23513712 -o output_haplotypecaller.g.vcf

22.3. output: output_haplotypecaller.g.vcf.idx and output_haplotypecaller.g.vcf

23. Perform joint genotyping on gVCF files produced by HaplotypeCaller using the Genome Analysis Toolkit

23.1. cd gatk

23.2. java -jar GenomeAnalysisTK.jar -T GenotypeGVCFs -R ~/dmel-all-chromosome-r6.20_masked.fasta --variant ~/output_haplotypecaller.g.vcf -o GVCF_output.vcf

23.3. output: GVCF_output.vcf.idx and GVCF_output.vcf

24. Filtering variants using the Genome Analysis Toolkit

24.1. cd gatk

24.2. java -jar GenomeAnalysisTK.jar -R ~/dmel-all-chromosome-r6.20_masked.fasta -T SelectVariants -nt 4 --variant ~/GVCF_output.vcf -o filtered_variants.vcf -select "QUAL > 20.0 && DP > 20"

24.3. output: filtered_variants.vcf.idx and filtered_variants.vcf

25. Compress file using htslib

25.1. cd htslib

25.2. ./bgzip ~/filtered_variants_output.vcf

25.3. Output: filtered_variants.vcf.gz

26. Index file using htslib

26.1. cd htslib

26.2. ./tabix -p vcf ~/filtered_variants.vcf.gz

26.3. Ouput: filtered_variants.vcf.gz.tbi

27. Determine the effect of the variants found using the Variant Effect Predictor (VEP)

27.1. Activate VEP

27.1.1. mysql -u root -p

27.1.2. Introduce password

27.1.3. In another terminal window:

27.1.4. source activate myperl

27.1.5. source ~/.profile

27.1.6. cd ~/src/ensembl/misc-scripts

27.1.7. ./ping_ensembl.pl

27.1.8. The following message should be printed in the terminal window:

"Installation is good. Connection to Ensembl works and you can query the human core database".

27.2. If password does not work: Open MySQL from System Preferences > Initialize Database > Type new password > Choose 'Use legacy password' > Start MySQL Server.

28. Dividing files using htslib and running VEP

28.1. cd ~/src/ensembl/misc-scripts

28.2. ~/tabix -h ~/filtered_variants.vcf.gz 2L:1-23513712 | ~/vep --cache --fork 4 --format vcf --species drosophila_melanogaster --o vep.txt

28.3. Output: vep.txt_summary.html and vep.txt

29. Filtering VEP file for non-synonymous mutations

29.1. cd ~/src/ensembl/misc-scripts

29.2. ~/filter_vep -i ~/vep.txt -o nonsynonymous.txt –format tab -y -f
"Consequence is nonsynonymous_variant"

29.3. Output: nonsynonymous.txt

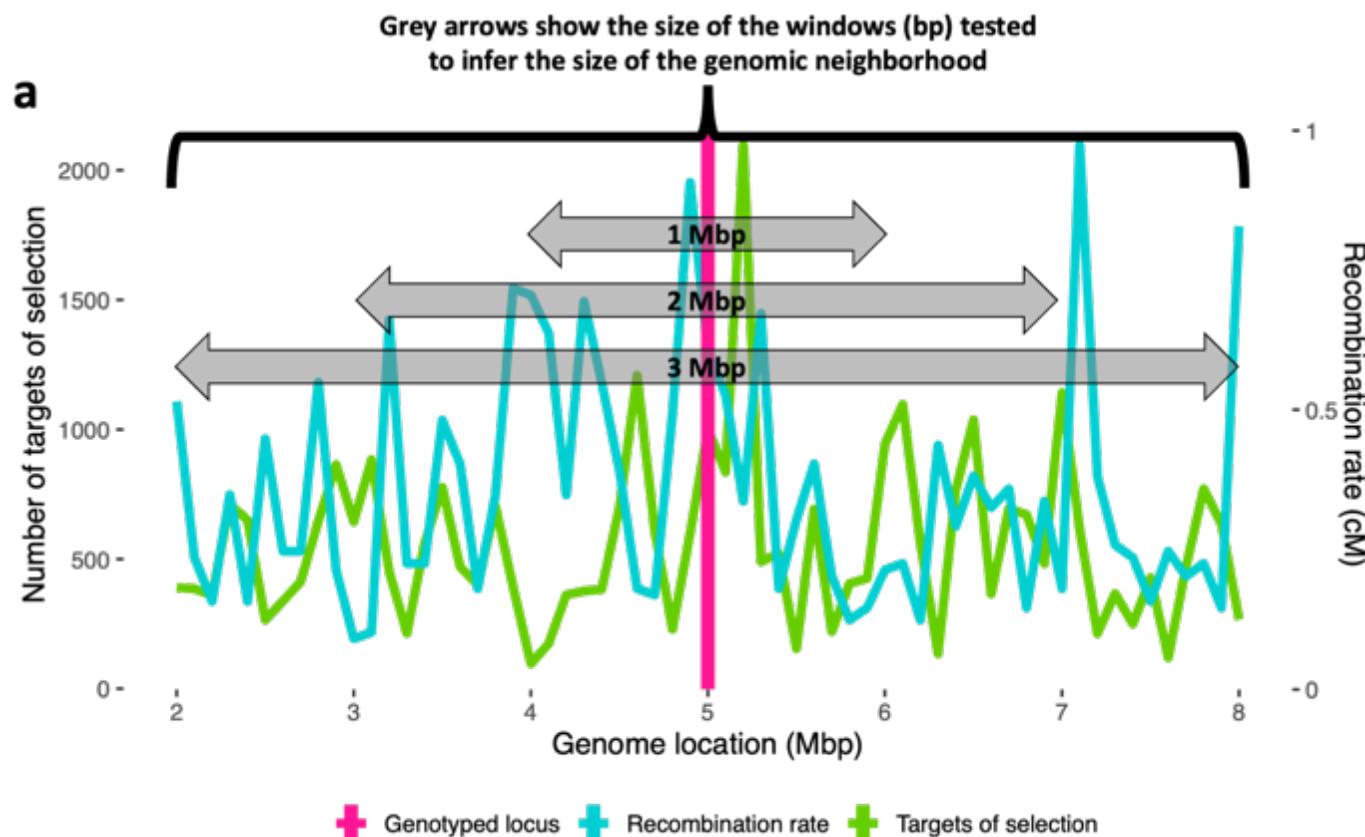
30. Delete comments in the file using terminal

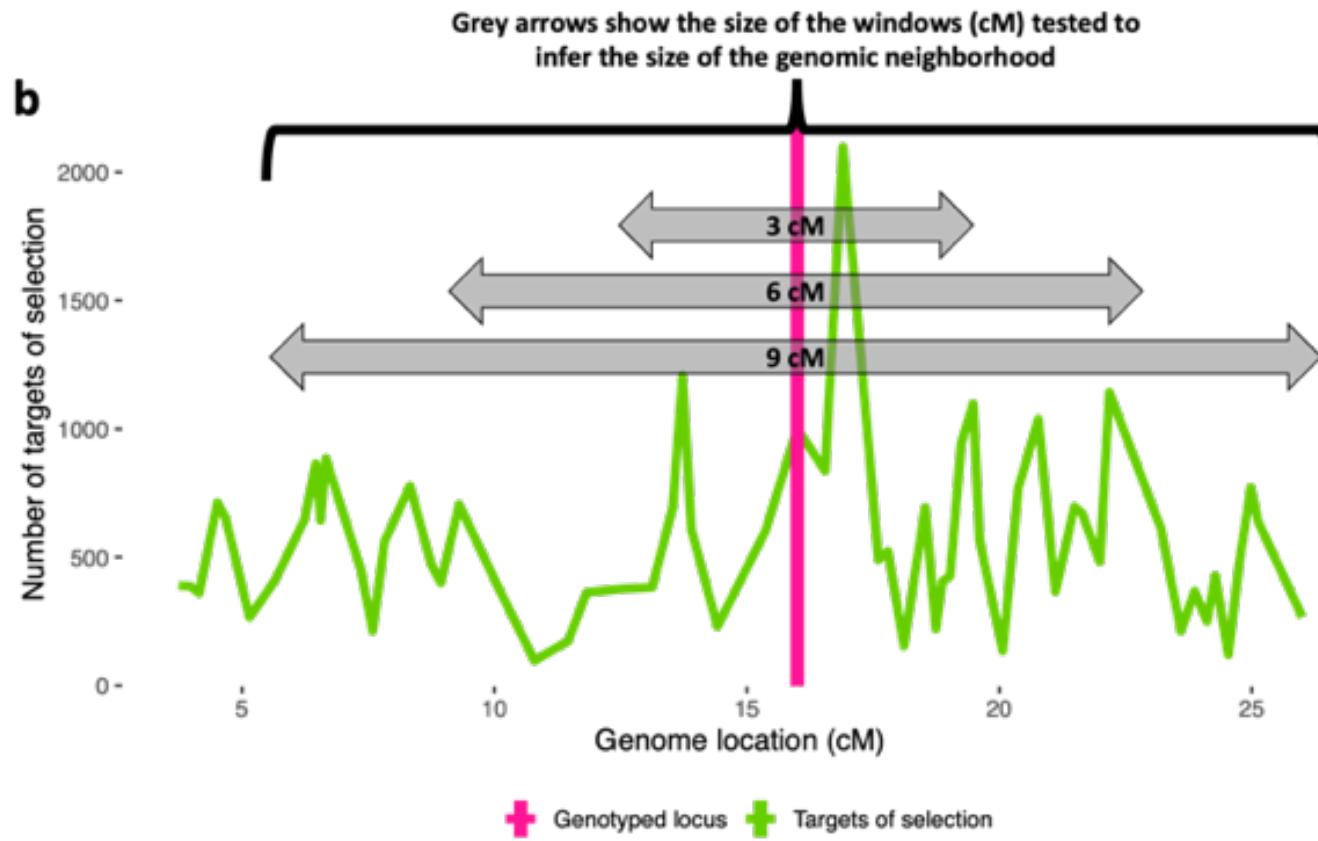
30.1. grep -v "^#" nonsynonymous.txt > nonsynonymous.csv

31. Download Ensembl data, format data, perform regression analysis and plot results
using the R script “test_australia_vs_ensembl.R”

2.9.3 Methods to investigate associative overdominance

2.9.3.1 Genomic neighbourhood

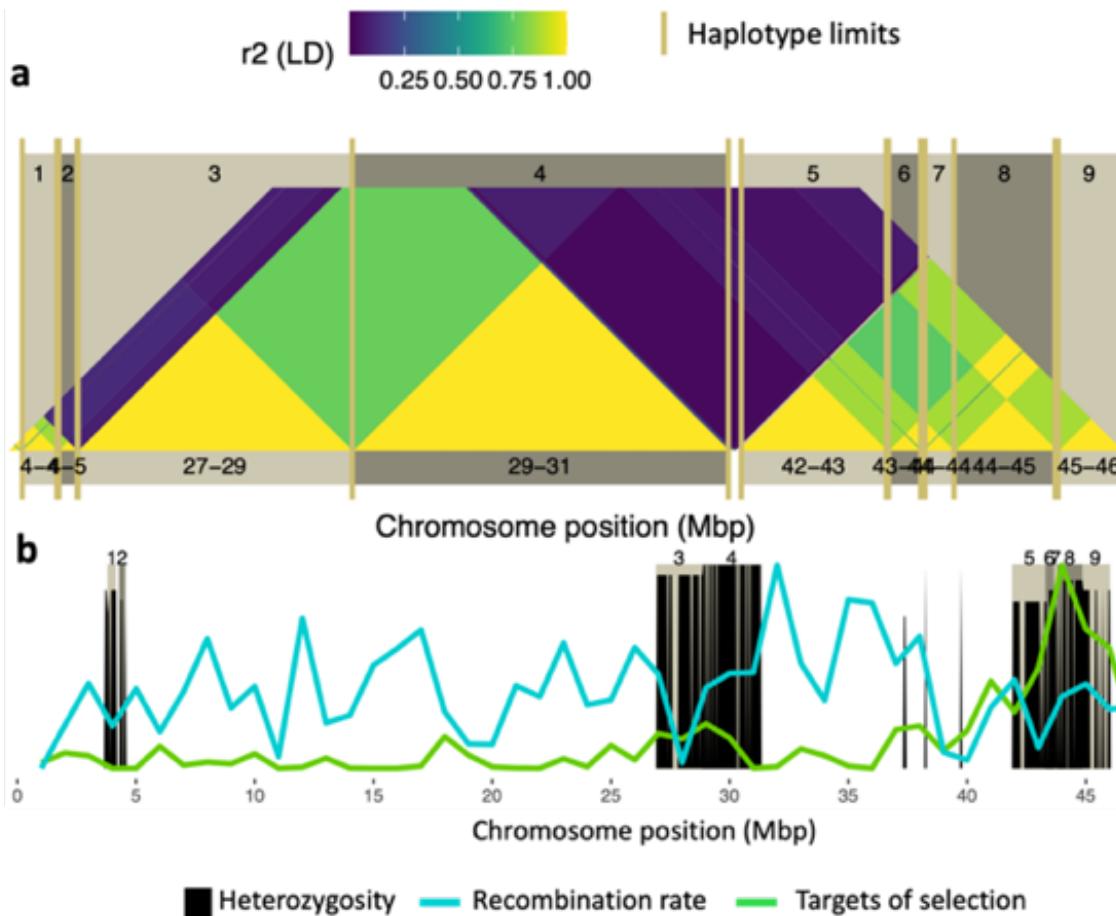




Supplementary Methods Figure 2.4 | Schematic representation of the approach to infer the size of the genomic neighbourhood. **a)** In my regression approach using physical distance (*i.e.*, base pairs; bp), I used as response variables F_{ST} and Heterozygosity (He) of each genotyped locus (pink column). For the first explanatory variable for the flies and as a proxy for the number of targets of selection within a gene, I used the difference between the number of non-synonymous (NS) and synonymous (S) mutations (*i.e.*, $NS - S$) and for the Chillingham cattle, I used the number of NS as the proxy for the number of targets of selection (green line). The second explanatory variable was recombination rate (in centiMorgans; cyan line). The explanatory variables were each measured in several windows of different sizes measured in bp (grey arrows). One multiple regression analysis was performed for each window size to infer the size of the genomic neighbourhood. **b)** In my regression approach using genetic distance (*i.e.*, recombination rate in centiMorgans; cM), I used as response variables F_{ST} and He of each genotyped locus (pink column). As explanatory variable I used my proxy for targets of selection (green line). The explanatory variable was measured in windows of different sizes measured in cM (grey arrows). One simple linear regression analysis was performed for each window size to infer the size of the genomic neighbourhood.

2.9.3.2 Haplotype identification

To investigate patterns of linked selection, I developed a program to visualize and identity haplotypes using the programming language R. By identifying haplotypes (*e.g.*, Supplementary Methods Figure 2.5), I was able to calculate for each haplotype: the amount of recombination across each haplotype, the number of proxies for targets of selection and the number of polymorphic loci. My program calculates linkage disequilibrium between pairs of loci as measured by the r^2 statistic (Hill & Robertson, 1968) and calculated by using the source code of the R package *snpStats* (Clayton, 2018).



Supplementary Methods Figure 2.5 | Example of output of my method to identify and visualize haplotypes in cattle. **a)** Linkage disequilibrium (LD) heatmap of the chromosome 29 of the Chillingham cattle, where the colour shading indicates the strength of LD (r^2). Genomic locations are shown on the horizontal axis, and haplotypes appear as bright yellow triangles and delimited by dark yellow vertical lines. Numbers identifying each haplotype are shown in the upper part of **a** and **b**. **b)** Visualisation of recombination rate (blue line), proxies for targets of selection (green line) and polymorphic loci (black columns) across chromosome 29.

3 Computer simulations to investigate linked selection mechanisms

3.1 INTRODUCTION

It is now widely accepted that natural selection can affect not only genetic variation with direct consequences on fitness, but can also affect adjacent neutral genetic variation due to genetic linkage (Smith & Haigh, 1974). This effect has been increasingly known as “linked selection”.

Associative overdominance (AOD) is a type of linked selection which was first proposed by Frydenberg (1963). AOD is thought to arise if a genome location has two or more haplotypes (groups of alleles at different loci in linkage disequilibrium that are inherited together) and each one of these haplotypes have various deleterious recessive alleles, and these mutations are located in different positions in each haplotype (*i.e.*, haplotypes have opposite combinations of deleterious mutations). This arrangement of deleterious mutations in the haplotypes will result in the expression of the deleterious recessive mutations (*i.e.*, reduction in fitness) in an individual carrying the same haplotype in both chromosomes (homozygote). In contrast, in individuals carrying two different haplotypes (heterozygote), deleterious mutations will not be expressed due to the opposite or “repulsion” arrangement of deleterious recessive mutations in the haplotypes. This mechanism produces an apparent overdominance (*i.e.*, heterozygous individuals have a higher fitness than homozygous individuals) that results in the maintenance of genetic diversity within populations (*e.g.*, heterozygosity; H_e) and the depression of genetic differentiation between populations (*e.g.*, F_{ST}). A detailed description of the genetic mechanisms of AOD and the elements and factors involved in the development of AOD can be found in Section 3.2.1.4.1 “Basic principles”.

The effects of AOD on genetic diversity sharply contrast with those of background selection (BGS; Charlesworth, 1994). BGS is a different type of linked selection in which the elimination of deleterious alleles from a population has the effect of removing genetic diversity at linked loci. BGS will therefore result in a lower H_e and greater F_{ST} than would be expected for a given effective population size (N_e).

Proposed AOD mechanisms (Latter, 1998; Zhao & Charlesworth, 2016; Becher *et al.*, 2020; Gilbert, Pouyet, *et al.*, 2020) involve the interaction between natural selection, genetic drift and recombination at multiple linked loci. Such a scenario cannot be completely solved analytically by theoretical models. To deal with this problem, I used computer simulations and in particular agent-based models (ABM). These models simulate a system using a collection of autonomous individual entities called agents, whose interactions are dictated by a set of rules (Bonabeau, 2002).

Here I created an ABM that enabled me to investigate and identify the evolutionary, demographic and genomic conditions under which AOD can occur, using realistic scenarios. Even though several simulation programs have been developed to model evolutionary processes (Hoban *et al.*, 2012) and some specialize in simulating recombination (Arenas, 2013) and others can be highly customised and modified (Haller & Messer, 2019), I decided to develop my own simulation program with greater flexibility that allowed me to:

- Expand the program as my knowledge of AOD grew.
- Parameterise the program with the particular genetic characteristics for the two species under study such as the number, location, frequencies, distribution of fitness effects, selection coefficients (s), and dominance coefficients (h) of deleterious alleles occurring in specific chromosomes, and the number and location of recombination events occurring between loci in specific chromosomes.
- Recreate the specific life history and demographics of the study populations, such as source populations, dispersal rate, number of generations, founder individuals, reproductive features, effective population size (Ne), census population size (Nc), sex differences and number of offspring.
- Develop a clear and transparent approach that does not work as a black box.
- Produce the specific outputs that are required to test my hypotheses.

The main characteristics of my program are:

- Each allele in each individual is an agent (*i.e.*, each allele is explicitly simulated), which provides a realistic approach.

- Each locus can be customisable regarding its allele frequencies, selection coefficients, and dominance coefficients.
- Number of loci and individuals to be simulated are only limited by computing resources.
- Recombination is accurately modelled and it is possible to use recombination maps as input.
- The ratio between effective population size and census population size can be easily controlled.
- The program is designed to efficiently run as many replicates as required and provide summary results across generations and replicates.
- The occurrence of AOD in the simulations is not controlled by any setting, rather AOD emerges from the interaction between recombination and selection.

I developed three different simulation models to investigate AOD: general (Section 3.2), fly (Section 3.3) and Chillingham (Section 3.4). In the rest of this chapter, each one of these models is described following the ODD (Overview, Design concepts, Details) protocol for describing individual- and agent-based models (Grimm *et al.*, 2006), as updated by Grimm *et al.* (2020). The ODD format addresses ten elements:

1. **Purpose and patterns.** Describes the purpose of the model and the patterns that serve as model evaluation criteria.
2. **Entities, state variables and scales.** Describes the different types of entities represented in the model. For each entity type, the state variables that characterize it are defined. The temporal and spatial resolution and extent of the model are also specified in this element.
3. **Process overview and scheduling.** Provides an overview of the processes in the model. Since this is only an overview, details of these processes are not included here but in element 7. Submodels.
4. **Design concepts.** Describes how eleven concepts important for the design of ABMs were considered in the model.
5. **Initialization.** Describes how all the model entities are created before the simulations start.

6. **Input data.** Describes the input files and how they are produced.
7. **Submodels.** Describes each submodel with sufficient detail so the model can be completely re-implemented.
8. **Calibration.** Describes the parameters that can be calibrated, the calibration justification and the calibration process.
9. **Validation.** Present the theoretical patterns that were reproduced by the model.
10. **Simulation experiments.** Provides all the details, and their justification, required to replicate the simulation results presented in Chapter 2.

The models were implemented in the programming language R v4 (R Core Team, 2021). The R code of the models, dummy datasets and the R scripts to produce the inputs files, analyse and plot the data, calibrate and validate the model are available on GitHub (<https://github.com/mijangos81/AOD>).

Chapter 2 describes three datasets: general, fly and Chillingham. The program described in this chapter addresses each of these. The general model serves as a foundation for the fly and Chillingham models. The general model is described in detail with the differences with the fly model described in Section 3.3 and the Chillingham model in Section 3.4.

3.2 GENERAL SIMULATIONS

3.2.1 Model description

3.2.1.1 Purpose and patterns

The purpose of the general model is to establish and characterise the evolutionary, demographic and genomic conditions in which linked selection via deleterious alleles impacts genetic diversity.

There are two mechanisms by which linked selection via deleterious alleles can impact neutral genetic diversity: background selection (BGS) and associative overdominance (AOD). A detailed explanation can be found in Section 3.2.1.4.1. The six patterns produced by each mechanism are described in Table 3.1.

Table 3.1 Patterns produced by mechanisms of linked selection via deleterious alleles. He ; heterozygosity, genetic diversity within populations. F_{ST} ; genetic differentiation between populations.

Pattern expected for linked neutral loci	Background selection	Associative overdominance
Bias in He relative to neutral expectations	negative	positive
Bias in F_{ST} relative to neutral expectations	positive	negative
Correlation between He and density of deleterious alleles	negative	positive
Correlation between F_{ST} and density of deleterious alleles	positive	negative
Correlation between He and recombination rate	positive	negative
Correlation between F_{ST} and recombination rate	negative	positive

3.2.1.2 Entities, state variables, and scales

Chromosomes are contained within individuals and individuals are contained within populations. Every individual has two chromosomes (*i.e.* diploid organism). The model simulates two populations simultaneously. In the model every allele is an autonomous agent. Alleles are coded as characters within a string (chromosome). The locations of the alleles within chromosomes (loci; singular locus) can be of two types: under selection or neutral. Loci under selection contain deleterious alleles and alternative

alleles but not neutral alleles. Neutral loci contain only neutral alleles. There are three kinds of alleles: deleterious (with detrimental effects on fitness and coded with the character “a”), alternative or “wild” (with no effects on fitness and coded with the character “A”) and neutral (with no effects on fitness and coded with numbers ranging from one to nine). Beneficial alleles are not considered in the model because it is thought that their impact on genome-wide genetic diversity is limited in the circumstances in which the model is intended to be used *i.e.*, small temporal scales (< 100 generations) and small effective population sizes ($N_e < 300$; Corbett-Detig *et al.*, 2015).

Alleles have four attributes that remain constant for a particular allele throughout the simulations: selection coefficient (s ; ranging from 0 to 1, where $s = 0$ means that allele has no effect on fitness and $s = 1$ means allele is lethal), dominance coefficient (h ; ranging from 0 to 1, where $h = 0$ is completely recessive and $h = 1$ is completely dominant), physical location (base pairs, bp) and genetic location (in recombination distance, centiMorgans, cM). Deleterious alleles have positive values of s and h . Alternative and neutral alleles have values of s and h equal to 0. Individuals have one attribute: sex (male or female).

Alleles occur in a two-dimensional space, horizontally defined by their location within the chromosome (*i.e.*, locus) and vertically defined by the individual in which they are contained. Each time step corresponds to one generation.

3.2.1.3 Process overview and scheduling

The general simulation involves three main processes: loading of input data (performed once at the beginning of the simulations), initialization (performed at the beginning of each simulation replicate), and four submodels (performed every generation; see Figure 3.1). In the general simulations four submodels are executed every generation in the following order:

1. Dispersal.
2. Sexual reproduction and recombination.

3. Natural selection only on the loci under selection (selection of offspring to become the parents of the next generation).
4. Calculation of genetic statistics.

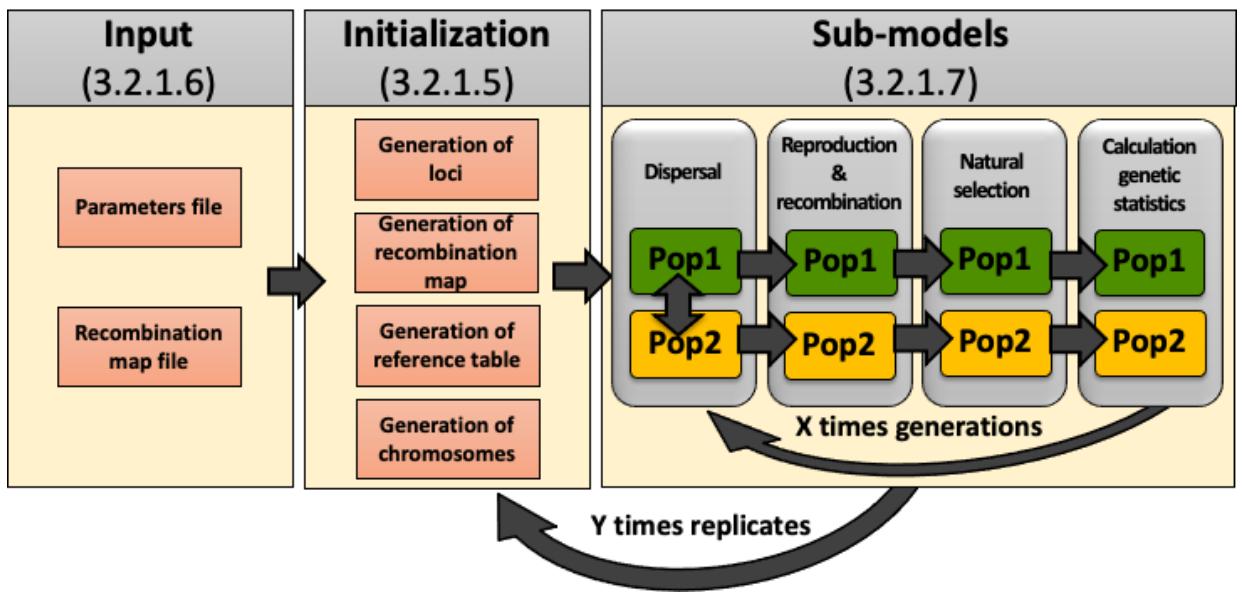


Figure 3.1 | Schematic representation of the processes executed in the general simulations. Information within parentheses correspond to the section in which the process is described.

The model follows standard conventions used in population genetic models (Wright, 1943; Crow & Kimura, 1970; Charlesworth & Charlesworth, 2010), see Section 3.2.1.7 for their justification:

- Dispersal between the two populations is symmetric and constant across generations.
- Dispersing individuals are drawn at random from each population.
- One male and one female are sampled at random without replacement to form mating pairs.
- All individuals reproduce.
- The number of offspring that each mating pair produces is sampled from a negative binomial distribution.
- Generations are not overlapping (*i.e.*, parents and offspring do not coexist).

- The size of the two populations remains constant throughout the simulation, *i.e.*, in each generation the entire population is replaced by sampling the same number of offspring as there were parents in the previous generation.
- Sex ratio is equal within each population.

Calculation of the fitness of offspring can be based on an additive fitness model or on a multiplicative fitness model (Bergen, 2015). Further explanation can be found in Section 3.2.1.7.3.

Selection of offspring to become the parents of the next generation can be based on a relative fitness model or on an absolute fitness model (Wallace, 1975; Agrawal & Whitlock, 2012; Lesecque *et al.*, 2012; Gilbert *et al.*, 2018). Further explanation can be found in section 3.2.1.7.3.

Description of the genetic statistics that are calculated can be found in Section 3.2.1.4.6.

Effective population size (N_e) is controlled by adjusting the variance of the negative binomial distribution that is used to determine the number of offspring that each mating pair produces (Crow & Kimura, 1970, p. 110). Further explanation can be found in sections 3.2.2.

Mutation is not considered in the model because the rate at which new deleterious mutations arise (1-2 per genome per generation; Haag-Liautard *et al.*, 2007; Keightley, 2012) is likely too small to produce any genome-wide effects in genetic diversity in the temporal scale (< 100 generations) in which the model is intended to be used.

The biological process of recombination involves two different events: gene conversion and crossing over. In the model, gene conversion events are not considered because the interval at which they disrupt linkage disequilibrium (< 500 bp; Comeron *et al.*, 2012) is likely too small to significantly disrupt linked selection via deleterious alleles (Campos *et al.*, 2017).

3.2.1.4 Design concepts

The following ODD design concepts were not relevant to the model and therefore not implemented: adaptation, objectives, learning, prediction, and sensing.

3.2.1.4.1 Basic principles

Here I describe the biological aspects that the model simulates, while this section might present basic information, I felt that this description is necessary to place the model within its larger context and be understandable to persons outside of my research area. This section is based on Mijangos *et al.*, (2017).

Genomes consist of sequences of deoxyribonucleic acid (DNA) that encode most of an organism's traits. DNA is packaged within structures called chromosomes. A diploid organism carries two copies of each chromosome (together are called homologous chromosomes), with one copy inherited from each of its parents. The location of a specific DNA sequence within the genome is called locus (with loci being the plural), and the different variants of DNA sequences at a locus are called "alleles". The word "gene" is often ambiguously used to make reference to both "locus" and "allele"; therefore, I avoid the use of gene in this chapter. An individual carrying different alleles at a particular locus is said to be heterozygote at that locus, while an individual carrying the same allele is a homozygote at that locus.

Neutral loci do not affect fitness (*i.e.*, survival or reproduction of living organisms) and therefore are not affected by natural selection. Neutral models frequently assume that loci are inherited by offspring as independent units from their parents (due to random mating, independent assortment of chromosomes and recombination). However, loci that are close to each other on the same chromosome are often inherited together in clusters of different loci termed haplotypes. This non-independent inheritance of loci will cause a non-random association between their alleles causing a linkage disequilibrium (LD) between loci. Conversely, the process of recombination shuffles alleles between haplotypes and hence dilutes the amount of LD between loci. Recombination occurs during meiosis, when homologous chromosomes temporarily pair and exchange segments between each other. As a result of recombination,

haplotypes with different combinations of alleles arise, which are then inherited by offsprings.

When LD occurs between two loci, one neutral and one under selection, theoretical expectations for neutral loci break down, so that sometimes neutral loci appear as if they are under selection (“linked selection”). Linked selection can cause the loss of genetic variation at a higher rate than expected by neutral models under two different scenarios: a) via selective sweeps, where a beneficial allele will be spread across the population along with the alleles at other loci that are linked to it (Smith & Haigh, 1974); and b) via background selection (BGS), where a deleterious allele will be eliminated from the population along with the alleles at other loci linked to it (Charlesworth, 1994).

However, in some circumstances linked selection may result in the maintenance of genetic variation rather than its loss (Rumball *et al.*, 1994). The current hypothesis explaining this phenomenon has been named associative overdominance (AOD). Overdominance arises when a heterozygote has a higher fitness than either homozygote. This is a type of balancing selection, where selection maintains two or more alleles in the population. In the case of AOD, apparent overdominance occurs if an individual has two different haplotypes (*i.e.*, the individual is heterozygous for the haplotypes), and each haplotype contains recessive deleterious alleles (“a”) at different positions (called “repulsion”; Figure 3.2). This arrangement of recessive deleterious alleles means that functionally normal alleles (alleles “A”) from one haplotype could mask the detrimental effects of deleterious alleles (“a”) in the alternative haplotype. Hence, a homozygote of either haplotype will be less fit than a heterozygote, resulting in the maintenance of both haplotypes. Eventually, this will break up due to recombination occurring within the haplotypes.

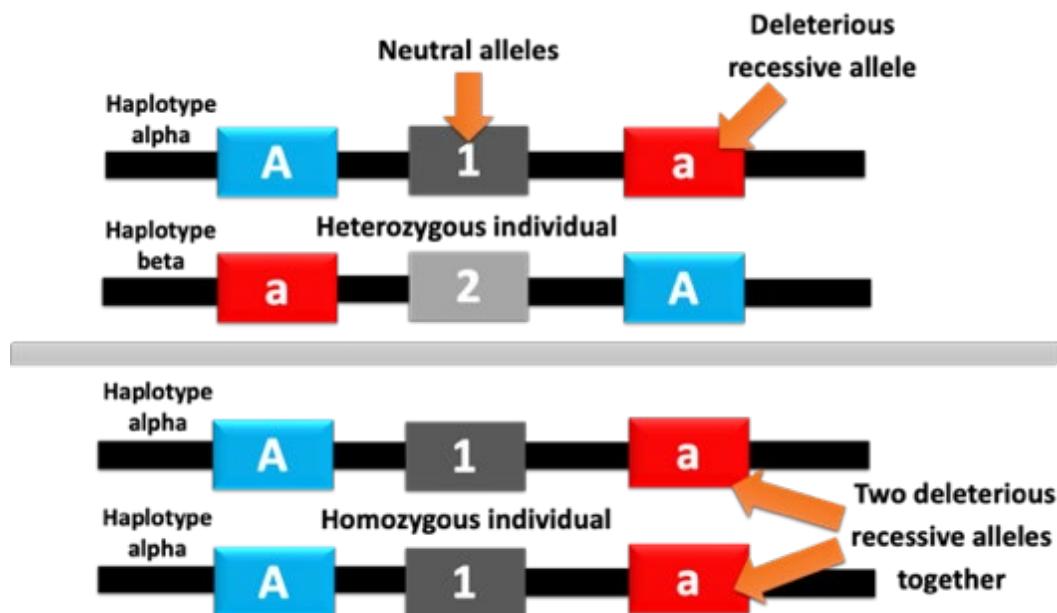


Figure 3.2 | Schematic representation of AOD's mechanism. The figure shows in the upper panel a heterozygous individual for a 3-locus haplotype, in which deleterious recessive alleles (red squares, “a”) are masked by the alternative alleles (blue squares, “A”) of the opposite haplotype. The lower panel shows a homozygous individual, in which the harmful effects of recessive deleterious alleles are expressed. As a result, both haplotypes, alpha and beta, will be maintained by AOD, along with alleles at a linked neutral locus (grey squares, “1” and “2”).

For linked selection to occur, it is necessary for a neutral locus to be located within a specific distance of the loci under selection, so recombination will be relatively slow to break up the LD between these loci. The genetic distance between two loci within a chromosome is measured in units called centiMorgans (cM). One cM corresponds to one percent of probability that the two loci will be separated by a recombination event in each meiosis.

Genetic drift is the random change of allele frequencies over time, including loss. The effects of genetic drift depend on N_e . The smaller N_e is, the stronger the effects of genetic drift. Haplotypes, just as alleles, are lost more rapidly in small populations than in large populations due to the increased effects of genetic drift. As a result, recombination will be less efficient in creating new combinations of alleles, because there are fewer available haplotypes upon which recombination could act. In the long

term, reductions in the number of haplotypes might lead to an increase in LD between loci.

The effects of natural selection depend on selection coefficients. The higher the selection coefficient, the stronger the effect of natural selection, especially if the allele is dominant. Whether an allele is expressed in the phenotype depends on its level of dominance (whether dominant or recessive). Recessive alleles are expressed only when they are present in a homozygote. In contrast, dominant alleles are expressed when an individual is homozygous or heterozygous for the dominant allele. BGS and selective sweeps will have different efficiency depending upon whether the deleterious alleles concerned are dominant or recessive, with the former resulting in accelerated rates of loss of genetic diversity. In the case of AOD, it is necessary that alleles to be at least partially recessive, so that they can be expressed in homozygotes, but not in heterozygotes and thus allowing an apparent overdominance.

3.2.1.4.2 Emergence

The key result of the model is the change in the frequency of the alleles in each locus that occurs from generation to generation. The frequencies of the alleles within and between populations are summarized by genetic statistics (Section 3.2.1.4.6). In the model, the change in the frequencies of the alleles emerge from the combination of the effects of simulating four evolutionary processes:

1. Gene flow (exchange of alleles between populations; Section 3.2.1.7.1).
2. Recombination (shuffling of alleles between homologous chromosomes; Section 3.2.1.7.2).
3. Natural selection (selection of individuals with the fittest combination of alleles; Section 3.2.1.7.3), and;
4. Genetic drift (effect of sampling alleles from generation to generation in finite populations; Section 3.2.1.7.3).

3.2.1.4.3 Interaction

All the interactions between the entities of the model are mediated (*i.e.*, an entity affects others indirectly). There are five interactions between the model entities:

1. Interaction between alleles in the same locus within each individual occurs in the “Calculation of the fitness of each offspring” process Section 3.2.1.7.3. The fitness of a locus within an individual is dependent on the alleles it contains.
2. Interaction between the loci as a group and the individual that contains them, occurs in the “Calculation of the fitness of each offspring” process Section 3.2.1.7.3. The fitness of an individual is dependent on the accumulative fitness of each locus that the individual contains.
3. Interaction between homologous chromosomes within each individual occurs in the “Sexual reproduction/recombination” process Section 3.2.1.7.2. Homologous chromosomes exchange alleles during the process of recombination.
4. Interaction between offspring occurs in the “Sampling/selection of offspring to become the parents of the next generation” process Section 3.2.1.7.3. The probability of becoming parents of the next generation is dependent on the fitness of the individual relative to the fitness of the other individuals in the population.
5. Interaction between populations occurs in the “Dispersal” process Section 3.2.1.7.1. Populations exchange individuals by symmetric dispersal.

3.2.1.4.4 Stochasticity

Justification for the use of stochasticity to determine the values in Table 3.2 can be found in the full description of each process.

Table 3.2 Points in the model on which stochasticity is used to determine values and sections where these are justified.

Values determined by stochasticity	Process of the model where stochastic values are used	Description
The probability of sampling a given neutral allele for each neutral locus.	Generation of chromosomes Section 3.2.1.5.3	All the neutral alleles have the same probability (1 / number of neutral alleles) to be chosen.
The individuals to disperse to the opposite population.	Dispersal Section 3.2.1.7.1	Sampled at random without replacement.
The formation of mating pairs.	Reproduction Section 3.2.1.7.2	One male and one female are sampled at random without replacement.
The number of offspring that each mating pair produces.	Reproduction Section 3.2.1.7.2	Sampled from a negative binomial distribution.
The number of recombination events per meiosis.	Recombination Section 3.2.1.7.2	Sampled from a Poisson distribution.
The assignment of one of the two gametes of each parent to each offspring.	Reproduction Section 3.2.1.7.2	Sampled at random.
Number to decide whether an offspring survives or dies in an absolute fitness model.	Selection Section 3.2.1.7.3	Sampled from a uniform distribution.
The offspring to become parents of the next generation in an absolute fitness model.	Selection Section 3.2.1.7.3	Sampled at random without replacement.
The probability of offspring to become parents of the next generation in a relative fitness model.	Selection Section 3.2.1.7.3	Determined by the fitness of all offspring.

3.2.1.4.5 Collectives

There are four collectives: loci are a collection of alleles of all the individuals in the same chromosome position; chromosomes are a collection of alleles; individuals are a collection of two chromosomes; and populations are a collection of individuals.

3.2.1.4.6 Observation

Genetic statistics were calculated by using the source codes of the functions of the R packages *Hierfstat* (Goudet, 2005), *SpadeR* (Chao *et al.*, 2015) and *snpStats* (Clayton, 2018). The full description of all the statistics and the equations used to calculate them can be found in Section 3.2.1.7.4.

Statistics calculated at the beginning of the simulation:

- Density of deleterious alleles per individual per centiMorgan (del/ind/cM).
- Dispersal rate (Holsinger, 2020, p. 93).
- Expected equilibrium F_{ST} under neutrality (Takahata, 1983).
- Expected equilibrium mutual information under neutrality (Sherwin *et al.*, 2017).
- Rate of loss of heterozygosity under neutrality (Crow & Kimura, 1970).

Statistics calculated for each locus (under selection and neutral) in each generation:

- Shannon index (Sherwin *et al.*, 2017).
- Observed heterozygosity (Nei, 1987).
- Expected heterozygosity (Nei, 1987).
- Expected heterozygosity under neutrality (Crow & Kimura, 1970, p. 329).
- Allele Frequency Difference (Berner, 2019).
- Shannon differentiation (Sherwin *et al.*, 2017).
- Mutual information (Sherwin *et al.*, 2017).
- F_{ST} (Nei, 1987).
- F_{ST} corrected for sample size (Nei, 1987).
- Jost's D (Jost, 2008).

Statistics calculated for each population in each generation:

- Genetic load per locus (Crow & Kimura, 1970).
- Multiplicative fitness of the population (Crow & Kimura, 1970, p. 329).
- Genetic load based on multiplicative fitness (Crow & Kimura, 1970, p. 329).
- Genetic load based on additive fitness (Kimura *et al.*, 1963).

Statistics calculated at the end of the simulation in each population:

- Linkage disequilibrium between each pair of loci using the r^2 statistic (Hill & Robertson, 1968).
- Percentage of loci in which the deleterious allele was eliminated.
- Number of deleterious alleles.
- Generation of input files for three methods to detect loci under selection based on F_{ST} outliers: *FDist2* (Beaumont & Nichols, 1996), *BayeScan* (Foll & Gaggiotti, 2008) and *OutFLANK* (Whitlock & Lotterhos, 2015).

Statistics calculated over all the simulation's replicates:

- Mean of all the statistics
- Standard deviation of all statistics

3.2.1.5 Initialization (Table 3.3)

The following sections describe in detail all the general simulation processes. The description of each process begins with a table containing a small description of the parameters that can be controlled in each process. However, the full explanation of the parameters and how they are used in the model are described in the text below the table and parameter names appear in an ***italic bold*** format. For clarity, I used the same parameter names as used in the actual code of the model. The names of the parameters were chosen to be self-explanatory.

Table 3.3 Parameters used in the initialization process

Parameter name	Description	Range of values
<i>number_loci_under_selection</i>	Number of loci under selection to simulate	Dependent on computing resources
<i>map_resolution</i>	Resolution of the recombination map (bp)	Dependent on genomic resources
<i>s_gral</i>	Selection coefficient of deleterious alleles	0-1
<i>h_gral</i>	Dominance coefficient of deleterious alleles	0-1
<i>q_gral</i>	Initial frequencies of deleterious alleles	0-1
<i>number_neutral_alleles</i>	Number of alleles per neutral locus. The same initial frequency (p) is used for each allele.	1-9
<i>pop_size</i>	Population size (must be even)	Dependent on computing resources

3.2.1.5.1 Generation of loci, assignment of genetic (cM) and physical location (bp) and generation of the recombination map

The processes of generation of loci, assignment of their genetic and physical location within chromosomes and generation of the recombination map are performed in five steps, which are described below including an example of a simple case.

In **Step 1** (Figure 3.3), the model uploads the recombination map input file: “gral_recom_map.csv”. Details about how this input file is generated can be found in the Section input data (3.2.1.6). The recombination map is a table where each row contains the number of centiMorgans (cM) occurring in every chromosome region of size = ***map_resolution***. Note that 1 cM corresponds to one percent of probability that two loci will be separated by a recombination event in each meiosis.

Step 1

$$\text{map_resolution} = 100,000 \text{ bp}$$

Model uploads the recombination map input file (gral_recom_map.csv). In this example, the file has two rows with a total recombination rate in the chromosome of 20 cM and a total length of 200,000 bp.

centiMorgans (cM)
10
10

Figure 3.3 | Schematic representation of Step 1.

In **Step 2** (Figure 3.4), the number of neutral loci to simulate is determined by the number of rows of the recombination map (one neutral locus per row). The number of loci under selection is determined by the parameter *number_loci_under_selection*. Next, the total number of loci to simulate (*i.e.*, neutral loci + loci under selection) are evenly distributed in each row of the recombination map.

Step 2

$$\text{number_loci_under_selection} = 6.$$

Neutral loci = 2 (one neutral locus for each row in gral_recom_map.csv).

Total number of loci to simulate = neutral loci + *number_loci_under_selection* = 8.

Total number of loci to simulate are evenly distributed in each row.

centiMorgans (cM)	Total number of loci to simulate
10	4
10	4

Figure 3.4 | Schematic representation of Step 2.

In **Step 3** (Figure 3.5), for each row of the recombination map, two divisions are performed: **a**) number of cM / number of loci, and; **b**) the size of the chromosome region that each row represents (*i.e.*, *map_resolution*) / number of loci.

Step 3

$$\text{map_resolution} = 100,000.$$

centiMorgans (cM)	Total number of loci to simulate	Division a (cM / number of loci)	Division b (<i>map_resolution</i> / number of loci)
10	4	2.5	25,000
10	4	2.5	25,000

Figure 3.5 | Schematic representation of Step 3.

In **Step 4** (Figure 3.6), to determine the genetic and physical location of each locus, cumulative sums are performed on the values of division **a** and division **b** for each locus, respectively. The results of these two divisions (**a** and **b**) are assigned to each one of the loci in each row. The values of division **a** are used in the recombination process to model the probability of a recombination event occurring in a given genomic location (Section 3.2.1.7.2). Loci that are located at the centre of each region of size ***map_resolution*** are set as neutral (highlighted in yellow in the below example).

Step 4

Locus	Division a (cM / number of loci)	Division b (<i>map_resolution</i> / number of loci)	Genetic location – cumulative sum division a	Physical location – cumulative sum division b
1	2.5	25,000	2.5	25,000
2	2.5	25,000	5.0	50,000
3	2.5	25,000	7.5	75,000
4	2.5	25,000	10.0	100,000
5	2.5	25,000	12.5	125,000
6	2.5	25,000	15.0	150,000
7	2.5	25,000	17.5	175,000
8	2.5	25,000	20.0	200,000

Figure 3.6 | Schematic representation of Step 4.

In order to model recombination accurately, we must account for the case when the total probability of recombination within a chromosome is less than an integer. For instance, in a chromosome of 20 cM, the probability of a recombination event occurring during a meiosis is 20%, and therefore the probability of a recombination event not occurring is 80%. In **Step 5** (Figure 3.7), to obtain the probability of a recombination event not occurring, the model subtracts the probability of the occurrence of a recombination event during a meiosis (*i.e.*, sum of the values of division **a** divided by 100) from the next integer. This value is then multiplied by 100 and inserted in the last row of the recombination map. If this row is chosen as the recombination point in the recombination process (Section 3.2.1.7.2), recombination does not occur. For example, if a chromosome of 20 cM is simulated, the last row of the recombination map will have a value of 80 and therefore 80% of the time recombination will not occur.

Step 5

$$\text{Sum of values of division a} / 100 = 20 / 100 = 0.2$$

$$\text{Next integer (1)} - 0.2 = 0.8 * 100 = 80$$

80 is inserted in the last row of the recombination map

Locus	Division a (cM / number of loci)	Division b (map_resolution / number of loci)	Genetic location – cumulative sum division a	Physical location – cumulative sum division b
1	2.5	25,000	2.5	25,000
2	2.5	25,000	5.0	50,000
3	2.5	25,000	7.5	75,000
4	2.5	25,000	10.0	100,000
5	2.5	25,000	12.5	125,000
6	2.5	25,000	15.0	150,000
7	2.5	25,000	17.5	175,000
8	2.5	25,000	20.0	200,000
	→ 80			

Figure 3.7 | Schematic representation of Step 5.

3.2.1.5.2 Generation of the reference table

The reference table has two functions: a) serve as a lookup table to determine the probability of sampling each allele in each locus in the “Generation of chromosomes” process (Section 3.2.1.5.3), and; b) as a lookup table to calculate the fitness of each locus in the “natural selection” process (Section 3.2.1.7.3). The reference table has as many rows as the total number of loci to simulate and three columns that contain: the selection coefficient (s ; determined by the parameter ***s_gral***), the dominance coefficient (h ; ***h_gral***) and the initial frequency of the deleterious allele (q ; ***q_gral***) of each locus. The same value of s , h and q are used for all the loci under selection. Neutral loci have an s and h equal to 0. All the neutral alleles have the same initial frequency which is given by $1 / \text{number_neutral_alleles}$.

3.2.1.5.3 Generation of chromosomes

The generation of a chromosome is performed in three steps: **i**) for loci under selection, a character, either an “a” (deleterious; with detrimental effects on fitness) or an “A” (alternative; with no effects on fitness), is sampled for each row of the reference table generated in the previous section. The column of the reference table containing the q values is used to determine the probability to sample an “a” (probability = q) or an “A” (probability = $p = 1 - q$) for each row; **ii**) for neutral loci, a number between 1 and ***number_neutral_alleles*** is sampled with equal probabilities (*i.e.*, $1 / \text{number_neutral_alleles}$), and; **iii**) the sampled characters (*i.e.*, “a”, “A” or a number) are concatenated in a single string. This process is repeated ***pop_size*** * 4 times, because the model simulates two populations and each individual has two chromosomes. Finally, two chromosomes are assigned at random to each individual. Note that each allele (under selection and neutral) is sampled separately from a binomial distribution (*i.e.*, in the simplest case, one allele is sampled from either one of two alleles with probabilities q and $p = 1 - q$), which results in that all loci being close to linkage equilibrium and to Hardy-Weinberg equilibrium.

3.2.1.6 Input data (Table 3.4)

Table 3.4 Parameters used in the input data process.

Parameter name	Description	Range of values
<i>map_resolution</i>	Resolution of the recombination map (bp)	Defined by the user
<i>c_gral</i>	Recombination rate per region of size <i>map_resolution</i>	Defined by the user
<i>windows_gral</i>	Number of rows of the recombination map	Defined by the user

The recombination map input file for the general simulations (“gral_recom_map.csv”) is generated automatically by the model. The model creates a table with a number of rows equal to the parameter *windows_gral*. Each row represents a chromosome region of size equal to *map_resolution* base pairs. The recombination rate occurring in each row is specified with the parameter *c_gral*. In the general simulations, all the rows have the same recombination rate.

3.2.1.7 Submodels

3.2.1.7.1 Dispersal (Table 3.5)

Table 3.5 Parameters used in the dispersal process.

Parameter name	Description	Range of values
<i>dispersal</i>	Whether dispersal occurs	TRUE or FALSE
<i>number_transfers</i>	Number of dispersing individuals in each transfer event	< population size
<i>transfer_each_gen</i>	Number of generations between dispersal events	< number of generations

Dispersal starts in the second generation. Firstly, *number_transfers* individuals are sampled at random from each population and then exchanged to the opposite population every *transfer_each_gen* generation(s). The sex of the individuals to be exchanged is alternated every time a dispersal event occurs. If two or more individuals are exchanged the same generation, half are males and half females.

3.2.1.7.2 Sexual reproduction and recombination

Sexual reproduction (Table 3.6)

Table 3.6 Parameters used in the sexual reproduction process.

Parameter name	Description	Range of values
number_offspring	Mean number offspring per mating	Dependent on computing resources
variance_offspring	Variance in the number of offspring per mating	Defined by the user

The number of offspring that each pair of parents produces is sampled from a negative binomial distribution with a mean = **number_offspring** and a dispersion parameter = **variance_offspring**. Offspring have the same probability of being either male or female.

The negative binomial distribution was chosen because it has two distinctive attributes:

1. It is a discrete distribution (*i.e.*, a distribution describing the probabilities of the occurrence of values from variables that can be counted) that is essentially the same as the Poisson distribution but its variance can be larger than its mean (Bolker, 2008) and can be explicitly controlled (by the dispersion parameter). The variance of this distribution is mean + (mean² / dispersion parameter). This attribute allows to control the ratio between the effective population size (N_e) and the census population size (N_c), as described in the text below.
2. The mean of the negative binomial can be explicitly controlled (by the parameter mean). This attribute allows to control the number of offspring per mating which is convenient when there is a need that each pair of parents produce enough offspring in each generation for the population not to become extinct.

N_e (the size of an idealised population that would have the same amount of inbreeding, or genetic drift, as the population under consideration; Kimura & Crow, 1963) depends on a number of factors described in Table 3.7 (Crow & Kimura, 1970; Charlesworth & Charlesworth, 2010). The model restricts explicitly each one of these factors, so the only factor affecting N_e is the variance in family size (lifetime production of offspring per individual).

Table 3.7 Factors affecting N_e and the model's approach to restrict/control these factors.

Factors affecting N_e	Approach used in the model to restrict/control factors affecting N_e
Geographic structure	Restricted by random mating
Genetic structure	Restricted by random mating
Fluctuations in population size	Restricted by using a constant population size
Reproduction system	Restricted by using only sexual reproduction
Mode of inheritance	Restricted by using only autosomal inheritance
Number of males and females	Restricted by an equal sex ratio
Age structure	Restricted by non-overlapping generations
Variance in reproductive success	Restricted by all individuals reproducing once
Variance in family size	Controlled by the variance of the negative binomial distribution

The calibration process of the ratio between effective population size and census population size (N_e/N_c) can be found in Section 3.2.2 and the validation of this approach can be found in Section 3.2.3.1.

Recombination (Table 3.8)

Table 3.8 Parameters used in the recombination process.

Parameter name	Description	Range of values
<i>recombination</i>	Whether recombination occurs	TRUE or FALSE

In actual biology, each meiosis produces four gametes, in which a recombination event or events may have occurred or not in each one of them (Figure 3.8). However, researchers who generate recombination maps, measure recombination rate by observing the number of recombination events that occurred not in the gametes but in the offspring (Comeron *et al.*, 2012). Consequently, in the model just two gametes are simulated, instead of four, by doing so recombination occurs only in the gametes that the offspring inherits.

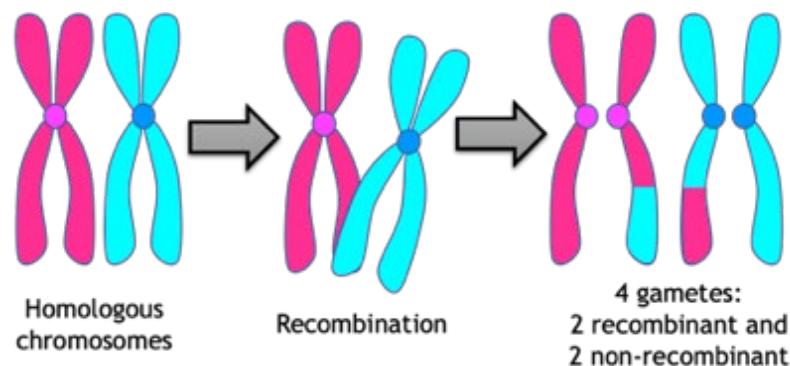


Figure 3.8 | Diagram showing the four gametes produced each meiosis, in this example, two recombinant gametes and two non-recombinant gametes.

For each offspring of each mating, a recombination calculation is performed on the two homologous chromosomes (gametes) of each parent to determine: the number of recombination events; whether recombination occurs and if it occurs determine the location of the recombination point; and which of the homologous chromosomes of each parent the offspring will inherit.

The recombination process is performed in three steps:

Step a. The number of recombination events is sampled from a Poisson distribution (Comeron *et al.*, 2012) with a mean and variance equal to the sum of all the recombination rates in the column of division **a** in the recombination map (see Section 3.2.1.5.1) divided by 100 and rounded to the next integer. If the number of recombination events is > 1, **Step b** is performed otherwise **Step c** is performed.

Step b. The point of recombination is determined by sampling one location (row) from the recombination map with a probability of choosing a given row equal to the recombination rate of each row. If the last row of the recombination map (which contains the probability of a recombination event not occurring) is chosen as the recombination point, recombination does not occur, as explained in the description of the generation of the recombination map in Section 3.2.1.5.1. If recombination occurs, chromosome sequences are split at the recombination point and switched (Figure 3.8). **Step b** is repeated as many times as the number of recombination events determined in **Step a**. If recombination does not occur, **Step c** is performed.

Step c. Independently of whether recombination occurred or not, one of the two gametes (that did or did not have recombination) from each parent is sampled at random and assigned to the offspring.

3.2.1.7.3 Selection (Table 3.9)

Table 3.9 Parameters used in the selection process.

Parameter name	Description	Range of values
<i>selection</i>	Whether selection occurs	TRUE or FALSE
<i>fitness_model</i>	Fitness model to use	“multiplicative” or “additive”
<i>selection_model</i>	Selection model to use	“relative” or “absolute”
<i>genetic_load</i>	Approximation of the genetic load of the proportion of the genome that is simulated. This parameter is used in the absolute fitness model	Defined by the user

For loci under selection with a deleterious allele “a” and an alternative allele “A”, the relative fitness (w) of each allele combination within a locus within an individual (*i.e.* genotype) are: $w_{AA} = 1$, $w_{Aa} = 1 - hs$, and $w_{aa} = 1 - s$, where h is the dominance coefficient of the deleterious allele and s is the selection coefficient of the deleterious allele. For neutral loci $w = 1$ for all genotypes.

The fitness of an individual (W) in a multiplicative model (***fitness_model*** = “multiplicative”) is the cumulative multiplication of the fitness of each locus in the individual (Charlesworth & Charlesworth, 2010, p. 166). This model assumes that fitness is multiplicative across loci, which could happen, for example, if different genes affect viability at different stages in the life cycle (Halliburton, 2004, p. 479).

The fitness of an individual (W) in an additive model (***fitness_model*** = “additive”) is 1 minus the summation of the amount of reduction in fitness contributed by each locus in the individual (Bergen, 2015). The amount of reduction in fitness contributed by each genotype are: $w_{AA} = 0$, $w_{Aa} = hs$, and $w_{aa} = s$. This model assumes that fitness is additive across loci, which could happen, for example, if two genes produce the same protein. In this case, fitness depends on the total amount of the protein produced (Halliburton, 2004, p. 479). Note that the additive model of fitness could lead to a negative value of fitness. In contrast, the multiplicative model never has a negative value of fitness; its minimum value is 0.

In a relative fitness model (***selection_model*** = “relative”), also called soft selection or density dependent selection (Wallace, 1975; Lesecque *et al.*, 2012), the fitness of each individual (W) is dependent on the fitness of other individuals in the population (*i.e.*, based on competition). In this model, first the fitness of each offspring (W) is recalculated as $W_{relative} = W / \text{summation of the fitness of all offspring}$. Afterwards, offspring are selected (equal sex ratio) to become parents of the next generation with a probability equal to their recalculated fitness ($W_{relative}$; Lesecque *et al.*, 2012). For example, if we had four individuals with fitnesses (W) of 0.1, 0.2, 0.3, and 0.2 the first individual would be selected on average $W_{relative} = 0.1 / (0.1 + 0.2 + 0.3 + 0.2) = 0.125$ of the time to become parent of the next generation.

In an absolute fitness model (*selection_model* = "absolute"), also called hard selection or density independent selection (Wallace, 1975; Lesecque *et al.*, 2012), genetic load is defined as "the loss of fitness resulting from deleterious alleles maintained by mutation-selection balance" (Agrawal & Whitlock, 2012). Genetic load effectively measures the fraction of the population that fails to survive or reproduce (Charlesworth & Charlesworth, 2010). The fitness of each individual (W) is independent of the fitness of other individuals in the population (Wallace, 1975). In this model, first a random number is sampled from a uniform distribution within an interval between 0 and *genetic_load*, which is an approximation of the genetic load of the proportion of the genome that is simulated. This approximation can be guided by equations to calculate genetic load described in the Section 3.2.1.7.4. Afterwards, individuals are set to death if the random number is larger than the total fitness of the individual (W ; Wang *et al.*, 1999). Finally, offspring are sampled at random (equal sex ratio) from the remaining live offspring to become parents of the next generation.

3.2.1.7.4 Calculation of genetic statistics (Table 3.10)

Table 3.10 Parameters used in the calculation of genetic statistics process.

Parameter name	Description	Range of values
<i>number_transfers</i>	Number of dispersing individuals	< population size
<i>transfer_each_gen</i>	Number of generations between transfer events	< number of generations
<i>pop_size</i>	Population size (must be even)	Dependent on computing resources
<i>Ne</i>	<i>Ne</i> value to be used in the equation of the expected F_{ST} (Equation 3.2)	Defined by the user
<i>Ne_dispersal</i>	<i>Ne</i> value to be used in the equation of rate of loss of heterozygosity (Equation 3.4)	Defined by the user
<i>ld_max_pairwise</i>	Maximum distance, in base pairs, at which pairwise LD should be calculated	< length of the chromosome
<i>ld_resolution</i>	Resolution, in basepairs, at which LD should be measured	< length of the chromosome
<i>region_size</i>	The size, in base pairs, of the regions at which the LD statistics and number of deleterious alleles are calculated	< length of the chromosome

Statistics calculated at the beginning of the simulation

Calculation of the number of deleterious alleles per individual per centiMorgan (del/ind/cM) is based on the number of heterozygous loci under selection in each individual and allele proportions. For this the Hardy-Weinberg equation (Hardy, 1908; Weinberg, 1908) was used and calculated as:

$$p^2 + 2pq + q^2 = 1, \quad \text{Equation 3.1}$$

where p is the frequency of the alternative allele ("A") and q is the frequency of the deleterious allele ("a"). From this equation the model first calculates the proportion of heterozygotes (*i.e.*, $2pq$) for each locus and then these values are averaged across all the loci. This proportion is then multiplied by the number of loci under selection to be simulated (**number_loci_under_selection**) and divided by the length of the chromosome in centiMorgans.

Dispersal rate (m) is the fraction of individuals in a population that is composed of dispersers or the probability that a randomly chosen individual in this generation came from a population different from the one in which it was found in the preceding generation (Holsinger, 2020, p. 93). Dispersal rate is calculated as (**number_transfers / transfer_each_gen**) / **pop_size**.

Expected F_{ST} under neutrality (Takahata, 1983) is calculated as:

$$F_{ST} = \frac{1}{4Ne m \left(\frac{n}{n-1} \right)^2 + 1}, \quad \text{Equation 3.2}$$

where **Ne** is effective populations size of each individual subpopulation, m is dispersal rate and n the number of subpopulations (always 2).

Expected mutual information for microsatellites data (Sherwin *et al.*, 2017) is calculated as:

$$I = \frac{0.22}{\sqrt{2Ne * m}} - \frac{0.69}{2Ne\sqrt{m}}, \quad \text{Equation 3.3}$$

where ***Ne*** is effective populations size and ***m*** is dispersal rate.

Rate of loss of heterozygosity per generation (Crow & Kimura, 1970, p. 329) is calculated as:

$$1 - \frac{1}{2Ne_{dispersal}}, \quad \text{Equation 3.4}$$

where ***Ne dispersal*** is the *Ne* of each subpopulation corrected for dispersal. A corrected *Ne* is used because high dispersal rates between populations can substantially affect *Ne* estimates (Wang & Whitlock, 2003). Neutral simulations can be used to infer ***Ne dispersal*** as described in Chapter 2, Supplementary Methods 2.9.1.

Statistics calculated for each locus (under selection and neutral) in each generation

Genetic statistics were calculated by using the source codes of the functions of the R packages *Hierfstat* (Goudet, 2005) and *SpadeR* (Chao *et al.*, 2015).

Observed heterozygosity corrected for sample size (Nei, 1987, pp. 164–165) is calculated as:

$$Ho = 1 - \sum_k \sum_i \frac{Pkii}{np}, \quad \text{Equation 3.5}$$

where *Pkii* represents the proportion of homozygote *i* in sample *k* and *np* the number of samples.

Expected heterozygosity corrected for sample size (Nei, 1987, pp. 164–165) is calculated as:

$$He = \frac{\tilde{n}}{\tilde{n} - 1} \left[1 - \sum_i \overline{p_i^2} - \frac{Ho}{2\tilde{n}} \right], \quad \text{Equation 3.6}$$

where $\tilde{n} = \frac{np}{\sum_k \frac{1}{n_k}}$ and $\overline{p_i^2} = \sum_k$

Expected heterozygosity under neutrality (Crow & Kimura, 1970, p. 329) is calculated as:

$$He_t = He_0 \left(1 - \frac{1}{2Ne_{dispersal}} \right)^t, \quad \text{Equation 3.7}$$

where He_0 is heterozygosity at generation 0 and t is the number of generations.

Overall heterozygosity corrected for sample size (Nei, 1987, pp. 164–165) is calculated as:

$$Ht = \sum_i \bar{p}_i^2 + \frac{He}{\tilde{n}np} - \frac{Ho}{2\tilde{n}np}, \quad \text{Equation 3.8}$$

where $\bar{p}_i = \sum_k \frac{p_{ki}}{np}$

Amount of heterozygosity among samples (Nei, 1987, pp. 164–165) is calculated as:

$$Dst = Ht - He, \quad \text{Equation 3.9}$$

Amount of heterozygosity among samples corrected for sample size (Nei, 1987, pp. 164–165) is calculated as:

$$Dst' = \frac{np}{np - 1} Dst \quad \text{Equation 3.10}$$

Overall heterozygosity corrected for sample size (Nei, 1987, pp. 164–165) is calculated as:

$$Ht' = He + Dst' \quad \text{Equation 3.11}$$

G_{ST} (Nei, 1987, pp. 164–165) is calculated as:

$$G_{ST} = \frac{Dst}{Ht} \quad \text{Equation 3.12}$$

$G_{ST'}$ corrected for sample size (Nei, 1987) is calculated as:

$$G_{ST'} = \frac{Dst'}{Ht'} \quad \text{Equation 3.13}$$

Jost's D (Jost, 2008) is calculated as:

$$Dest = \left(\frac{np}{np - 1} \right) \left(\frac{Ht' - He}{1 - He} \right) \quad \text{Equation 3.14}$$

Shannon index (Sherwin *et al.*, 2017) is calculated as:

$$^1H = - \sum_{i=1}^n p_i \log p_i, \quad \text{Equation 3.15}$$

where n is the number of alleles and p_i is the frequency of the i^{th} allele.

Allele Frequency Difference (Berner, 2019) is calculated as:

$$AFD = \frac{1}{2} \sum_{i=1}^n |(p_{i1} - p_{i2})|, \quad \text{Equation 3.16}$$

where n is the total number of different alleles observed at the polymorphism, and the p_i terms specify the frequency of allele i in the two populations.

Mutual information (Sherwin *et al.*, 2017) is calculated as:

$$I = \left(1^1 H_\gamma - \overline{1^1 H_\alpha} \right), \quad \text{Equation 3.17}$$

where $\overline{1^1 H_\alpha}$ denotes the average of each within-population entropy $1^1 H_\alpha$, and $1^1 H_\gamma$ is calculated as for $1^1 H_\alpha$, but after the allelic dataset has been pooled over all populations.

Shannon differentiation (Sherwin *et al.*, 2017) is calculated as:

$$\frac{I}{\ln K}, \quad \text{Equation 3.18}$$

where I is mutual information and K is the number of equal-sized populations. *i.e.*, mutual information normalized to a scale from 0 to 1.

Statistics calculated for each population in each generation

Genetic load per locus (Crow & Kimura, 1970, p. 299) is calculated as:

$$L = 2hspq + sq^2, \quad \text{Equation 3.19}$$

where p is the frequency of the alternative allele, q the frequency of the deleterious allele, h is dominance and s is selection coefficient.

Multiplicative fitness of the population (Charlesworth & Charlesworth, 2010, p. 166) is calculated as:

$$\bar{W} = \prod_{i=1}^l 1 - L_i, \quad \text{Equation 3.20}$$

where L_i is the load at locus i and l is the number of loci in the population, *i.e.*, the cumulative multiplication of 1 minus the load of each locus in the population.

Genetic load based on multiplicative fitness (Charlesworth & Charlesworth, 2010, p. 166) is calculated as:

$$L = 1 - \bar{W}, \quad \text{Equation 3.21}$$

where \bar{W} is the mean fitness of the population.

Genetic load based on additive fitness (Kimura *et al.*, 1963, p. 302) is calculated as:

$$L = \sum L_i, \quad \text{Equation 3.22}$$

where L_i is the load at locus i , i.e., the summation of the genetic load of each locus in the population.

Statistics calculated at the end of the simulation in each population

Pairwise linkage disequilibrium is calculated between each pair of loci within a distance = *ld_max_pairwise* using the r^2 statistic (Hill & Robertson, 1968). r^2 was calculated by using the source code of the R package *snpStats* (Clayton, 2018).

Number of deleterious alleles per region of size = *region_size*.

Statistics calculated over all the simulation's replicates

Mean of all the statistics

Standard deviation of all statistics.

3.2.2 Calibration of the effective size to census size ratio (Table 3.11)

Table 3.11 Parameters used in the process for calibration of the Ne/Nc ratio.

Parameter name	Description	Range of values
<i>selection</i>	Whether selection occurs	TRUE or FALSE
<i>dispersal</i>	Whether dispersal occurs	TRUE or FALSE
<i>number_offspring</i>	Mean number offspring per mating	Dependent on computing resources
<i>variance_offspring</i>	Variance in the number of offspring per mating	Defined by the user
<i>pop_size</i>	Population size (must be even)	Dependent on computing resources
<i>Ne_dispersal</i>	Ne value to be used in the equation of rate of loss of heterozygosity (Equation 3.7)	Defined by the user
<i>gen_number</i>	Number of generations	Dependent on computing resources
<i>number_iterations</i>	Number of simulations replicates	Dependent on computing resources

If the user requires that the Ne/Nc ratio to be equal to 1, a large enough value in the parameter ***variance_offspring*** (*e.g.*, > 1,000) should be used. The corroboration of this point can be found in the next section, which describes the validation of the Ne/Nc ratio of the model.

If the user requires that the Ne/Nc ratio to be different from 1, a calibration process can be performed for this end. Using a Ne/Nc ratio different from 1 is especially useful when studying dispersal because dispersal rates are based on census population size (see Section 3.2.1.7.4). This calibration involves running neutral simulations (*i.e.*, no dispersal and no selection) with different values of the dispersion parameter (***variance_offspring***) until finding the value that satisfies the equation of the neutral rate of loss of heterozygosity (*i.e.*, Equation 3.7).

The steps for the calibration process are as follows:

1. Setting the desired effective population size (*Ne_dispersal*).
2. Setting the desired census population size (*pop_size*).
3. Setting the desired mean number of offspring for each mating (*number_offspring*).
4. Setting the parameters *dispersal* and *selection* to FALSE.
5. Setting the value of the variance of the number of offspring per mating to the value of which is to be tested (*variance_offspring*).
6. Running the simulations for 30 generations (*gen_number*) and for 1,000 replicates (*number_iterations*).
7. The model at the end of the simulations will plot equation Equation 3.7 and the mean heterozygosity per generation per population. See Figure 3.9 for an example.
8. If the trajectory of heterozygosity does not match the trajectory of Equation 3.7, the user should repeat simulations with a different value for the parameter *variance_offspring* until finding the value that satisfies this equation.

3.2.3 Validation

The model was validated by testing the simulations against various theoretical expectations for natural selection, gene flow, genetic drift and recombination. The validation was partially based on equations and methods described in the validation report of the simulation program EvolGenius v6.1 (Kliman, 2014).

3.2.3.1 The effective size to census size ratio

As described by Crow and Kimura (1970, p. 110), in a sexual reproducing population of constant size and random mating, the effective population size (Ne) is equal to the census population size (Nc), that is $Ne/Nc = 1$, if the variance in lifetime production of offspring per individual is equal to two, as demonstrated by the following equation:

$$Ne = \frac{4Nc - 2}{\sigma^2 + 2}, \quad \text{Equation 3.23}$$

where Nc is the census population size and σ^2 is the variance in lifetime production of offspring per parent.

The binomial variance is:

$$\sigma^2 = 2Nc \left(\frac{1}{Nc}\right) \left(1 - \frac{1}{Nc}\right) = \frac{2(Nc - 1)}{Nc}, \quad \text{Equation 3.24}$$

Substituting Equation 3.24 into Equation 3.23 gives, to a very close approximation, $Ne = Nc$.

In the model, the negative binomial distribution is used to determine the number of offspring that each mating produces, the justification for using this distribution can be found in Section 3.2.1.7.2. In programming language R, the negative binomial distribution is described by the mean (named in the model as *number_offspring*), which determines the average number of offspring per mating, and the dispersion parameter (named in the model as *variance_offspring*), which determines the variance of the number of offspring per mating. The variance of this distribution is $mean + (mean^2 / dispersion)$ as described in Bolker (2008, p. 167).

In the model, controlling the number of offspring per mating (*number_offspring*) is convenient because it allows to produce enough offspring in each generation, so populations do not become extinct. However, in the model the mean number of offspring per mating each generation is always equal to two for two reasons: a) the population size remains constant from generation to generation; and b) there is no variance in reproductive success because all individuals reproduce once. By using a large enough value in the parameter *variance_offspring* (e.g., > 1,000), the variance of the negative binomial distribution is very close to two.

To validate that the effective population size (Ne) in the model agrees with theoretical expectations, the model was run for 30 generations and 1,000 replicates, with no dispersal between populations, no selection and a $\text{variance_offspring} = 10,000$. Results of the simulations are compared against the theoretical expectation of the neutral rate of loss of heterozygosity (Equation 3.7) and simulations from the program *EASYPOL* (Balloux, 2001) using the same parameters used in the simulations of the model (Figure 3.9).

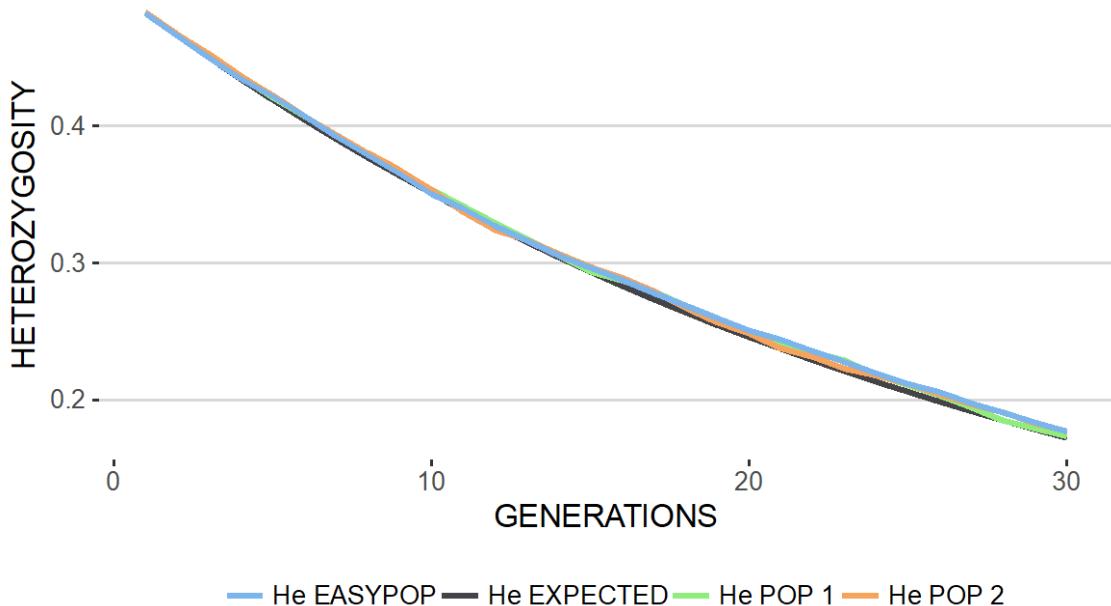


Figure 3.9 | Plot showing that the trajectory of the rate of loss of heterozygosity of the model (*He POP 1* and *He POP 2*) matches both the rate of loss of heterozygosity calculated by the theoretical equation (*He EXPECTED*; Equation 3.7) and the rate of loss of heterozygosity of the simulation program *EASYPOL* (Balloux, 2001).

3.2.3.2 Heterozygote advantage (overdominance)

If the heterozygote has the highest fitness, allele frequencies (p and $q = p - 1$) should reach an equilibrium (Hedrick, 2011, p. 136), as follows:

$$p(\text{allele } a)_{eq} = \frac{s_{aa}}{s_{AA} + s_{aa}}, \quad \text{Equation 3.25}$$

where s_{aa} and s_{AA} are the selection coefficients for the respective homozygous genotypes. The model was tested using the following values: $N\epsilon = 1,000$, $s_{aa} = 0.4$, $s_{AA} = 0.1$. Simulations were run for 100 generations and 100 replicates. Observed values were in agreement with values predicted by theory: mean observed $p(\text{allele } a)_{eq} = 0.79 (+/- 0.0098)$ and expected $p(\text{allele } a)_{eq} = 0.8$.

3.2.3.3 Fixation time for a neutral allele

Following Kimura & Ohta (1969), the probability that a given neutral allele will ultimately fix is its starting proportion (p). The average number of generations until fixation is:

$$\bar{t}_1(p) = \frac{-4 N\epsilon q \ln q}{p}, \quad \text{Equation 3.26}$$

where p = initial frequency of allele "A", q = initial frequency of allele "a" and $N\epsilon$ is effective population size. The model was tested using the following values: $N\epsilon = 100$, $p = 0.5$, $q = 0.5$. Simulations were run for 1,000 generations and 100 replicates. Observed values were in agreement with theory: mean observed $\bar{t}(0.5) = 278 (+/- 150)$ generations and expected $\bar{t}(0.5) = 277$ generations.

3.2.3.4 Directional selection

The recurrence equation for allele proportion is (Halliburton, 2004, p. 136):

$$p(t) = \frac{p^2 w_{AA} + pq w_{Aa}}{p^2 w_{AA} + 2pq w_{Aa} + q^2 w_{aa}}, \quad \text{Equation 3.27}$$

where t is a measure of time in generations, $p(t)$ is allele proportion of allele “A” at generation t , p = proportion of alternative allele “A” at $t - 1$, q = proportion of deleterious allele “a” at $t - 1$, w_{AA} = fitness homozygote “AA”, w_{Aa} = fitness heterozygote “Aa” and w_{aa} = fitness homozygote “aa”. The model was tested for a recessive deleterious allele with the following values: $w_{AA} = 1$, $w_{Aa} = 1$, $w_{aa} = 0.9$, $t = 20$, $p = 0.1$, $q = 0.9$, $\mathbf{Ne} = 1,000$. Simulations were run for 100 replicates. Observed values were in agreement with theory: mean observed $p(20) = 0.36$ and expected $p(20) = 0.36$.

3.2.3.5 Recombination

Decay of linkage disequilibrium (D) as a function of recombination rate (c) can be expressed as follows (Charlesworth & Charlesworth, 2010, p. 380):

$$D_t = (1 - c)^t D_0, \quad \text{Equation 3.28}$$

where t is a measure of time in generations, D_t is LD at generation t , D_0 is initial LD. The model was tested using the following values: $c = 0.01$, $t = 50$, $D_0 = 0.25$, $\mathbf{Ne} = 1,000$. Simulations were run for 100 replicates. Observed values were in agreement with theory: mean observed $D_{50} = 0.15$ and expected $D_{50} = 0.15$.

3.2.4 Simulation experiment

The purpose of the general model is to establish and characterise the evolutionary, demographic and genomic conditions in which linked selection via deleterious alleles impacts genetic diversity. Based on previous and preliminary simulations and on previous research (Zhao & Charlesworth, 2016; Schou *et al.*, 2017; Becher *et al.*, 2020; Gilbert, Pouyet, *et al.*, 2020), the four main parameters on which AOD depends are:

1. Effective population size (N_e).
2. Dominance (h).
3. Selection coefficient (s).
4. Number of deleterious alleles per individual per centiMorgan (del/ind/cM).

Four different values for h , s and del/ind/cM and five values for N_e were tested in the general simulations. These simulations used a Latin square design (a matrix where each cell is a unique combination), to test all the possible combinations of values of these four parameters (a total of 320 combinations).

In the text below are described the parameter values, and their justification, that were used in the general simulations presented in Chapter 2 (Table 3.12).

Table 3.12 Parameters used in the general simulations.

Parameter name	Process	Description	Value
<i>pop_size</i>	Initialisation Section 3.2.1.5	Population size	10, 50, 100, 150, 200
<i>number_loci_under_selection</i>	Initialisation Section 3.2.1.5	Number of loci under selection	2000
<i>number_neutral_alleles</i>	Initialisation Section 3.2.1.5	Number of alleles per neutral locus	9
<i>s_gral</i>	Initialisation Section 3.2.1.5	s of deleterious alleles	0.0001, 0.0005, 0.001, 0.005
<i>h_gral</i>	Initialisation Section 3.2.1.5	h of deleterious alleles	0, 0.1, 0.3, 0.5
<i>q_gral</i>	Initialisation Section 3.2.1.5	Initial frequencies of deleterious alleles	0.1465
<i>dispersal</i>	Dispersal	Whether dispersal occurs	TRUE

	Section 3.2.1.7.1		
<i>number_transfers</i>	Dispersal Section 3.2.1.7.1	Number of dispersing individuals	1, 1, 1, 3, 2
<i>transfer_each_gen</i>	Dispersal Section 3.2.1.7.1	Interval of number of generations in which dispersal occurs	10, 2, 1, 2, 1
<i>number_offspring</i>	Reproduction Section 3.2.1.7.2	Mean number offspring per mating	10
<i>variance_offspring</i>	Reproduction Section 3.2.1.7.2	Coefficient that determines the variance in the number of offspring per mating	1,000
<i>recombination</i>	Recombination Section 3.2.1.7.2	Whether recombination occurs	TRUE
<i>fitness_model</i>	Selection Section 3.2.1.7.3	Fitness model to use	"multiplicative"
<i>selection</i>	Selection Section 3.2.1.7.3	Whether selection occurs	TRUE
<i>selection_model</i>	Selection Section 3.2.1.7.3	Selection model to use	"relative"
<i>Ne</i>	Statistics Section 3.2.1.7.4	<i>Ne</i> value to be used in the equation of the expected F_{ST} (Equation 3.2)	10, 50, 100, 150, 200
<i>Ne_dispersal</i>	Statistics Section 3.2.1.7.4	<i>Ne</i> value to be used in the equation of rate of loss of He (Equation 3.7)	15, 75, 150, 225, 300
<i>map_resolution</i>	Input Data Section 3.2.1.6	Resolution of the recombination map	100,000
<i>c_gral</i>	Input Data Section 3.2.1.6	Recombination rate per region of size <i>map_resolution</i>	0.25, 0.5, 1, 10
<i>windows_gral</i>	Input Data Section 3.2.1.6	Number of rows of the recombination map	50

number_iterations	Overall	Number of simulations replicates	250
gen_number	Overall	Number of generations	62

3.2.4.1 Effective population size (Ne)

In order to maximize the use of available computing power, a $Ne/Nc = 1$ was used. In the general simulations five values of Ne were simulated (10, 50, 100, 150 and 200). These values were chosen to compare results with previous research (Schou *et al.*, 2017) and because these values are representative of real population sizes of many species of conservation concern. I recognise that $Ne/Nc = 1$ is rarely achieved in wild populations (Frankham, 1995), but the processes I simulate are known to be controlled by Ne not Nc (Crow & Kimura, 1970; Frankham, 1995).

3.2.4.2 Dominance coefficient (h)

The main methods to investigate the fitness effects of deleterious alleles are based on the effect that these alleles have on the fitness of heterozygotes *i.e.*, the product of dominance (h) and selection coefficient (s ; Huber *et al.*, 2018). However, different values of h and s can yield similar patterns in the studied population. Given these uncertainties, it is necessary to test a wide range of values of h . Four values of dominance (0, 0.1, 0.3 and 0.5), ranging from semidominant ($h = 0.5$) to completely recessive ($h = 0$), based on reported values (Manna *et al.*, 2011; Huber *et al.*, 2018) were tested in the general simulations. The maximum value that was tested (0.5) was chosen based on Charlesworth (2015): “Much evidence suggests that h for deleterious alleles is mostly nonzero but < 0.5 ”. Zhao and Charlesworth (2016) found that as h increases from 0 to 0.5, the effect of AOD (*i.e.*, higher heterozygosity; He) declines, and that the limit at which the effect of AOD disappears is when $h > 0.37$. A similar pattern was found in the preliminary simulations, as shown in Figure 3.10.

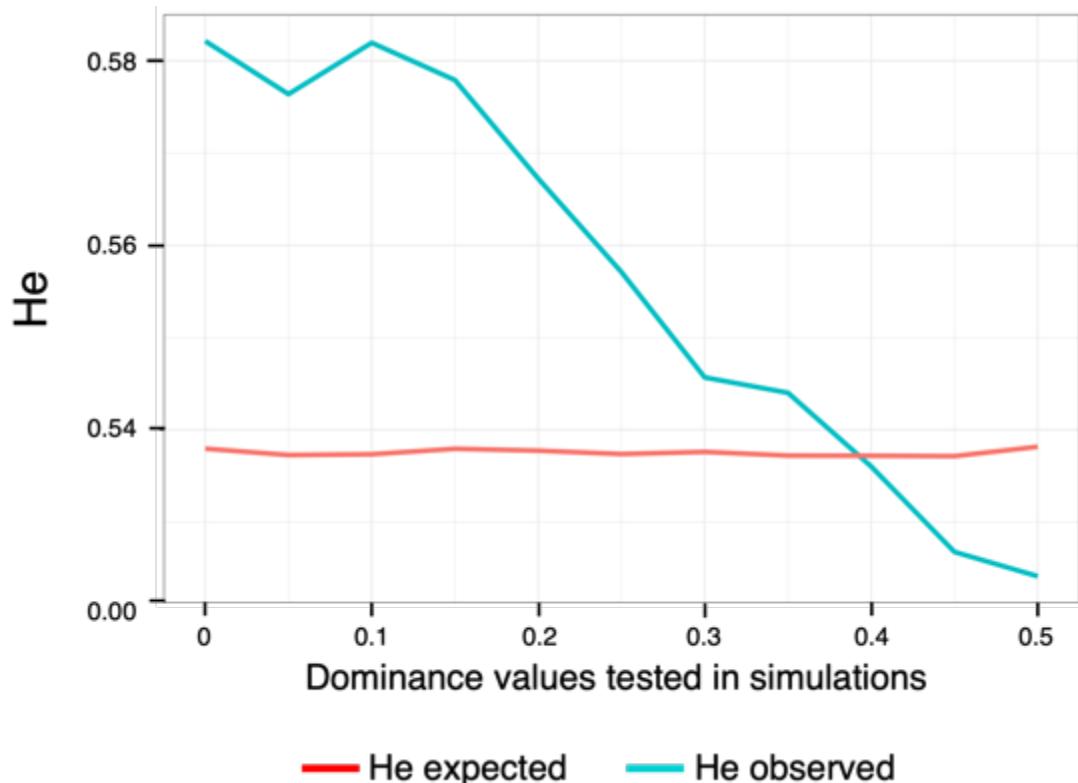


Figure 3.10 | Effects of different values of dominance (h) on heterozygosity (He) in the simulations using a selection coefficient of 0.01.

3.2.4.3 Selection coefficient (s)

Four values of the selection coefficient (s ; 0.0001, 0.0005, 0.001 and 0.005) were tested based on reported values in humans and *D. melanogaster* (Manna *et al.*, 2011; Huber *et al.*, 2017; Huber *et al.*, 2018). These values of s represent arguably the range between two biological extremes of very weak and strong selection. The average s has been estimated to be 70- to 110-fold more deleterious in humans than in *Drosophila* (Huber *et al.*, 2017). The current estimate of mean $h * s$ for new non-synonymous mutations in *D. melanogaster* is ~ 0.001 with a coefficient of variation of ~ 2 (Kousathanas & Keightley, 2013; Charlesworth, 2015).

Preliminary simulations showed that as s increases, the effect of AOD is stronger (*i.e.*, He is higher), as shown in Figure 3.11. Note that the square in the upper right hand of the Figure 3.11 is missing because the simulated populations became extinct.

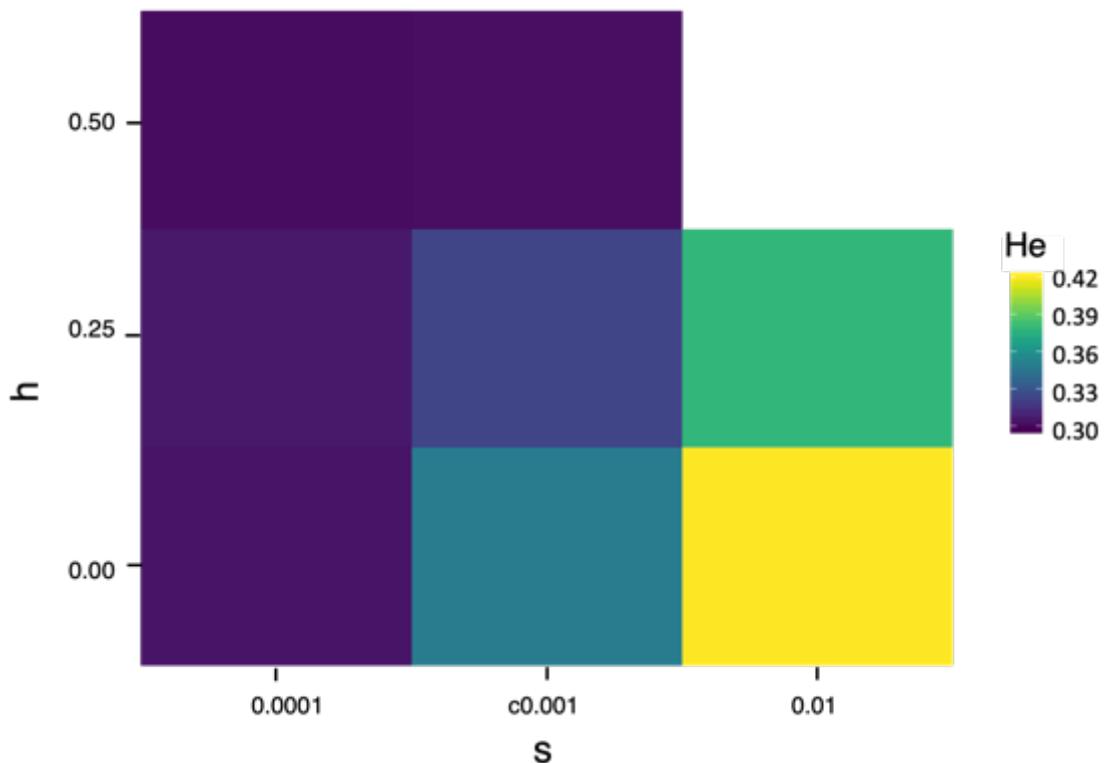


Figure 3.11 | Effects of different values of dominance (h) and selection coefficient (s) on heterozygosity (He) in the simulations.

3.2.4.4 Number of deleterious alleles per individual per centiMorgan (del/ind/cM)

With the aim of simulating realistic recombination rates (c) between loci, I focused on the number of deleterious alleles per individual per centiMorgan (del/ind/cM) rather than on c alone.

Table 3.13 shows realistic values of del/ind/cM in human and *Drosophila*. In the literature, the number of cM is reported as the mean of the two sexes. Therefore, in Table 3.13, the total number of cM in *Drosophila* is divided by two because recombination only occurs in females.

Table 3.13 Realistic values of number of deleterious alleles per individual.

Species	Human	<i>Drosophila</i>	Domestic yak	Wild yak
del/ind	800-1,000 ^{1,9}	5,000 ⁵	1,202–1,470 ¹⁰	974 –1,019 ¹⁰
cM	3,615 ²	277/2 ⁶	2,435 ⁸ (cow)	2,435 ⁸ (cow)
Coding genes	20,418 ³	13,931 ³	21,880 ³ (cow)	21,880 ³ (cow)
Exons	233,785 ⁴	88,403 ⁷	193,892 ³ (cow)	193,892 ³ (cow)
del/ind/cM	0.27	36.2	0.6	0.42
del/ind/gene	0.041	0.359	0.067	0.047
del/ind/exon	0.004	0.057	0.007	0.005

¹ (Charlesworth & Charlesworth, 2010)

² (Kong *et al.*, 2002)

³ (Zerbino *et al.*, 2017)

⁴ (Sakharkar *et al.*, 2004)

⁵ (Haddrill *et al.*, 2010)

⁶ (Comeron *et al.*, 2012)

⁷ (Lee *et al.*, 2004)

⁸ (Ma *et al.*, 2015)

⁹ (Chun & Fay, 2009)

¹⁰ (Xie *et al.*, 2018)

cM; centiMorgans

del; deleterious alleles

ind; individual

Even though del/ind/cM in humans is small compared to *Drosophila*, loci in the human genome are relatively more clumped than in *Drosophila*, this means than in some genome regions del/ind/cM in humans would be much larger than the mean. The following four values of del/ind/cM were tested in the simulations: 1, 10, 20 and 40. The explanation about how these values were controlled is described in the next paragraphs.

In the model, each individual will carry the deleterious allele in just a fraction of the total number of loci to be simulated because the frequency of the deleterious allele (q) is < 1 . We can use the Hardy Weinberg equation ($p^2 + 2pq + q^2 = 1$, where p is the frequency of the alternative allele and the q is the frequency of the deleterious allele) to demonstrate this fact. For instance, in a simulation where a chromosome contains 2,000 loci with deleterious alleles, with a mean value of $q = 0.15$, on average 25.5% of the individuals would be heterozygotes ($2pq$) and 2.25% would be homozygotes (q^2) for the deleterious allele. In this example, on average each individual would carry the deleterious allele in 27.75% (25.5% + 2.25%) of the 2,000 loci, that is each individual would carry deleterious alleles in 555 loci.

In the simulations, as noted above, del/ind/cM depends on three variables: the initial frequency of the deleterious allele (q), the total number of loci with deleterious alleles to be simulated and the length of the chromosome in centiMorgans. Therefore, two of these three variables should be maintained constant in the simulations to isolate the effect that del/ind/cM might have on AOD. In the general simulations q ($q = 0.15$) and the total number of loci with deleterious alleles that were simulated ($n = 2,000$) were kept constant across all the combinations. Four chromosome lengths in centiMorgans were tested: 12.5, 25, 50 and 500. Further justification for the values used for these parameters is presented in the following paragraphs.

Number of loci with deleterious alleles to be simulated. According to Zhao & Charlesworth (2016) and previous simulations that were performed, the magnitude of AOD is proportional to genetic load. In this simulation context, this means that the greater the number of loci under selection to be simulated or the higher the product of $q * h * s$ are, the stronger the effect of AOD will be. This also means that if too few loci are

used, AOD would not occur. A constant value of 2,000 loci with deleterious alleles were simulated for all the scenarios. Using 2,000 loci enables realistic scenarios in terms of genetic load and del/ind/cM across the ranges of the different values for q , h and s that were tested in the general simulations (Table 3.14).

Table 3.14 Tabulation of values of genetic load and del/ind/cM for different combinations of values of q , h and s that were tested in the general simulations.

Number of loci with deleterious alleles	Initial frequency of deleterious allele (q)	Length of the chromosome (cM)	h	s	del/ind/cM	Genetic load (Equation 3.21)
2000	0.1465	12.5	0.00	0.0001	40	0.06
2000	0.1465	25	0.00	0.0001	20	0.06
2000	0.1465	50	0.00	0.0001	10	0.06
2000	0.1465	500	0.00	0.0001	1	0.06
2000	0.1465	12.5	0.10	0.0005	40	0.27
2000	0.1465	25	0.10	0.0005	20	0.27
2000	0.1465	50	0.10	0.0005	10	0.27
2000	0.1465	500	0.10	0.0005	1	0.27
2000	0.1465	12.5	0.30	0.0010	40	0.52
2000	0.1465	25	0.30	0.0010	20	0.52
2000	0.1465	50	0.30	0.0010	10	0.52
2000	0.1465	500	0.30	0.0010	1	0.52
2000	0.1465	12.5	0.50	0.0050	40	0.98
2000	0.1465	25	0.50	0.0050	20	0.98
2000	0.1465	50	0.50	0.0050	10	0.98
2000	0.1465	500	0.50	0.0050	1	0.98

Initial allele frequency of the deleterious allele (q). Zhao and Charlesworth (2016) showed that q has a moderate effect on the strength of AOD. The same pattern is seen in preliminary simulations, as shown in Figure 3.12 and Figure 3.13. Note that squares in the upper right hand of the figures are missing because the simulated populations became extinct. A constant value of $q = 0.15$ for all the combinations of values was used to isolate the effect that del/ind/cM might have on AOD. Note that a q value = 0.15 is within the range of the frequency of deleterious alleles per gene that would be normally found in *Drosophila* and human individuals (Table 3.13, column del/ind/cM). Note that a q value = 0.15 is within the range of the frequency of deleterious alleles per gene that

would be normally found in Drosophila and human individuals (Supplementary information Table 3.13, column del/ind/cM).

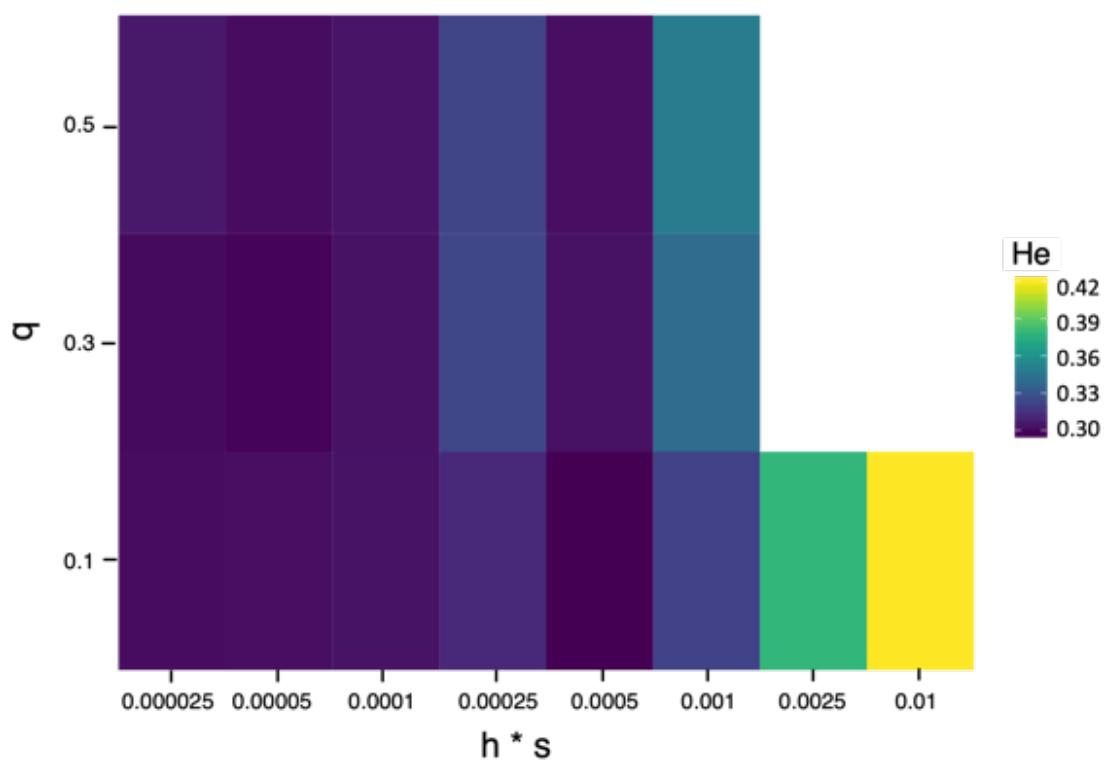


Figure 3.12 | Effects of different values of dominance (h) and selection coefficient (s) and initial frequency of the deleterious allele (q) on heterozygosity (He) in the simulations.

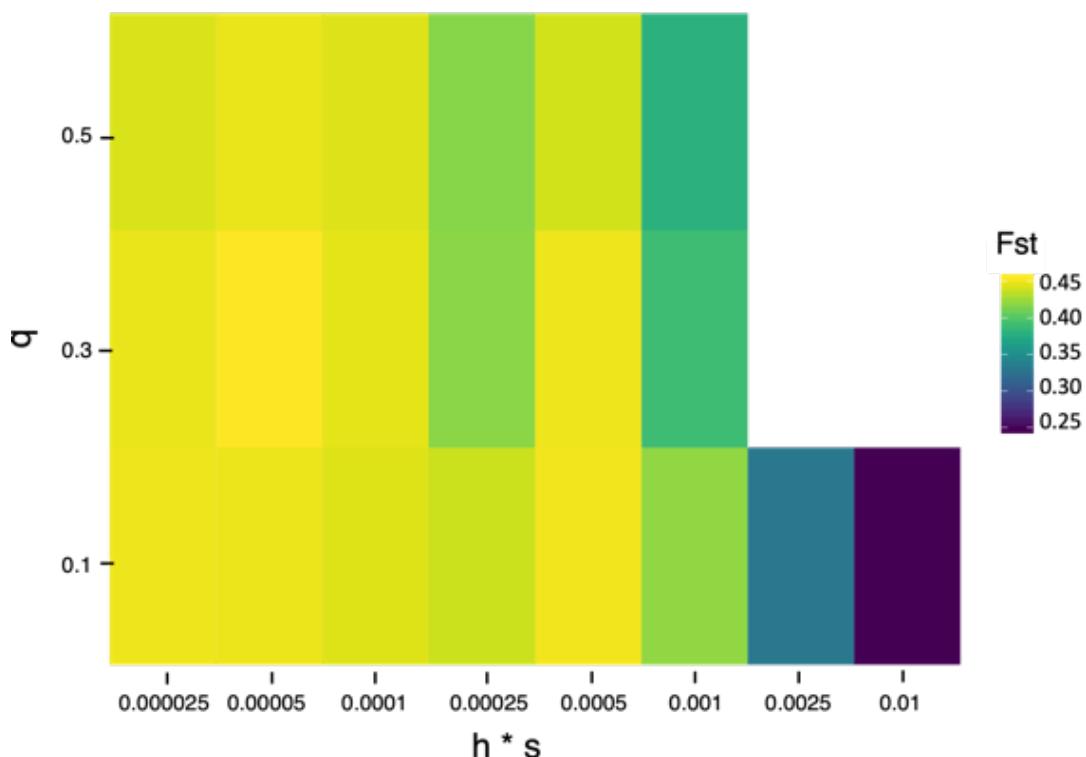


Figure 3.13 | Effects of different values of dominance (h) and selection coefficient (s) and initial frequency of the deleterious allele (q) on F_{ST} in the simulations.

3.2.4.5 Number of offspring per pair of parents

In the model, the lower the mean number of offspring per pair of parents, the stronger the effect of AOD (i.e., lower F_{ST} and higher He), as shown in Figure 3.14. A constant value of 10 offspring per pair of parents was used for all the combinations.

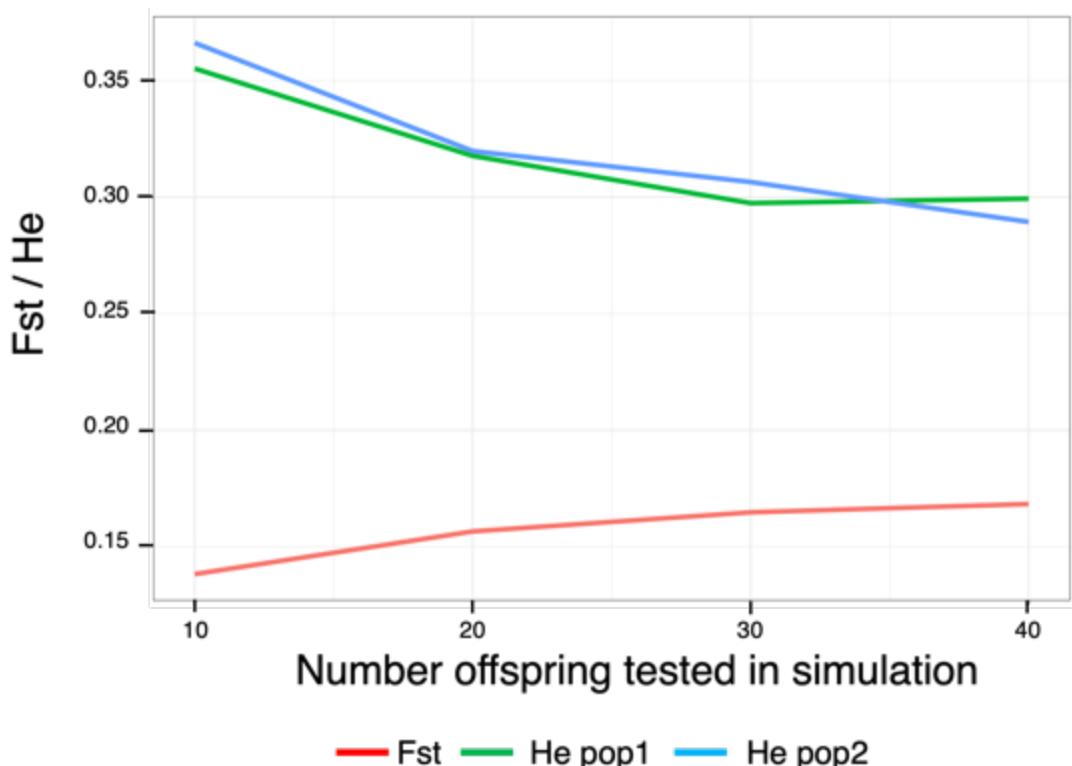


Figure 3.14 | Effects of different values of mean number of offspring per pair of parents on heterozygosity (He) and F_{ST} in the simulations.

3.2.4.6 Variance in the number of offspring per mating

A value of 10,000 for the parameter ***variance_offspring*** was used, which controls the variance of the number of offspring per mating. This value gives a variance of 2 in the number of offspring per mating to obtain a $Ne/Nc = 1$, as discussed in the validation of the Ne/Nc ratio section 3.2.3.1. Figure 3.15 ($Ne = 150$) and Figure 3.16 ($Ne = 10$) show that in neutral simulations, using a value of 10,000 for the parameter

variance_offspring, the rate of loss of heterozygosity per generation of the simulations agreed with the expected rate of loss of heterozygosity under neutrality (Equation 3.7).

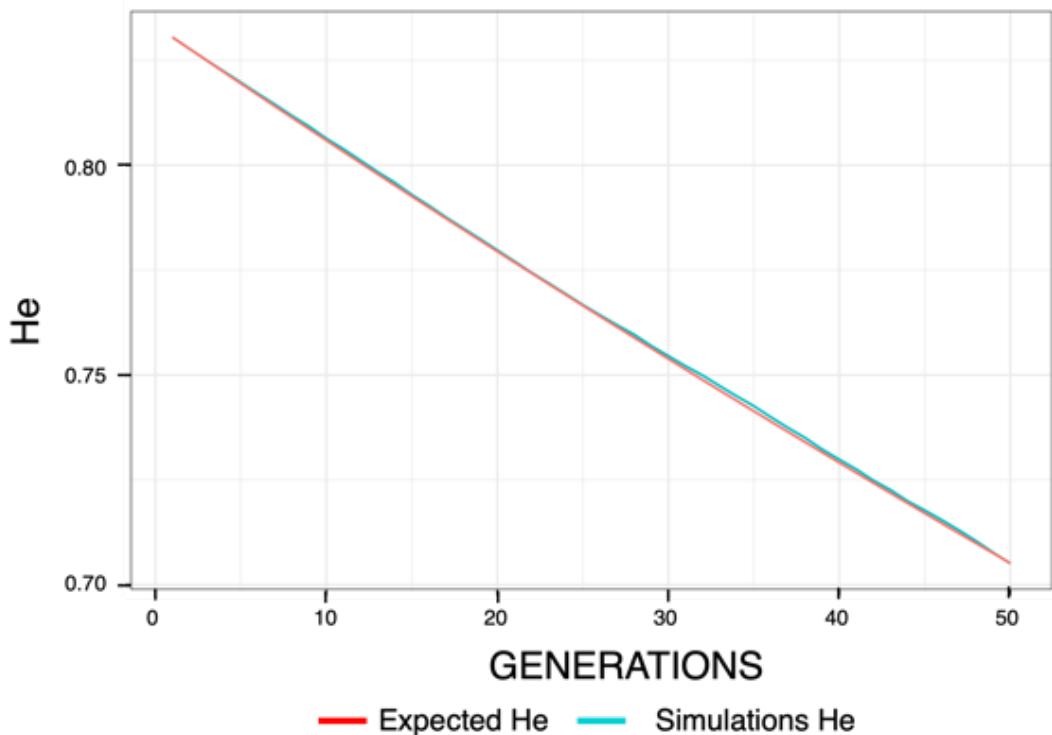


Figure 3.15 | Trajectories of heterozygosity (H_e) over generations in neutral simulations using an $N_e = 150$ and a value of the parameter *variance_offspring* of 10,000 (blue line) and the equation of the neutral rate of loss of heterozygosity (red line; Equation 3.7).

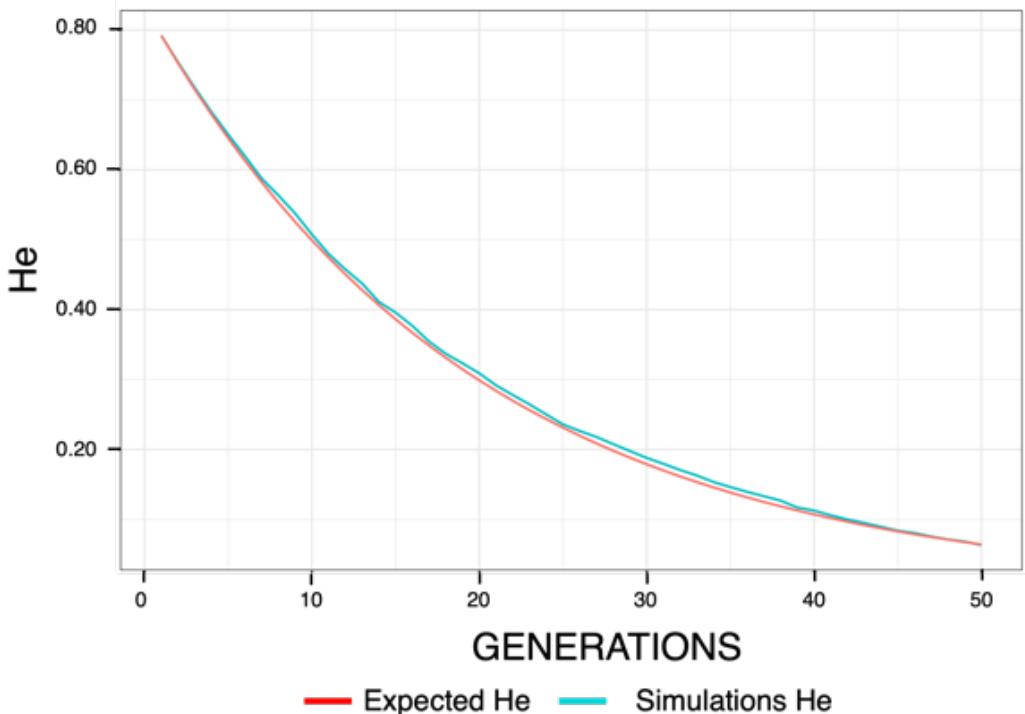


Figure 3.16 | Trajectories of heterozygosity (H_e) over generations in neutral simulations using an $N_e = 10$ and a value of the parameter ***variance_offspring*** of 10,000 (blue line) and the equation of the neutral rate of loss of heterozygosity (red line; Equation 3.7).

3.2.4.7 Dispersal rate

A constant dispersal rate of 0.01 was used for all the combinations. This value was chosen for the operationally feasible range of population sizes and dispersal rates (Table 3.15). The number of generations in Table 3.15 were calculated as the expected time at which populations connected by symmetrical gene flow reach half of drift-dispersal equilibrium F_{ST} (T1) and twice this period (T2) based on the following equation (Whitlock, 1992):

$$t_{1/2} = \frac{\ln(1/2)}{\ln\left[(1-m)^2\left(1 - \frac{1}{2N_e}\right)\right]}, \quad \text{Equation 3.29}$$

where $t_{1/2}$ is the expected time to reach the half drift-dispersal equilibrium F_{ST} value, and m is the dispersal rate. Expected equilibrium F_{ST} was calculated using equation Equation 3.2 and dispersal rate (m ; fraction of individuals in a population that is composed of immigrants; Holsinger, 2020, p. 93) was calculated as (dispersers per X generation / dispersal in X generation) / N_e or (***number_transfers / transfer_each_gen***) / (***population_size_dispersal***). Figure 3.17 to Figure 3.20 show how F_{ST} in the simulations reaches its equilibrium approximately at the generation predicted by the theoretical equation.

Table 3.15 Parameters explored to calculate the number of generations to reach drift-dispersal equilibrium in the simulations. The parameters that were used in the general simulations are highlighted in grey.

Ne	Dispersers per X generation	Dispersal in X generation	Dispersal Rate	T1 (generations)	T2 (generations)	Expected F_{ST}
10	2	5	0.0400	6	12	0.135
50	2	1	0.0400	8	16	0.030
150	6	1	0.0400	9	18	0.010
300	12	1	0.0400	9	18	0.005
10	1	5	0.0200	8	16	0.238
50	1	1	0.0200	14	28	0.059
150	3	1	0.0200	16	32	0.020
300	6	1	0.0200	17	34	0.010
10	1	10	0.0100	10	20	0.385
50	1	2	0.0100	23	46	0.111
100	1	1	0.0100	28	56	0.059
150	3	2	0.0100	30	60	0.040
200	2	1	0.0100	31	62	0.030
10	1	20	0.0050	12	24	0.556
50	1	4	0.0050	35	70	0.200
150	3	4	0.0050	52	104	0.077
300	3	2	0.0050	60	120	0.040
10	1	30	0.0033	12	24	0.652
50	1	6	0.0033	42	84	0.273
150	2	4	0.0033	70	140	0.111
300	2	2	0.0033	84	168	0.059

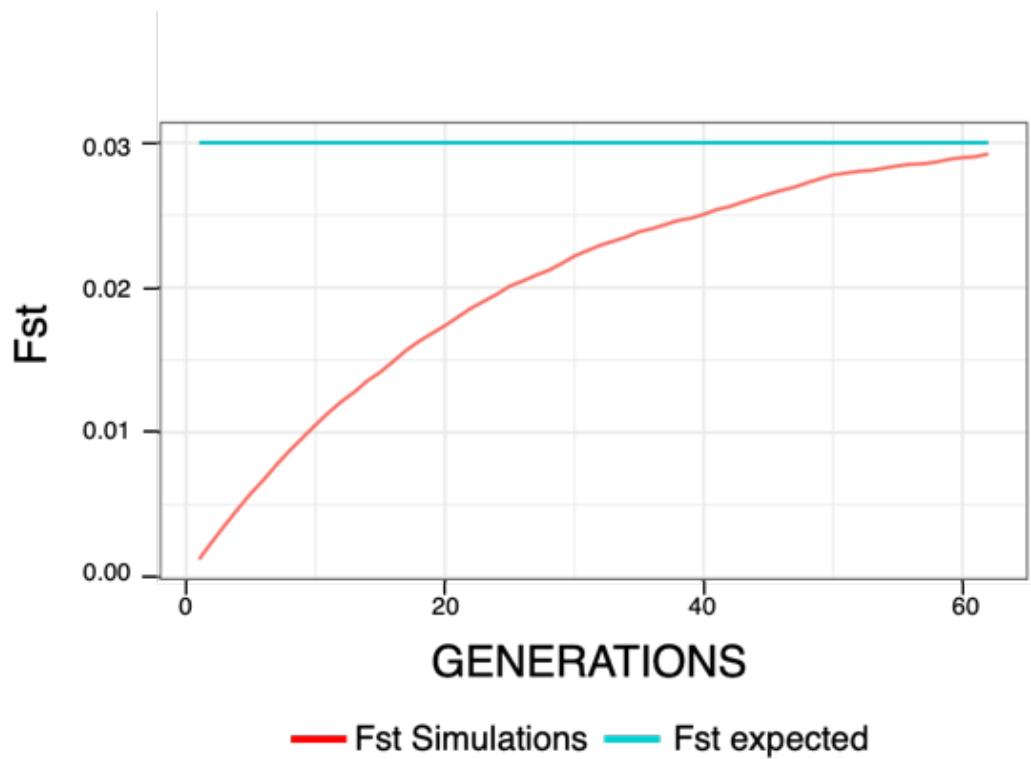


Figure 3.17 | Plot showing that the trajectory of F_{ST} over generations in the simulations using an $N_e = 200$ and a dispersal rate = 0.01 (red line) reaches an equilibrium as predicted by the theoretical F_{ST} equation (Equation 3.2; blue line).

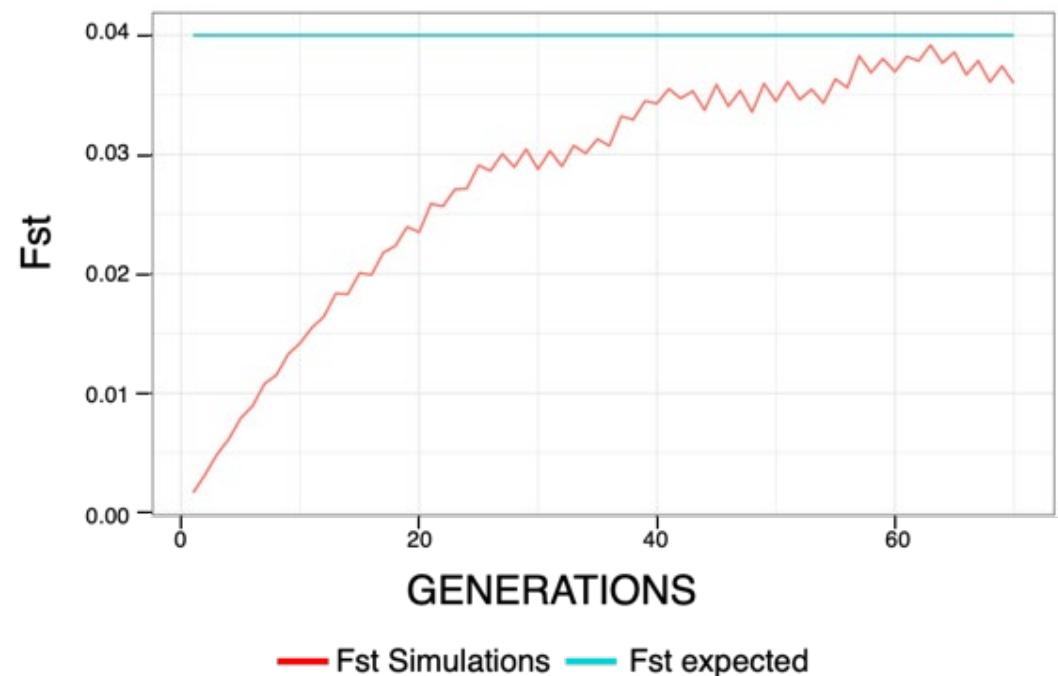


Figure 3.18 | Plot showing that the trajectory of F_{ST} over generations in the simulations using an $N_e = 150$ and a dispersal rate = 0.01 (red line) reaches an equilibrium as predicted by the theoretical F_{ST} equation (Equation 3.2; blue line).

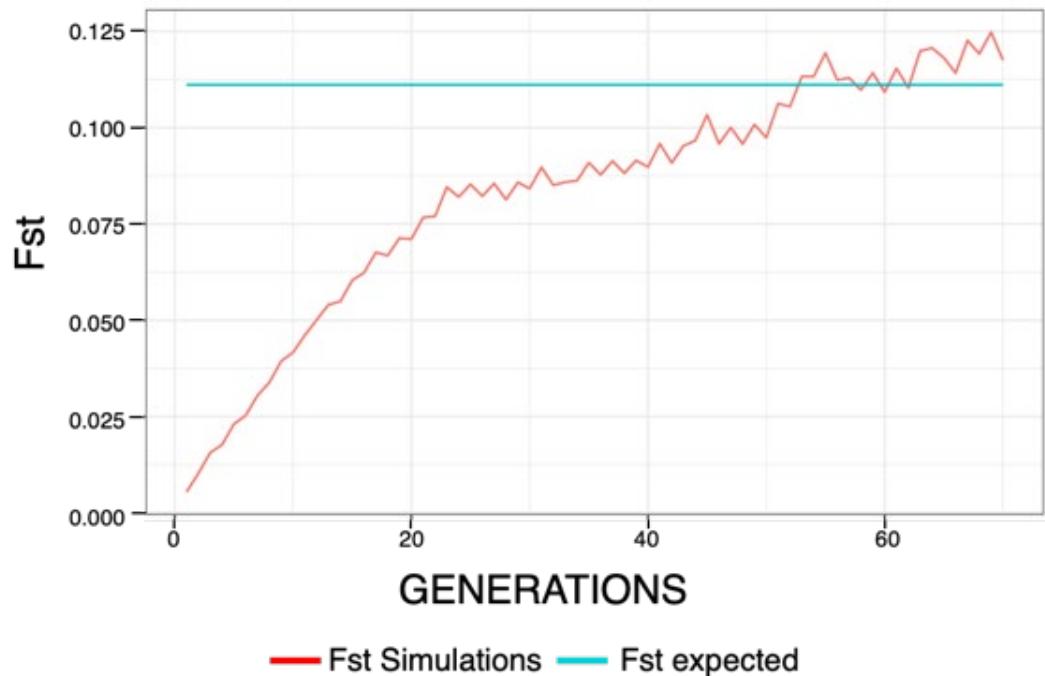


Figure 3.19 | Plot showing that the trajectory of F_{ST} over generations in the simulations using an $N_e = 50$ and a dispersal rate = 0.01 (red line) reaches an equilibrium as predicted by the theoretical F_{ST} equation (Equation 3.2; blue line).

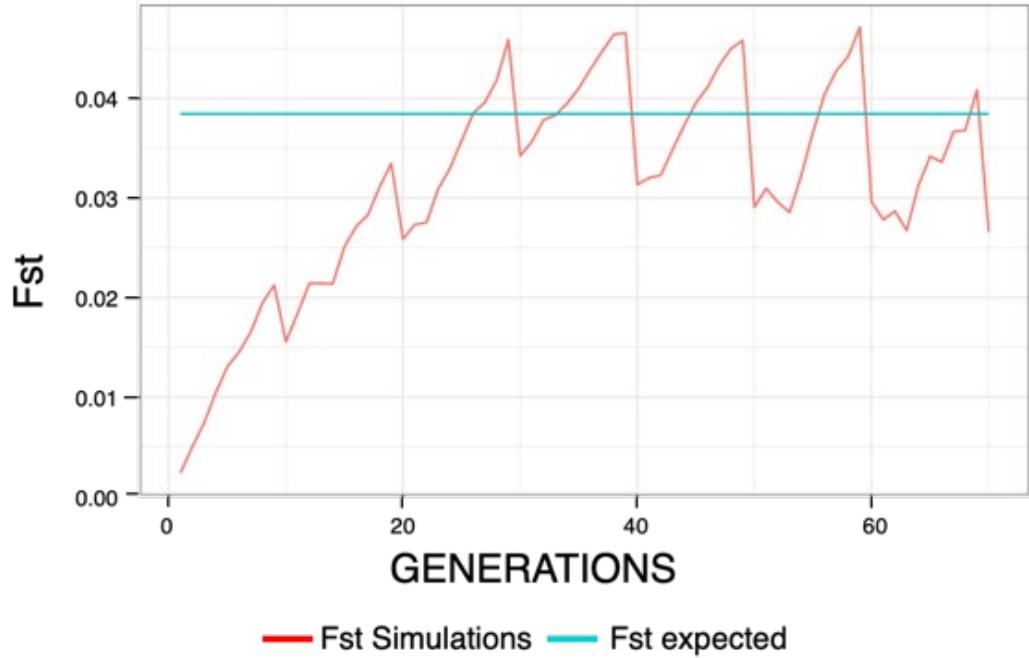


Figure 3.20 | Plot showing that the trajectory of F_{ST} over generations in the simulations using an $N_e = 10$ and a dispersal rate = 0.01 (red line) reaches an equilibrium as predicted by the theoretical F_{ST} equation (Equation 3.2; blue line). The saw-toothed appearance of the trajectory of F_{ST} in the simulations is due to the intermittent dispersal (one individual every tenth generation; Table 3.15).

3.2.4.8 Number of generations

In all the combinations, populations reproduce during 62 generations which is the number of generations required to reach drift-dispersal equilibrium for the largest Ne (200) that was tested (see Table 3.15).

3.2.4.9 Number of neutral loci and number of neutral alleles

A chromosome of 5 Mbp (*windows_gral* = 50 and *map_resolution* = 100,000 bp) with one neutral locus every 100 Kbp (for a total of 50 neutral loci) was used in all the combinations (see Section 3.2.1.5). In neutral simulations, heterozygosity and F_{ST} increase as the number of alleles per neutral locus increases, as shown in Figure 3.21 and Figure 3.22, where different number of alleles were simulated. The values of heterozygosity and F_{ST} correspond to the 50th generation. Nine alleles with equal frequencies were used in all the combinations of the general simulations.

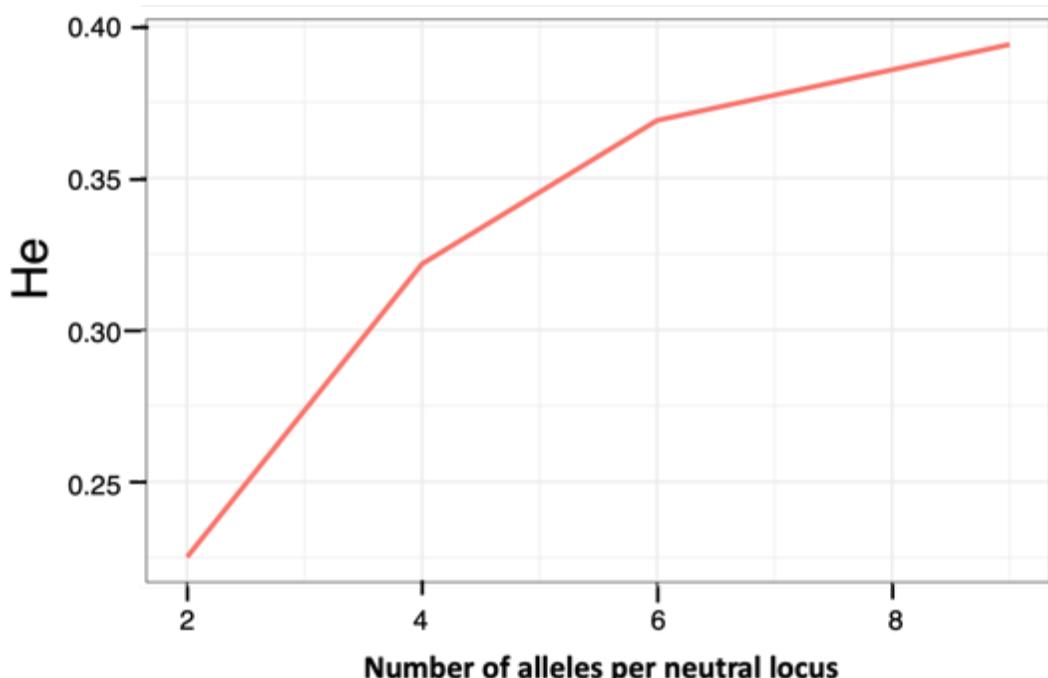


Figure 3.21 | The effect of the number of alleles on heterozygosity (He) in neutral simulations.

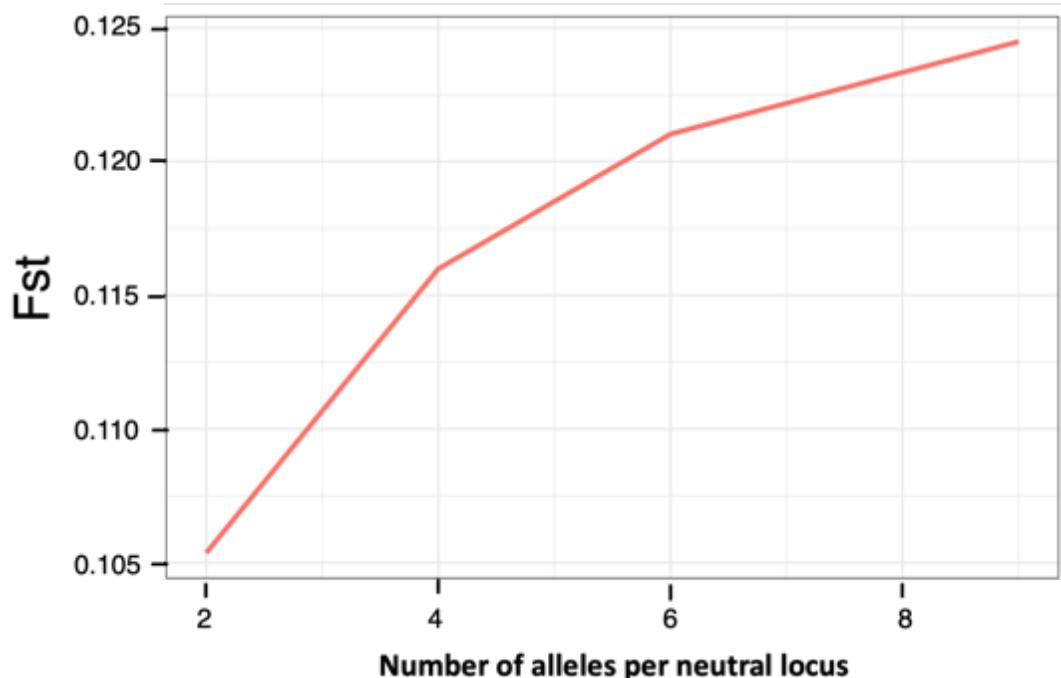


Figure 3.22 | The effect of the number of alleles on F_{ST} in neutral simulations.

3.2.4.10 Recombination

To generalize the results as much as possible, in the simulations recombination occurs in males and females, as in most species. Description of the recombination process can be found in Section 3.2.1.7.2.

3.2.4.11 Fitness model

It has been suggested that a multiplicative fitness model fits observed data significantly better than alternative models (Lobkovsky *et al.*, 2019). Therefore, in the general simulations the multiplicative fitness model (***fitness_model*** = “multiplicative”) was used. For the description of fitness models see Section 3.2.1.7.3.

3.2.4.12 Selection model

In most real situations, it is likely that the action of natural selection is a combination of both selection models, absolute and relative. However, probably highly prolific organisms with low investment/parental care might be more subject to soft selection which is better modelled by the relative fitness model, while organisms that are less prolific and have a high investment/parental care might be subject to hard selection

which is better modelled by the absolute fitness model (Gilbert *et al.*, 2018). Description of the selection process can be found in Section 3.2.1.7.3.

In the general simulations, the effect of AOD is stronger when using the relative selection model compared to the absolute selection model (Figure 3.23). Additionally, when using the absolute selection model, populations became extinct if the number of loci under selection to be simulated or the product of $q * h * s$ load was too high. In order to test all the different combinations without the population becoming extinct, the relative selection model was used for all the combinations.

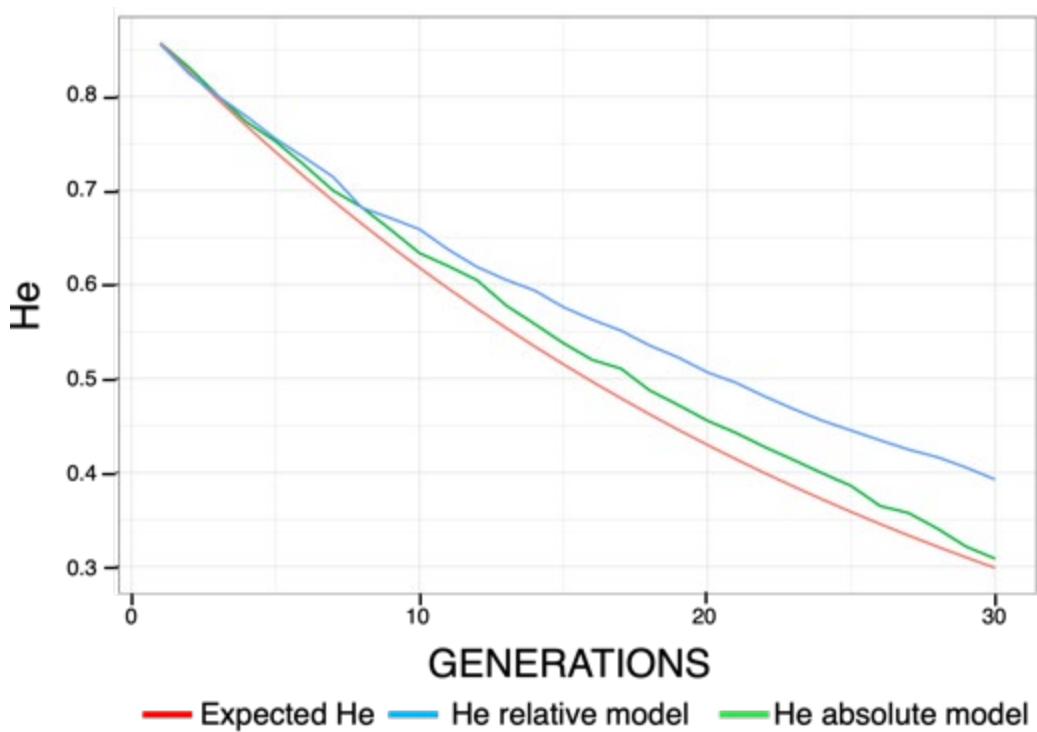


Figure 3.23 | Effect of using different selection models on heterozygosity (H_e) in the simulations.

3.2.4.13 Number of iterations

Power analyses were performed to decide how many replicates of every simulation scenario should be run. For this analysis, I followed Cohen (2013) statistical power analysis approach with the help of the R package *pwr* (Champely *et al.*, 2018). These power analyses were performed using the results of preliminary simulations of all the possible combinations of the following parameters values:

- Dominance (h): 0, 0.25, 0.5.
- Selection coefficient (s): 0.01, 0.001, 0.0001.
- Initial allele frequency of the deleterious allele (q): 0.1, 0.3, 0.5.

I focused on a power of 0.8 to detect a target difference of +/- 0.05 of the statistics (Table 3.16) between a given simulation scenario and the neutral expectation with a significant level of 0.05. First, from the preliminary simulations, the mean standard deviation for each statistic was calculated. The standard deviations of the statistics calculated among the different combinations of values were similar (Table 3.16). Then the effect size was calculated by dividing the target difference between the simulation scenario and the neutral expectation (0.05) by the standard deviation of each statistic. Finally, the mean effect size of the statistics was used to calculate in *pwr* the number of replicates required.

Table 3.16 Standard deviation (sd) and effect size of the statistics calculated in the simulations.

Statistic	Standard deviation (sd)	Difference between means	Effect size
Shannon differentiation	0.30	0.05	0.17
Shannon index	0.35	0.05	0.14
Mutual information	0.21	0.05	0.24
Heterozygosity	0.22	0.05	0.23
Fst	0.24	0.05	0.21
Fst corrected for sample size	0.24	0.05	0.21
Jost's D	0.32	0.05	0.16
Mean	0.28		0.18

The power analysis indicated that with a mean effect size of 0.18, 249 replicates are needed to detect a difference of +/- 0.05 between the simulation result and the neutral expectation with a power of 0.8 at a significance level of 0.05, as shown in Figure 3.24 and Figure 3.25. 250 replicates were used in the general simulations.

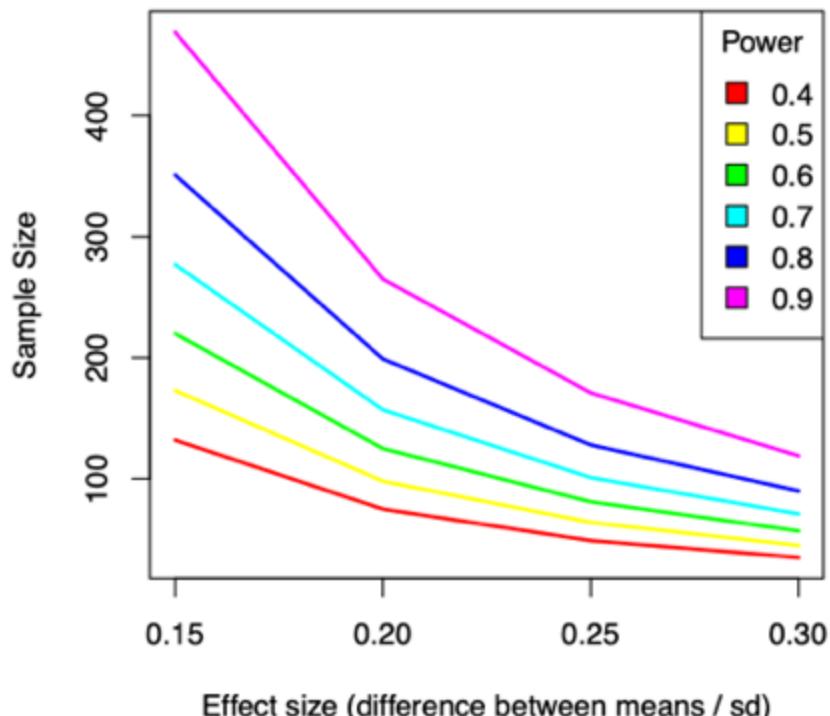


Figure 3.24 | Effects of different values of sample size and effect sizes on the power of identifying a difference between the genetic statistics produced by simulations and the expected value.

One-sample t test power calculation

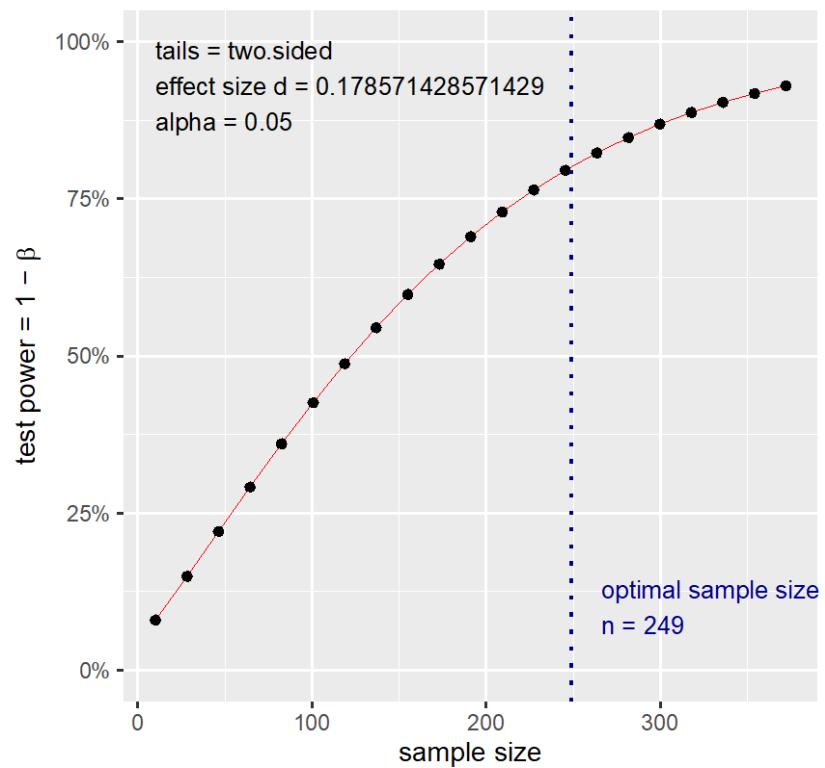


Figure 3.25 | Sample size and power to detect a difference between simulation and neutral expectation (Sig = alpha = 0.05; two tailed).

3.3 FLY SIMULATIONS

The general model serves as foundation for the fly model. To avoid repetition of information, the general model is described in detail in Section 3.2, and in this section only the issues that are different from the general model are stated.

3.3.1 Model description

3.3.1.1 Purpose

The purpose of the fly model is to establish whether realistic parameter values can explain patterns of genetic variation within and between populations observed in the empirical fly experiment (Chapter 2), and determine whether AOD is a probable explanation for these findings:

- Mean H_e in the high dispersal regime was 61% (coefficient of variation 30%) higher than expected, in the moderate dispersal regime it was 69% (cv 12%) higher than expected and in the low dispersal regime it was 70% (cv 10%) higher than expected.
- Mean F_{ST} in the high dispersal regime was 61% (cv 30%) lower than expected, while in the moderate dispersal and low regimes it was 69% (cv 12%) to 70% (cv 10%) lower than expected, respectively.

3.3.1.2 Process overview and scheduling

The fly model involves four main processes: loading of input data (performed once at the beginning of the simulations), initialization (performed at the beginning of each simulation replicate), a pre-adaptation phase (with two submodels) and a dispersal phase (with four submodels; see Figure 3.26). In the fly model, the simulations start with a pre-adaptation phase with the aim to approximate the empirical population history of the fly experiment (see methods in Chapter 2 for a description). In this phase, two large isolated populations (*i.e.*, without dispersal), reproduce during several non-overlapping generations and remain with a constant population size.

In the pre-adaptation phase two processes are executed every generation in the following order:

1. Sexual reproduction and recombination only occurring in females.
2. Natural selection only on the loci under selection (selection of offspring to become the parents of the next generation).

After the pre-adaptation phase ends, the dispersal phase starts where two new populations are formed sampling a subset of the individuals from the two original isolated populations. Individuals can be sampled for both populations from the same original isolated population or from different original isolated populations, as in the fly experiment (see Chapter 2 Extended Data Figure 2.1). The two newly formed populations have a constant population size, reproduce in non-overlapping generations and are connected by symmetric dispersal. In the dispersal phase, four processes are executed every generation in the following order:

1. Dispersal.
2. Sexual reproduction and recombination only occurring in females.
3. Natural selection only on the loci under selection (selection of offspring to become the parents of the next generation).
4. Calculation of genetic statistics.

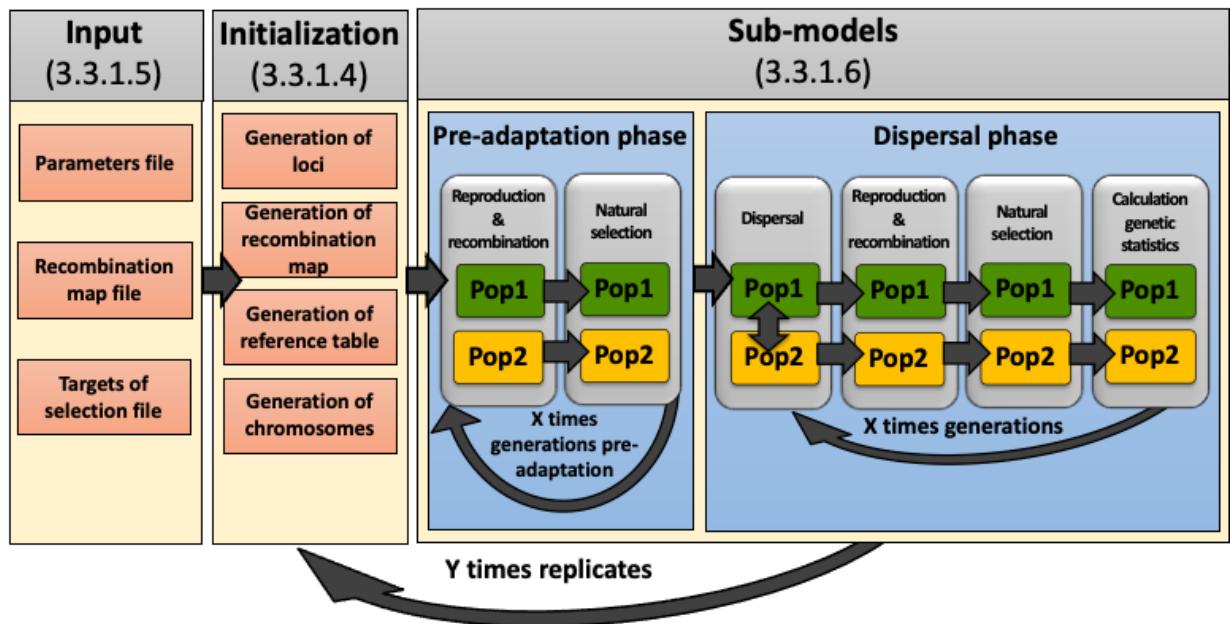


Figure 3.26 | Schematic representation of the simulation processes in the fly simulations. Information within parentheses correspond to the section in which the process is described.

3.3.1.3 Design concepts

3.3.1.3.1 Stochasticity

Justification for the use of stochasticity to determine the values in Table 3.17 can be found in the full description of each process.

Table 3.17 Points in the model on which stochasticity is used to determine values and sections where these are justified.

Values determined by stochasticity	Process of the model where stochastic values are used	Description
The value of the selection coefficient (s) for each locus under selection.	Initialisation Section 3.3.1.4	Sampled from a log normal distribution or a gamma distribution. $s = 0$ means that allele has no effect on fitness and $s = 1$ means allele is lethal.
The value of the dominance coefficient (h) for each locus under selection.	Initialisation Section 3.3.1.4	Sampled from a normal distribution. $h = 0$ is completely recessive and $h = 1$ is completely dominant.
The probability of sampling the deleterious allele “a” (q) or the alternative allele “A” ($1-q$) for each locus under selection when chromosomes are generated.	Initialisation Section 3.3.1.4	Determined by solving q from the following equation: $s(1-2h)q^2+hs(1+u)q-u=0$, where u is the mutation rate per generation per site (Crow & Kimura, 1970, pp 260). Note that the u value is only used in this equation.
The probability of a recombination event occurring in a specific chromosome location per generation per individual.	Reproduction/rec ombination Section 3.3.1.6.2	Determined by the recombination rate in each locus.

3.3.1.4 Initialization (Table 3.18)

The description of each process begins with a table containing a small description of the parameters that can be controlled in each process. However, the full explanation of the parameters and how they are used in the model are described in the text below the table and parameter names appear in an ***italic bold*** format. For clarity, I used the same parameter names as used in the actual code of the model. The names of the parameters were chosen to be self-explanatory.

Table 3.18 Parameters used in the initialization process.

Parameter name	Description	Range of values
<i>chromosome_name</i>	Name of the chromosome to be simulated	Defined by the user (must be character)
<i>targets_factor</i>	Factor to sample the number of loci under selection from the input file "targets_of_selection.csv"	Defined by the user
<i>map_resolution</i>	Resolution of the recombination map (bp)	Dependent on genomic resources
<i>location_loci_real_data</i>	Locations of loci from a real dataset	Defined by the user
<i>freq_loci_real_data</i>	Frequency of each one of the alleles of the loci from a real dataset	Defined by the user
<i>s_distribution</i>	Name of the distribution to use to sample the values of the selection coefficient (<i>s</i>)	"gamma" or "log_normal"
<i>gamma_shape</i>	Shape of the gamma distribution from where <i>s</i> values are sampled	Defined by the user
<i>gamma_scale</i>	Scale of the gamma distribution from where <i>s</i> values are sampled	Defined by the user
<i>log_mean</i>	Mean of the log normal distribution from where <i>s</i> values are sampled	Defined by the user
<i>log_sd</i>	Standard deviation of the log normal distribution from where <i>s</i> values are sampled	Defined by the user
<i>h_method</i>	Method to determine the values of the dominance coefficient (<i>h</i>)	"distribution" or "equation"

<i>theta_rate</i>	Value for the variable rate of the dominance equation to calculate h	Defined by the user
<i>intercept</i>	Value for the intercept of the equation to determine h (maximum value of h when $s = 0$)	0-1
<i>dominance_mean</i>	Mean of the normal distribution from where h values are sampled	0-1
<i>dominance_sd</i>	Standard deviation of the normal distribution from where h values are sampled	Defined by the user
<i>mutation_rate</i>	Mutation rate per generation per site. Value only used in the equation to determine q	Defined by the user
<i>number_neutral_alleles</i>	Number of alleles per neutral loci. The same initial frequency (p) is used for each allele	1-9
<i>pop_size</i>	Population size (must be even)	Dependent on computing resources

3.3.1.4.1 Generation of loci, assignment of their genetic location (cM) and physical location (bp) and generation of the recombination map

To generate loci and assign their locations, the model first uploads the targets of selection input file: “targets_of_selection.csv”. Details about how this input file is generated can be found in the Section Input data (3.3.1.5.2). The file of targets of selection contains all the coding sequences (CDS) of the study organism and the number of potential targets for selection in each CDS. This file contains five columns:

1. Identification number of each CDS.
2. Chromosome name at which each CDS belongs.
3. CDS start point (bp).
4. CDS end point (bp).
5. Number of potential targets of selection.

Next, the model extracts from the file of targets of selection those CDS that belong to the specific chromosome to be simulated (*chromosome_name*). To determine the number of loci under selection to be simulated and their physical location (bp), the model samples random locations within each CDS equal to ***targets_factor*** * number of potential targets of selection contained in each CDS.

To determine the number of neutral loci to be simulated and their physical location (bp), the model extracts the end point of the last CDS to be used as the total length of the chromosome (bp). Then, one neutral locus is created every 100 Kbp starting from 50 Kbp.

A recombination map is generated by using the results of the cross multiplications to be used as input for the recombination process which is explained further below. To assign a genetic location (cM) to each locus (neutral and under selection), the model uploads the recombination map input file: “live_pops_recom_map.csv”. Details about how this input file is generated can be found in the section: Input data (Section 3.3.1.5.1). The recombination map is a table where each row contains the number of cM occurring in every chromosome region of size = *map_resolution*. Then, the physical distance (bp) between each pair of contiguous loci is calculated. Next, the number of cM between each pair of loci is obtained by cross multiplication: (physical distance between pair of loci * number of cM in that region) / *map_resolution*. Finally, a cumulative sum (a vector whose elements are the cumulative sums of the elements) of the cross multiplications is used to determine the genetic location (cM) of each locus. To account for the case when the probability of the total recombination rate within a chromosome is less than an integer the same approach is used as in the general simulations (Section 3.2.1.5.1).

Additionally, the user can load into the model the real location and frequency of loci from an empirical dataset through the parameters: ***location_loci_real_data*** and ***freq_loci_real_data*** (see for input example the bottom of Table 3.23).

3.3.1.4.2 Generation of the reference table

The reference table has the same characteristics as in the general simulations (Section 3.2.1.5.2), however in the fly simulations the procedure differs in how q , s and h are determined.

Selection coefficient (s) values for each locus under selection can be sampled from either: a) a gamma distribution ($s_distribution = "gamma"$) with a mean = **$gamma_shape$** and a variance = **$gamma_scale$** (Huber *et al.*, 2017), or; b) a log normal distribution ($s_distribution = "log_normal"$) with a mean = $\log(log_mean)$ and a standard deviation = $\log(log_sd)$ (Kousathanas & Keightley, 2013; Charlesworth, 2015).

Dominance coefficient (h) values for each locus under selection can be determined by either: a) sampling a normal distribution ($h_method = "distribution"$) with a mean = **$dominance_mean$** and a standard deviation = **$dominance_sd$** (Charlesworth, 2015), or; b) calculating its value as function of its selection coefficient (s) using the following equation (Huber *et al.*, 2018):

$$h = f(s) = \frac{1}{\frac{1}{\theta_{intercept}} - \theta_{rate} \times s}, \quad \text{Equation 3.30}$$

where $\theta_{intercept}$ defines the value of h at $s = 0$. The $\theta_{intercept}$ is the maximum possible value of h , which is reached when $s = 0$. θ_{rate} determines how quickly h approaches zero with smaller selection coefficients. These parameters are named in the model ***intercept*** and ***theta_rate*** respectively. This equation models the hypothesis that deleterious alleles with higher selection coefficient tend to have less detrimental fitness effects on heterozygous individuals (*i.e.*, have smaller dominance coefficients).

Initial frequency of the deleterious allele at loci under selection (q) values are based on the mean frequency of a recessive deleterious variant sampled from a large population in mutation-selection equilibrium, which is approximated by the following quadratic equation (Crow & Kimura, 1970, p. 260):

$$s(1 - 2h)q^2 + hs(1 + u)q - u = 0, \quad \text{Equation 3.31}$$

where u is the mutation rate per generation per site. The model solves this quadratic equation for q for each locus. The parameter name for u in the model is ***mutation_rate***.

Neutral loci have an s and h equal to 0, and q is equal to $1/\text{number_neutral_alleles}$.

3.3.1.4.3 Generation of chromosomes

The generation of chromosomes process is the same as in the general simulations (Section 3.2.1.5.3).

3.3.1.5 Input data (Table 3.19)

Table 3.19 Parameters used in the process of loading the input data.

Parameter name	Description	Range of values
<i>map_resolution</i>	Resolution of the recombination map (bp)	Dependent on genomic resources

3.3.1.5.1 Recombination map

The recombination map of *D. melanogaster* was provided by Comeron *et al.* (2012). The recombination map contains the recombination rates (cM/Mbp/female meiosis) at 100 Kbp scale (*i.e.*, resolution), therefore in this map the parameter ***map_resolution*** = 100,000 bp. The chromosomal regions in Comeron *et. al* (2012) were estimated using the *Drosophila melanogaster* BDGP assembly release 5 (Hoskins *et al.*, 2007). Because the model uses the BDGP release 6 (dos Santos *et al.*, 2015) coordinates as reference to determine the location for the potential targets for selection, the coordinates of the recombination map from the BDGP release 5 were converted to the BDGP release 6 using the *FlyBase* (Thurmond *et al.*, 2019) converter tool (flybase.org/convert/coordinates).

3.3.1.5.2 Targets of selection

The file of targets of selection contains all the coding sequences (CDS) of the study organism and the number of potential targets for selection in each CDS. This file contains five columns:

1. Identification number of each CDS.
2. Chromosome name to which each CDS belongs.
3. CDS start point (bp).
4. CDS end point (bp).
5. Number of potential targets of selection.

To generate this file, I first retrieved from the Ensembl's website (www.ensembl.org; Yates *et al.*, 2020) the following files using the R package *biomaRt* (Durinck *et al.*, 2009):

- The gene transfer format (GTF) file: "Drosophila_melanogaster.BDGP6.22.98.gtf", which contains the location of each coding sequence (*i.e.*, exons) of every chromosome.
- The locations of all non-synonymous and synonymous mutations from all the chromosomes. Ensembl's *D. melanogaster* database contains over 6.7 million SNPs from two populations, one comprising 37 lines from North Carolina (Jordan *et al.*, 2007) and the other comprising 15 lines from Malawi (Begin & Lindfors, 2005).

Using the above files, I used the programming language R to determine the number of non-synonymous and synonymous mutations in each CDS.

To identify proxies for putative targets of selection for the fly simulations, I relied on non-synonymous mutations (*NS*; mutations resulting in substitution of a different amino acid sequence) and synonymous mutations (*S*; mutations resulting in the same amino acid sequence). It is generally assumed that most *NS* can range from being highly deleterious to neutral and *S* are generally neutral (Sella *et al.*, 2009). Under this assumption, neutral *NS* and *S* contribute similarly to a coding region's polymorphism, whereas highly deleterious *NS* are rapidly eliminated by natural selection and thus contribute negligibly to a coding region's polymorphism. Therefore, the ratio of the

number of *NS* to the number of *S* (*NS/S*) occurring within a coding region reflects the intensity of selection at which newly arising *NS* are exposed in a given coding region (Sella *et al.*, 2009). For instance, a coding region with a *NS/S* ratio < 1 denotes that newly arising *NS* are highly deleterious and thus rapidly eliminated. Consequently, as proxy for the putative number of targets of selection within a coding region, I subtracted the number of *S* from the number of *NS* in each coding region, so a negative result would indicate a lower number of targets for selection and a positive result a higher number of targets for selection in a given coding region.

3.3.1.6 Submodels

3.3.1.6.1 Pre-adaptation phase (Table 3.20)

Table 3.20 Parameters used in the pre-adaptation process.

Parameter name	Description	Range of values
<i>adaptation</i>	Whether pre-adaptation phase occur	TRUE or FALSE
<i>gen_number_adaptation</i>	Number of generations of the pre-adaptation phase	Dependent on computing resources
<i>pop_size_adaptation</i>	Population size in the pre-adaptation phase (must be even)	Dependent on computing resources
<i>pop_size_dispersal</i>	Population size in the dispersal phase (must be even)	Dependent on computing resources
<i>same_line</i>	Whether the new formed populations are sampled from the same founding population or from different founding populations	TRUE or FALSE

The fly simulations start with a pre-adaptation phase (*adaptation* = TRUE). In this phase, two isolated populations (without dispersal), or lines, with a constant population size (*pop_size_adaptation*; half males and half females), reproduce during *gen_number_adaptation* non-overlapping generations. When the pre-adaptation phase ends, two populations are formed by sampling at random, *pop_size_dispersal* individuals (half males and half females) from the original isolated populations. Individuals can be sampled at random for both populations from the same original

isolated population (*same_line* = TRUE) or from different original populations (*same_line* = FALSE).

3.3.1.6.2 Sexual reproduction and recombination

Sexual reproduction (Table 3.21)

Table 3.21 Parameters used in the sexual reproduction process.

Parameter name	Description	Range of values
<i>number_offspring</i>	Mean number offspring per mating	Dependent on computing resources
<i>variance_offspring</i>	Variance in the number of offspring per mating	Defined by the user

All females mate, but 20% of males do not mate, 30% mate once, 30% mate twice, and 20% mate three times, as reported by Markow & Sawka (1992) in their experimental fly populations. All the other variables in the reproduction process are the same as in the general simulations (section 3.2.1.7.2).

Recombination (Table 3.22)

Table 3.22 Parameters used in the recombination process.

Parameter name	Description	Range of values
<i>recombination</i>	Whether recombination occurs	TRUE or FALSE
<i>recombination_males</i>	Whether recombination occurs in males	TRUE or FALSE

In the fly model recombination occurs just in females (*recombination_males* = FALSE). If recombination occurs, it happens at the rate set by the recombination map (Section 3.3.1.5.1). All the other variables in the recombination process are the same as in the general simulations (Section 3.2.1.7.2).

3.3.2 Simulation experiment

The purpose of the fly simulations is to establish whether realistic parameter values can explain patterns of genetic variation within and between populations observed in the empirical fly experiment (Chapter 2), and determine whether AOD is a probable explanation for these findings:

- Mean H_e in the high dispersal regime was 61% (coefficient of variation 30%) higher than expected, in the moderate dispersal regime it was 69% (cv 12%) higher than expected and in the low dispersal regime it was 70% (cv 10%) higher than expected.
- Mean F_{ST} in the high dispersal regime was 61% (cv 30%) lower than expected, while in the moderate dispersal and low regimes it was 69% (cv 12%) to 70% (cv 10%) lower than expected, respectively.

In the text below are described the parameter values, and their justification, that were used in the fly simulations presented in Chapter 2 (Table 3.23).

Table 3.23 Parameters and values used in the fly simulations.

Parameter name	Process	Description	Value
<i>pop_size</i>	Initialisation Section 3.3.1.4	Population size	50
<i>number_neutral_alleles</i>	Initialisation Section 3.3.1.4	Number of alleles per neutral loci	4
<i>location_loci_real_data</i>	Initialisation Section 3.3.1.4	Location of loci from a real dataset	Values shown at the end of the table
<i>freq_loci_real_data</i>	Initialisation Section 3.3.1.4	Frequency of loci from a real dataset	Values shown at the end of the table
<i>chromosome_name</i>	Initialisation Section 3.3.1.4	Name of the chromosome	"2L"
<i>targets_factor</i>	Initialisation Section 3.3.1.4	Factor to sample potential targets for selection	1

<i>s_distribution</i>	Initialisation Section 3.3.1.4	Name of the distribution to use to sample s	"log_normal"
<i>gamma_shape</i>	Initialisation Section 3.3.1.4	Shape of the gamma distribution from where s values are sampled	NA
<i>gamma_scale</i>	Initialisation Section 3.3.1.4	Scale of the gamma distribution from where s values are sampled	NA
<i>log_mean</i>	Initialisation Section 3.3.1.4	Mean of the log normal distribution from where s values are sampled	log(0.002)
<i>log_sd</i>	Initialisation Section 3.3.1.4	Standard deviation of the log normal distribution from where s values are sampled	log(4)
<i>h_method</i>	Initialisation Section 3.3.1.4	Method to determine the values of h	"distribution"
<i>theta_rate</i>	Initialisation Section 3.3.1.4	Value for the variable rate of the dominance equation to calculate h	NA
<i>intercept</i>	Initialisation Section 3.3.1.4	Value for the intercept of the equation to determine h	NA
<i>dominance_mean</i>	Initialisation Section 3.3.1.4	Mean of the normal distribution from where h values are sampled	0.25
<i>dominance_sd</i>	Initialisation Section 3.3.1.4	Standard deviation of the normal distribution from where h values are sampled	$\sqrt{0.001}$
<i>mutation_rate</i>	Initialisation Section 3.3.1.4	Mutation rate per generation per site	5×10^{-5}

<i>adaptation</i>	Pre-adaptation Section 3.3.1.6.1	Whether to use the adaptation phase	TRUE
<i>gen_number_adaptation</i>	Pre-adaptation Section 3.3.1.6.1	Number of generations in the adaptation phase	50
<i>same_line</i>	Pre-adaptation Section 3.3.1.6.1	Should the pair of populations be sampled from the same founding population?	TRUE
<i>pop_size_adaptation</i>	Pre-adaptation Section 3.3.1.6.1	Population size in the adaptation phase	1000
<i>dispersal</i>	Dispersal Section 3.2.1.7.1	Whether dispersal occurs	TRUE
<i>number_transfers</i>	Dispersal Section 3.2.1.7.1	Number of dispersing individuals	1
<i>transfer_each_gen</i>	Dispersal Section 3.2.1.7.1	Interval of number of generations in which dispersal occurs	8
<i>number_offspring</i>	Reproduction Section 3.3.1.6.2	Mean number offspring per mating	10
<i>variance_offspring</i>	Reproduction Section 3.3.1.6.2	Variance in the number of offspring per mating	0.4
<i>recombination</i>	Recombination Section 3.3.1.6.2	Whether recombination occurs	TRUE
<i>recombination_males</i>	Recombination Section 3.3.1.6.2	Whether recombination occurs in males	FALSE
<i>fitness_model</i>	Selection Section 3.2.1.7.3	Fitness model to use	"multiplicative"
<i>selection</i>	Selection Section 3.2.1.7.3	Whether selection occurs	TRUE

<i>selection_model</i>	Selection Section 3.2.1.7.3	Selection model to use	"relative"
<i>Ne</i>	Statistics Section 3.2.1.7.4	<i>Ne</i> value to be used in the equation of the expected F_{ST} (Equation 3.2)	14.3
<i>Ne_dispersal</i>	Statistics Section 3.2.1.7.4	<i>Ne</i> value to be used in the equation of rate of loss of H_e (Equation 3.7)	16
<i>map_resolution</i>	Input Data Section 3.3.1.5.1	Resolution of the recombination map	100,000
<i>number_iterations</i>	Overall	Number of simulations replicates	400
<i>gen_number</i>	Overall	Number of generations	34

Note that the below input values are formatted based on the programming language R and the "c" below refers to an R function not recombination rate. These values are the actual input that was used in the simulation model.

```
location_loci_real_data:c(2373262,4960235,7040194,8325394,8348440,11015745,12  
507696,13153885,14615705,14995570,20706003)  
freq_loci_real_data:list(c(0.43,0.57),c(0.38,0.62),c(0.4,0.6),c(0.04,0.09,0.06,0.16,0.15,  
0.5),c(0.21,0.79),c(0.43,0.57),c(0.11,0.41,0.23,0.13,0.12),c(0.66,0.34),c(0.53,0.47),c(0.3  
,0.05,0.41,0.24),c(0.03,0.74,0.15,0.08))
```

3.3.2.1 Dominance (h)

In the fly simulations, the normal distribution method was used to sample the dominance coefficient (h) of each locus. For this distribution a mean (***dominance_mean***) of 0.25 was used based on recent estimates of h (García-Dorado & Caballero, 2000; Manna *et al.*, 2011), and a standard deviation (***dominance_sd***) of $\sqrt{0.001}$ based on Charlesworth (2015).

3.3.2.2 Selection coefficient (s)

The log-normal distribution was used to sample the selection coefficient (s) of each locus. For this distribution a mean (*log_mean*) of $\log(0.002)$ and a standard deviation (*log_sd*) of $\log(4)$ were used, which gives a mean of the selection coefficient of 0.005. These values together with the values of h obtained from a normal distribution in the above section resulted in a mean value of h^* s of 0.001 and a coefficient of variation of 2. These values provided the best fit for values of h^* s in *D. melanogaster* (Charlesworth, 2015).

3.3.2.3 Initial allele frequency of the deleterious allele (q)

The initial allele frequency of the deleterious allele (q) for each locus was determined using Equation 3.31 based on Crow & Kimura (1970, p. 260). A value of 5×10^{-5} for the mutation rate (*mutation_rate*) was used, which together with the values of h , s and *targets_factor* (see next section) results in a number of deleterious alleles per individual per centiMorgan (del/ind/cM) of 40. This value is close to what would be found in a typical fly individual (Table 3.13). Note that by using the above equation, deleterious alleles with very small s have a $q > 1$. Therefore, a maximum q value of 0.5 was set.

3.3.2.4 Number of loci under selection to be simulated

The number of loci under selection is controlled by the parameter *targets_factor*. For this parameter a value of 1 was used, which together with the values of h , s and q results in a number of deleterious alleles per individual per centiMorgan (del/ind/cM) of 40. This value is close to what would be found in a typical fly individual (Table 3.13).

3.3.2.5 Fitness model

It has been suggested that a multiplicative fitness model fits observed data significantly better than alternative models (Lobkovsky *et al.*, 2019). Therefore, in the fly simulations, the multiplicative fitness model (*fitness_model* = “multiplicative”) was used. For the description of fitness models see Section 3.2.1.7.3.

3.3.2.6 Selection model

In most real situations, it is likely that the action of natural selection is a combination of both selection models, absolute and relative. However, probably highly prolific organisms with low investment/parental care, such as *Drosophila*, might be more subject to soft selection which is better modelled by the relative fitness model (Gilbert *et al.*, 2018). Therefore for the fly simulations the relative model (**selection_model** = “relative”) was used. Description of the selection process can be found in Section 3.2.1.7.3.

3.3.2.7 Effective population size (*Ne*)

Ne is controlled by adjusting the variance in the number of offspring per mating (**variance_offspring**; see Section 3.2.1.7.2). For this a parameter, a **variance_offspring** of 0.4 was used, which gives an Ne/Nc ratio = 14.3/50, as in the fly experiment (see Section 2.3.2 in Chapter 2).

3.3.2.8 Census population size (*Nc*)

In the fly experiment, each of the four original stock lines had a census population size (*Nc*) of approximately 2,000 individuals. In the fly simulations, as discussed above an Ne/Nc ratio = 14.3/50 was used. Based on this ratio, for the pre-adaptation phase a population size (**pop_size_adaptation**) of 1,000 was used. For the dispersal phase, a population size (**pop_size**) of 50 was used, as in the actual fly experiment (see Section 2.3.2 in Chapter 2).

3.3.2.9 Dispersal rate

The same dispersal rate as in the experimental fly populations under low dispersal was used ($m = 0.0025$, populations exchange one dispersing individual every 8th generation). See Section 3.2.1.7.1 for the description of the dispersal process.

3.3.2.10 Number of generations

Populations reproduced during 50 generations during the pre-adaptation phase and 34 generations for the dispersal phase, as in the fly experiment (see Section 2.3.2 in Chapter 2).

3.3.2.11 Chromosome simulated

Simulations were based on the chromosome arm 2L because this chromosome had the greatest number of loci genotyped in the fly experiment.

3.3.2.12 Number of neutral loci and number of neutral alleles

In the simulation model the number of neutral loci to simulate is determined by the number of rows of the recombination map (one neutral locus per row; Section 3.2.1.5.1). Recombination rate in the recombination map that was used for the fly simulations (Comeron *et al.*, 2012) was measured every 100 Kbp (*i.e.*, **map_resolution** = 100,000 bp). As mentioned above, the fly simulations were based on the chromosome arm 2L, which has a length of 231 Kbp. Therefore, 231 neutral loci were simulated in the fly simulations. Additionally, the location and number of alleles at time zero (T0) of the loci genotyped in the fly experiment (Holleley, 2009) were used, see values below. Note that the below input values are formatted based on the programming language R and the “c” refers to an R function not recombination rate. These values are the actual input that was used in the simulation model.

- **location_loci_real_data**:c(2373262,4960235,7040194,8325394,8348440,11015
745,12507696,13153885,14615705,14995570,20706003)
- **freq_loci_real_data**:list(c(0.43,0.57),c(0.38,0.62),c(0.4,0.6),c(0.04,0.09,0.06,0.1
6,0.15,0.5),c(0.21,0.79),c(0.43,0.57),c(0.11,0.41,0.23,0.13,0.12),c(0.66,0.34),c(0
.53,0.47),c(0.3,0.05,0.41,0.24),c(0.03,0.74,0.15,0.08)).

For other neutral loci than the above, four alleles with equal frequencies were used ($q = 0.25$).

Using or not using the allele frequencies of the experiment had no effect on F_{ST} (Data not shown). However, using the experiment allele frequencies has an effect on He , which gives a better approximation to the observed pattern in the real data (Figure 3.27 and Figure 3.28).

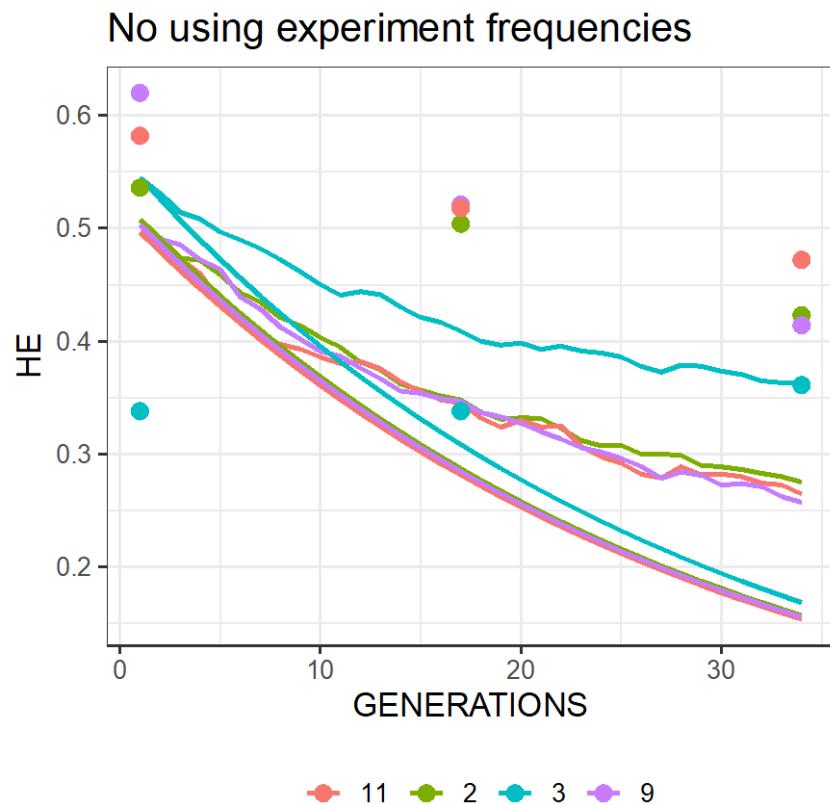


Figure 3.27 | Simulations in which the empirical allele frequencies of the fly experiment were not used as input. Each colour represents a different microsatellite in the fly experiment. Points are the actual heterozygosity (He) values in the fly experiment, straight lines are the expected He values based on **Equation 3.7** and irregular lines are the He of the simulations.

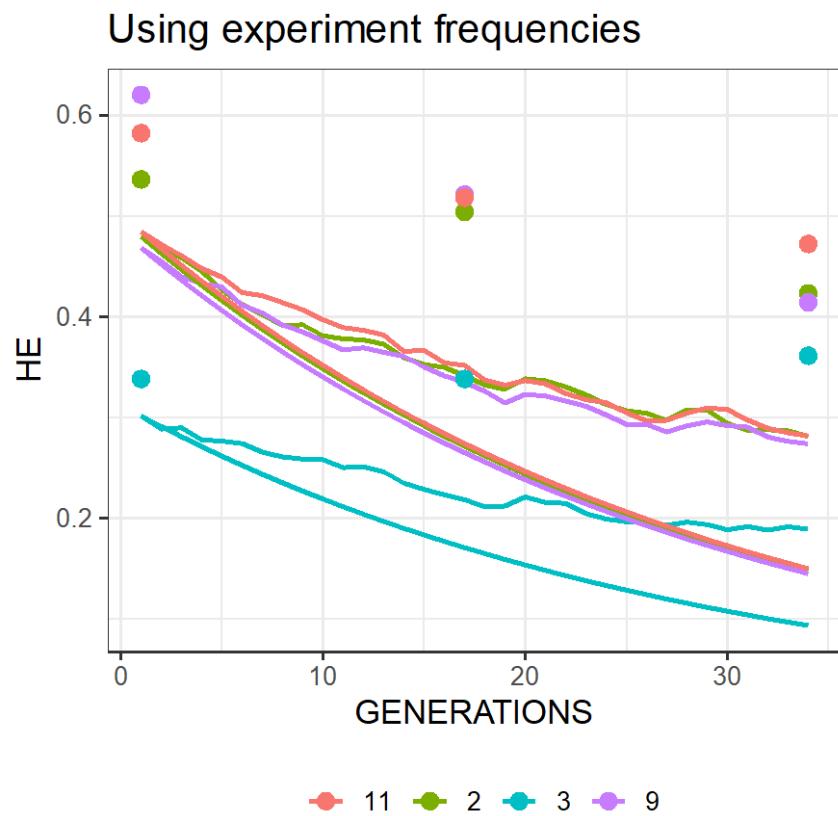


Figure 3.28 | Simulations in which the actual allele frequencies of the fly experiment were used as input. Each colour represents a different microsatellite in the fly experiment. Points are the actual heterozygosity (He) values in the fly experiment, straight lines are the expected He values based on Equation 3.7 and irregular lines are the He of the simulations.

3.3.2.13 Recombination

In the fly simulations recombination occurs just in females (`recombination_males = FALSE`) as in the actual biology of *D. melanogaster* (see Section 3.3.1.6.2 for the description of the recombination process).

3.3.2.14 Pre-adaptation phase

The pre-adaptation phase was used in the fly simulations, as in the fly experiment. For the description of the pre-adaptation phase see Section 3.3.1.6.1. The variance of the effect of AOD between loci is greater using the preadaptation phase (Figure 3.29) than not using it (Figure 3.30). The former approximates better the real data.

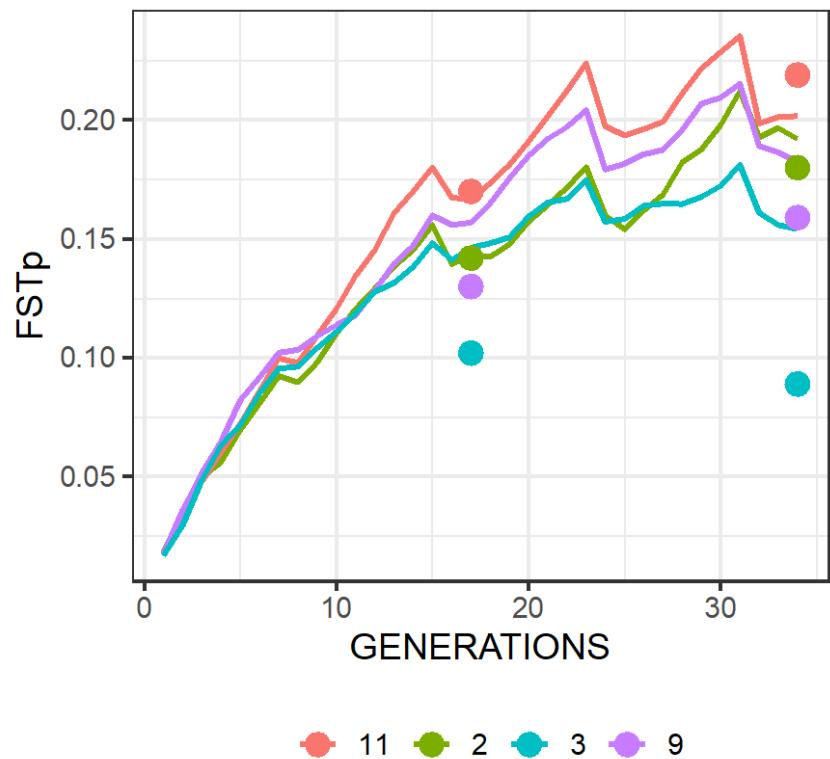


Figure 3.29 | Simulations in which the pre-adaptation phase was used. Each colour represents a different microsatellite in the fly experiment. Points are the actual F_{ST} values in the fly experiment and lines are the F_{ST} values of the simulations.

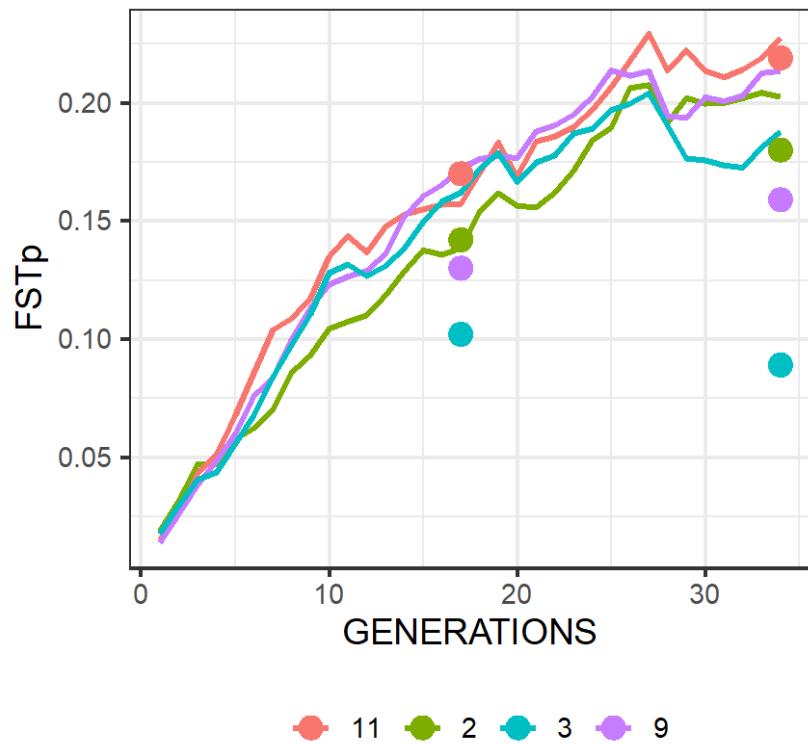


Figure 3.30 | Simulations in which the pre-adaptation phase was not used. Each colour represents a different microsatellite in the fly experiment. Points are the actual F_{ST} values in the fly experiment and lines are the F_{ST} values of the simulations.

3.4 CHILLINGHAM SIMULATIONS

The general and the fly models serve as foundation for the Chillingham model. To avoid repetition of information, only the issues that are different from the general and fly models are mentioned in this section.

3.4.1 Model description

3.4.1.1 Purpose

The purpose of the Chillingham simulations is to establish whether realistic parameter values can explain patterns of genetic diversity observed in the Chillingham cattle (Chapter 2), and determine whether AOD is a possible explanation. The patterns observed in the Chillingham cattle are:

- The proportion of polymorphic loci (9.1%; 70,150 polymorphic of 772,488 SNPs) was higher than predicted by neutral models (2.4%; Williams *et al.*, 2016).
- In contrast to other cattle breeds (Williams *et al.*, 2016), polymorphic loci are clustered in blocks in specific chromosome regions.

3.4.1.2 Process overview and scheduling

The Chillingham model involves three main processes: loading of input data (performed once at the beginning of the simulations), initialization (performed at the beginning of each simulation replicate), and three submodels (performed every generation; see Figure 3.31). In the Chillingham model three submodels are executed every generation in the following order:

1. Sexual reproduction and recombination.
2. Natural selection (selection of offspring to become the parents of the next generation).
3. Calculation of genetic statistics.

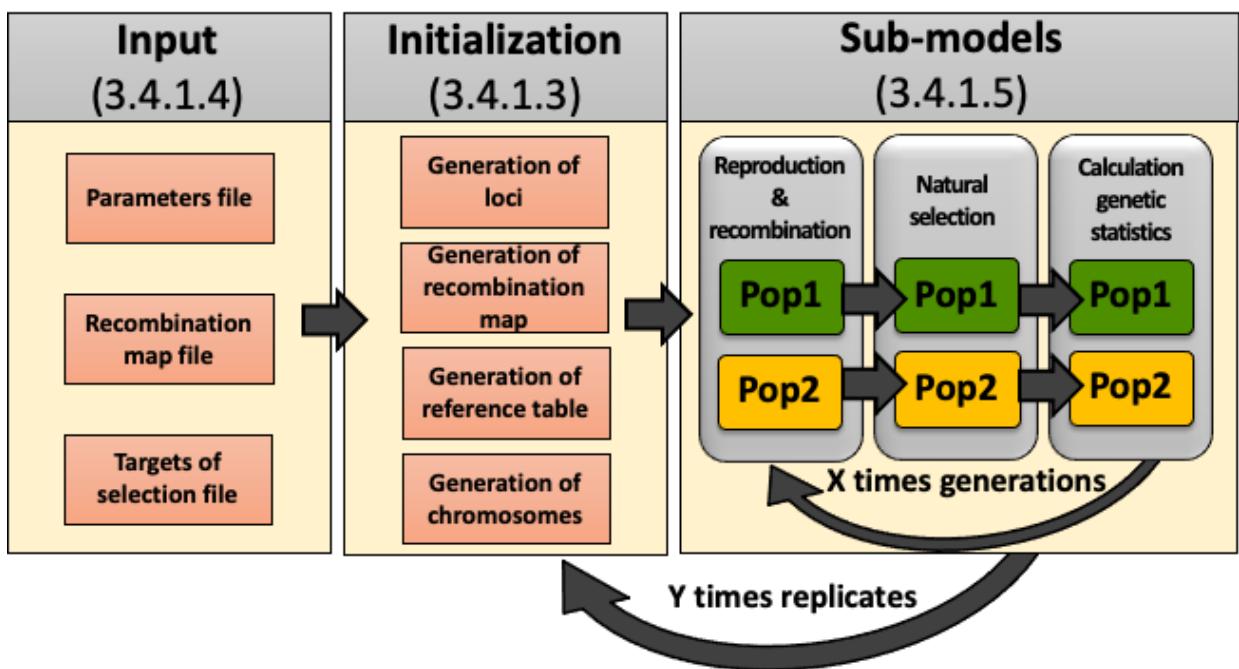


Figure 3.31 | Schematic representation of the simulation processes in the Chillingham simulations. Information within parentheses correspond to the section in which the process is described.

3.4.1.3 Initialization

Initialization process is the same as in the fly simulations (Section 3.3.1.4).

3.4.1.4 Input data

The description of each process begins with a table containing a small description of the parameters that can be controlled in each process. However, the full explanation of the parameters and how they are used in the model are described in the text below the table and parameter names appear in an ***italic bold*** format. For clarity, I used the same parameter names as used in the actual code of the model. The names of the parameters were chosen to be self-explanatory.

3.4.1.4.1 Recombination map (Table 3.24)

Table 3.24 Parameters used in the recombination map process.

Parameter name	Description	Range of values
<i>map_resolution</i>	Resolution of the recombination map (bp)	Dependent on genomic resources

The recombination map of *B. taurus* was obtained from (Ma *et al.*, 2015). This recombination map contains the recombination rate between each pair of consecutive SNPs calculated separately in males and females. To obtain the recombination rate per region, the model first calculates the mean of the recombination rate between the two sexes in each genomic location in which recombination was measured and then it adds all the recombination rates occurring in each region of size = *map_resolution*.

3.4.1.4.2 Targets of selection

To generate this file, I first retrieved from the Ensembl's website (www.ensembl.org; Yates *et al.*, 2020) the following files using the R package *biomaRt* (Durinck *et al.*, 2009):

- The gene transfer format (GTF) file: "Bos_taurus.ARS-UCD1.2.99.gtf", which contains the location of each coding sequence (CDS; *i.e.*, exons) of every chromosome.
- The locations of all non-synonymous mutations from all the chromosomes. Ensembl's *B. taurus* database contains 104 million SNPs.

Using the above files, I used R to determine the number of non-synonymous mutations in each CDS.

To identify proxies for putative targets for selection for the Chillingham simulations, the same process was used as for the fly simulations (Section 3.3.1.5.2), with the following alterations: proxies for targets for selection were based on only non-synonymous mutations because in species with small N_e , such as cattle (MacEachern *et al.*, 2009), selection is less efficient in eliminating deleterious mutations than in species with large N_e (Galtier, 2016; Chen, Glémén, *et al.*, 2017). This makes the approach I used in the fly simulations (subtracting synonymous from non-synonymous mutations) less

informative. For instance, in *D. melanogaster* 42% of coding regions have a ratio of < 1, whereas in *B. taurus* only 1% of coding regions have a ratio of < 1.

3.4.1.5 Submodels

3.4.1.5.1 Sexual reproduction and recombination

Sexual reproduction

Sexual reproduction is modelled as in the general simulations (Section 3.2.1.7.2).

Recombination (Table 3.25)

Table 3.25 Parameters used in the recombination process.

Parameter name	Description	Range of values
<i>recombination</i>	Whether recombination occurs	TRUE or FALSE
<i>recombination_males</i>	Whether recombination occurs in males	TRUE or FALSE

In the Chillingham model recombination occurs in both sexes (*recombination_males* = TRUE and *recombination* = TRUE). Recombination is modelled as in the fly simulations (Section 3.3.1.6.2).

3.4.2 Simulation experiment

The purpose of the Chillingham simulations is to establish whether realistic parameter values can explain patterns of genetic diversity observed in the Chillingham cattle (Chapter 2), and determine whether AOD is a possible explanation. The patterns observed in the Chillingham cattle are:

- The proportion of polymorphic loci (9.1%; 70,150 polymorphic of 772,488 SNPs) was higher than predicted by neutral models (2.4%; Williams *et al.*, 2016).
- In contrast to other cattle breeds (Williams *et al.*, 2016), polymorphic loci are clustered in blocks in specific chromosome regions.

In the text below are described the parameter values, and their justification, that were used in the Chillingham simulations presented in Chapter 2 (Table 3.26).

Table 3.26 Parameters used in the Chillingham simulations.

Parameter name	Process	Description	Value
<i>pop_size</i>	Initialisation Section 3.4.1.3	Population size	8
<i>number_neutral_alleles</i>	Initialisation Section 3.4.1.3	Number of alleles per neutral loci	2
<i>location_loci_real_data</i>	Initialisation Section 3.4.1.3	Location of loci from a real dataset	NA
<i>freq_loci_real_data</i>	Initialisation Section 3.4.1.3	Frequency of loci from a real dataset	NA
<i>chromosome_name</i>	Initialisation Section 3.4.1.3	Name of the chromosome	"18"
<i>targets_factor</i>	Initialisation Section 3.4.1.3	Factor to sample potential targets for selection	0.3
<i>s_distribution</i>	Initialisation Section 3.4.1.3	Name of the distribution to use to sample <i>s</i>	"log_normal"
<i>gamma_shape</i>	Initialisation Section 3.4.1.3	Shape of the gamma distribution from where <i>s</i> values are sampled	NA

<i>gamma_scale</i>	Initialisation Section 3.4.1.3	Scale of the gamma distribution from where s values are sampled	NA
<i>log_mean</i>	Initialisation Section 3.4.1.3	Mean of the log normal distribution from where s values are sampled	0.02
<i>log_sd</i>	Initialisation Section 3.4.1.3	Standard deviation of the log normal distribution from where s values are sampled	4
<i>h_method</i>	Initialisation Section 3.4.1.3	Method to determine the values of h	"distribution"
<i>theta_rate</i>	Initialisation Section 3.4.1.3	Value for the variable rate of the dominance equation to calculate h	NA
<i>intercept</i>	Initialisation Section 3.4.1.3	Value for the intercept of the equation to determine h	NA
<i>dominance_mean</i>	Initialisation Section 3.4.1.3	Mean of the normal distribution from where h values are sampled	0.1
<i>dominance_sd</i>	Initialisation Section 3.4.1.3	Standard deviation of the normal distribution from where h values are sampled	0.03
<i>mutation_rate</i>	Initialisation Section 3.4.1.3	Mutation rate per generation per site	$1*10^{-3}$
<i>adaptation</i>	Pre-adaptation Section 3.3.1.6.1	Whether to use the adaptation phase	FALSE
<i>gen_number_adaptation</i>	Pre-adaptation Section 3.3.1.6.1	Number of generations in the adaptation phase	NA

<i>same_line</i>	Pre-adaptation Section 3.3.1.6.1	Should the pair of populations be sampled from the same founding population?	NA
<i>pop_size_adaptation</i>	Pre-adaptation Section 3.3.1.6.1	Population size in the adaptation phase	NA
<i>dispersal</i>	Dispersal Section 3.2.1.7.1	Whether dispersal occurs	FALSE
<i>number_transfers</i>	Dispersal Section 3.2.1.7.1	Number of dispersing individuals	NA
<i>transfer_each_gen</i>	Dispersal Section 3.2.1.7.1	Interval of number of generations in which dispersal occurs	NA
<i>number_offspring</i>	Reproduction Section 3.4.1.5.1	Mean number offspring per mating	10
<i>variance_offspring</i>	Reproduction Section 3.4.1.5.1	Variance in the number of offspring per mating	1,000
<i>recombination</i>	Recombination Section 3.4.1.5.1	Whether recombination occurs	TRUE
<i>recombination_males</i>	Recombination Section 3.4.1.5.1	Whether recombination occurs in males	TRUE
<i>fitness_model</i>	Selection Section 3.2.1.7.3	Fitness model to use	"multiplicative"
<i>selection</i>	Selection Section 3.2.1.7.3	Whether selection occurs	TRUE
<i>selection_model</i>	Selection Section 3.2.1.7.3	Selection model to use	"absolute"
<i>Ne</i>	Statistics Section 3.2.1.7.4	<i>Ne</i> value to be used in the equation of the	NA

		expected F_{ST} (Equation 3.2)	
<i>Ne_dispersal</i>	Statistics Section 3.2.1.7.4	Ne value to be used in the equation of rate of loss of He (Equation 3.7)	8
<i>map_resolution</i>	Input Data Section 3.4.1.4	Resolution of the recombination map	100,000
<i>number_iterations</i>	Overall	Number of simulations replicates	250
<i>gen_number</i>	Overall	Number of generations	67

3.4.2.1 Dominance (h)

In the Chillingham simulations, a normal distribution was used to sample dominance (h) values for each locus, with a mean (***dominance_mean***) of 0.15. A lower h value than in the fly simulations was used based on two findings from recent research that suggest:

- Mutations are likely more deleterious (*i.e.*, have a higher selection coefficient) in more complex species, such as humans, than in less complex species such as *Drosophila* (Huber *et al.*, 2017).
- Mutations with higher selection coefficient are more recessive (*i.e.*, have a lower dominance; Huber *et al.*, 2018).

A standard deviation (***dominance_sd***) of $\sqrt{0.001}$ was used based on Charlesworth (2015).

3.4.2.2 Selection coefficient (s)

To sample the selection coefficient (s) for each locus, the log-normal distribution with a mean (***log_mean***) of $\log(0.02)$ and a standard deviation (***log_sd***) of $\log(4)$ was used, which gives a mean of the selection coefficient of 0.05. These values together with the values of h obtained from a normal distribution in the above section, results in a mean value of $h * s = 0.008$ and a coefficient of variation of 2.1. For comparison, the average s (0.005 as used in the fly simulations, Section 3.3.2.2) has been estimated to be 70- to 110-fold more deleterious in humans than in *Drosophila* (Huber *et al.*, 2017).

3.4.2.3 Initial allele frequency of the deleterious allele (q)

The initial allele frequency of the deleterious allele (q) for each locus under selection was determined using Equation 3.7 based on Crow & Kimura (1970, p. 260). A value of 1×10^{-3} for the mutation rate (**mutation_rate**) was used, which together with the values of h , s and **targets_factor** (see next section) result in a number of deleterious alleles per individual per centiMorgan (del/ind/cM) of 1.1. This value is close to what would be found in a typical cattle individual (Table 3.13). Note that by using the above equation, deleterious alleles with very small s have a $q > 1$. Therefore, a maximum q value of 0.5 was set.

3.4.2.4 Number of loci under selection to be simulated

The number of loci under selection is controlled by the parameter **targets_factor**. For this parameter, a value of 0.03 was used, which together with the values of h , s and q result in a number of deleterious alleles per individual per centiMorgan (del/ind/cM) of 1.1. This value is close to what would be found in a typical cattle individual (Table 3.13).

3.4.2.5 Fitness model

It has been suggested that a multiplicative fitness model fits observed data significantly better than alternative models (Lobkovsky *et al.*, 2019). Therefore, in the Chillingham simulations, the multiplicative fitness model (**fitness_model** = “multiplicative”) was used. For the description of fitness models see Section 3.2.1.7.3.

3.4.2.6 Selection model

In most real situations, it is likely that the action of natural selection is a combination of both selection models, absolute and relative. However, organisms such as cattle that are not highly prolific and have a high investment/parental care might be subject to hard selection, which is better modelled by the absolute fitness model (Gilbert *et al.*, 2018).

Description of the selection process can be found in Section 3.2.1.7.3.

3.4.2.7 Effective population size (N_e)

N_e is controlled by the adjusting of the variance in the number of offspring per mating using the parameter ***variance_offspring***. For this a parameter, a value of ***variance_offspring*** = 10,000 was used, which gives an N_e/N_c ratio = 1. See next section for further justification.

3.4.2.8 Census population size (N_c)

Because an N_e/N_c ratio = 1 was used, a census population size of 8, as estimated in the Chillingham cattle (Williams *et al.*, 2016) was used.

3.4.2.9 Dispersal rate

The two populations in the Chillingham model do not exchange individuals (***dispersal*** = FALSE).

3.4.2.10 Number of generations

Populations reproduced during 67 generations as in the Chillingham cattle (Williams *et al.*, 2016). At the time of sampling, the cattle had a small census population size (approximately 50 males and 50 females), had experienced significant inbreeding and had an estimated effective population size of N_e = 8 (calculated from genetic data, and therefore compensating for any overlap of generations; Visscher *et al.*, 2001). We confirmed this N_e by using the linkage disequilibrium method in NeEstimator V2.1 (Do *et al.*, 2014).

3.4.2.11 Chromosome simulated

Simulations were based on the chromosome 18 because this chromosome had signatures of AOD according to my analyses (see Chapter 2 Figure 2.4 and Appendix 1: Investigating AOD in each Chillingham chromosome).

3.4.2.12 Number of neutral loci and number of neutral alleles

In the simulation model the number of neutral loci to simulate is determined by the number of rows of the recombination map (one neutral locus per row; Section 3.2.1.5.1). The resolution of the recombination map that was used for the Chillingham

simulations (Ma *et al.*, 2015) was set to 100 Kbp (*i.e.*, ***map_resolution*** = 100,000 bp; see section 3.4.1.4.1 for the description of this parameter). As mentioned above, the Chillingham simulations were based on the chromosome 18, which has a length of 658 Kbp. Therefore, 658 neutral loci were simulated in the Chillingham simulations. All the neutral loci had two alleles with equal frequencies (0.5) because the empirical dataset are single nucleotide polymorphisms (SNPs; *i.e.*, biallelic).

3.4.2.13 Recombination

In the Chillingham simulations recombination occurs in both sexes, as in the actual biology of cattle. See Section 3.3.1.6.2 for the description of the recombination process.

3.4.2.14 Pre-adaptation phase

The pre-adaptation phase was not used in the Chillingham simulations.

3.5 CONCLUSION

Computer simulations allow us to explore scenarios that cannot be solved analytically by equation-based theoretical models (DeAngelis & Rose, 1992). One such scenario is the combined interaction between natural selection, genetic drift, gene flow and recombination on multiple loci. Here, I developed an agent-based model (ABM) to simulate this scenario. In this model, the agents are each of the alleles carried by each individual and their interactions are dictated by population genetics theory. The model was first designed to reproduce the predictions of individual theoretical models regarding selection, genetic drift, gene flow and recombination. Then to this design were incorporated realistic values for several parameters of importance and the genetic and demographic characteristics for the two species under study. Finally, the model was allowed to run for several replicates using different value combinations of the main parameters driving linked selection.

Perhaps the most critical aspect of ABM is the emergence behaviour arising from the interaction of the agents (Grimm *et al.*, 2020). In the model, the emergence of associative overdominance (AOD) was the most important output of the model. By using this model, I was able to answer the following questions:

What are the evolutionary, demographic and genomic conditions in which linked selection via deleterious alleles impacts genetic variation within and between populations?

Using my general simulation model (Section 3.2), I explored 320 scenarios with different combinations of the main drivers of linked selection and showed that:

- In most scenarios, the effects of associative overdominance (AOD) overcame the effects of background selection (BGS; Extended Data Figure 2.3).
- AOD depressed genetic differentiation between populations (Extended Data Figure 2.2)
- AOD decreased the rate of loss of genetic variation within populations (Figure 2.1).

- AOD was more prevalent in those scenarios with lower dominance coefficients (h), higher selection coefficient (s), lower effective population size (N_e) and a higher number of targets of selection as measured by the number of deleterious alleles per individual (del/ind/cM).

Can computer simulations parameterised with realistic values explain patterns of genetic diversity observed in live populations?

Using my species-specific simulation model, I explored realistic scenarios regarding the genetic and demographic characteristics of two species (fly Section 3.3 and Chillingham Section 3.4) and showed that:

- Fly simulations reproduced the variance and amount of H_e and F_{ST} observed in the loci genotyped in the fly experiment (Figure 2.2).
- Chillingham simulations replicated the location of genomic regions of high polymorphism observed in the Chillingham cattle (Figure 2.4).

This model will be a valuable resource to predict possible effects of linked selection and to anyone investigating linked selection, including AOD, BGS, selective sweeps (Elyashiv *et al.*, 2016) and the Hill–Robertson interference (Castellano *et al.*, 2016). In addition, this model will be extended, improved, and incorporated into an already established software: dartR (Gruber *et al.*, 2018).

3.6 AUTHORS CONTRIBUTIONS

Luis Mijangos developed the simulation model; Simon Watt run the simulations in the National Computational Infrastructure (NCI); Luis Mijangos analysed the data; Luis Mijangos wrote the manuscript with support from Bill Sherwin, Harvinder Sidhu and Clare Holleley. Zlatko Jovanoski, Isaac Towers, Simon Watt, Bill Sherwin, Harvinder Sidhu and Clare Holleley supervised Luis Mijangos. Richard Nichols provided critical feedback in the development of the model. Aaron Adamack introduced Luis Mijangos to the development of simulations. All authors discussed the results and contributed to the final manuscript.

4 Fragmentation by large dams and implications for the future viability of small platypus populations

Note that this chapter is formatted for submission to Molecular Ecology.

4.1 ABSTRACT

Populations of the evolutionarily unique platypus (*Ornithorhynchus anatinus*) have experienced major declines and local extinctions, from a range of historical and recent interacting threats, including fur trade, land clearing, water resource developments (dams and extractions), invasive species, fire, pollution, and urbanisation. Although spending most of their time in the water, platypuses can move over land. Still significant uncertainties remain whether large dams across the platypus' distribution pose barriers to movement, limiting gene flow and dispersal, which are essential evolutionary and ecological processes. I examined disruption of gene flow between platypus groups below and above large dams (wall height > 10 m). Platypuses were surveyed above and below four dams, matched to three adjacent rivers without dams. Platypus groups were differentiated genetically, above and below dams (F_{ST} between 0.021 and 0.053), comparable to differentiation between groups on adjacent river systems (F_{ST} between 0.035 and 0.045). This indicates that high dams represent significant barriers for platypus movements. Further, F_{ST} between populations was significantly correlated with the year in which the dam was built, increasing by 0.012 every generation (7.9 years), reflecting the effects of these barriers on platypus genetics. Platypus genetic diversity was similar to other species with small population sizes, such as the Koala (*Phascolarctos cinereus*). This study provides evidence of gene flow restriction which jeopardises the long-term viability of platypus populations, when populations are fragmented by large dams. Avoiding building large new dams on rivers would mitigate such impacts in other rivers, while assisted translocations may be beneficial between populations above and below dams.

4.2 INTRODUCTION

Semi-aquatic platypus (*Ornithorhynchus anatinus*), along with echidnas, belong to the order Monotremata, the most species-scarce ($n = 5$) and most basal branch of the mammalian group, which diverged from marsupials and eutherians 166 Mya (Warren *et al.*, 2008). Platypuses have a unique combination of features, including oviparity, venomous spurs in males, electroreception used to locate freshwater macroinvertebrates, biofluorescent pelage, and multiple sex chromosomes (five pairs instead of one; Veyrunes *et al.*, 2008; Bino *et al.*, 2019; Anich *et al.*, 2021). The uniqueness and rarity of platypus features (*sensu* Pavoine *et al.*, 2005), and the longest evolutionary history in mammals (97.6 million years; Isaac *et al.*, 2007), make it arguably the most irreplaceable mammal existing today.

There is increasing evidence of larger numbers of platypuses in historical times, and ongoing declines and extinctions of local populations (Grant & Fanning, 2007; Bino *et al.*, 2019; Hawke *et al.*, 2019). Declines are likely driven by multiple and synergistic threats including river regulation, habitat modification, climate change, pollution, by-catch mortality, and predation by invasive species (Grant & Fanning, 2007; Bino *et al.*, 2019; Hawke *et al.*, 2019). Continued declines due to current and future climate change are predicted as a result of increased frequency and severity of droughts (Bino *et al.*, 2019), as well as elevated water temperature conditions which could lead to the loss of more than 30% of suitable habitat by 2070 (Klamt *et al.*, 2011). The platypus is currently listed as ‘Near Threatened’ by the International Union for Conservation of Nature (IUCN; Woinarski & Burbidge, 2016), ‘Endangered’ in South Australia (*National Parks and Wildlife Act 1972*), ‘Vulnerable’ in Victoria (SAC, 2020), and is currently being assessed for national listing under the *Environmental Protection and Biodiversity Conservation Act 1999*.

Water resource developments, such as dams and water extractions are one of the more serious threats for platypus conservation, given their likely broad impact on habitat (Grant & Temple-Smith, 2003). Dams are widespread across much of the platypus distribution, where as many as 77% (383 out of 495) of the Australian large dams (wall height > 10 m; ancold.org.au) coincide within the regions where platypuses occur (Figure

4.1 a; see also Bino *et al.*, 2020). Adverse effects of dams extend over large areas both upstream and downstream. Water impoundments behind high dams form wind-exposed, deep, and standing (lentic) ecosystems, representing poor foraging and burrowing habitat for platypuses, given lower productivity of macroinvertebrate prey species (Grant & Llewellyn, 1991; Bethge *et al.*, 2003; Grant, 2004; Grant & Fanning, 2007; Marchant & Grant, 2015). Below dams, altered natural flow regimes can significantly impact on platypus abundances and demographics (Hawke *et al.*, 2021), with cold water affecting foraging and burrowing habitat and reducing diversity and abundance of macroinvertebrate prey (Grant & Llewellyn, 1991; Chester & Norris, 2006; Grant & Fanning, 2007; Brooks *et al.*, 2011). Dams are also predicted to be a major barrier for dispersal, with potential long-term ramifications for gene flow, however this aspect remains largely unquantified.

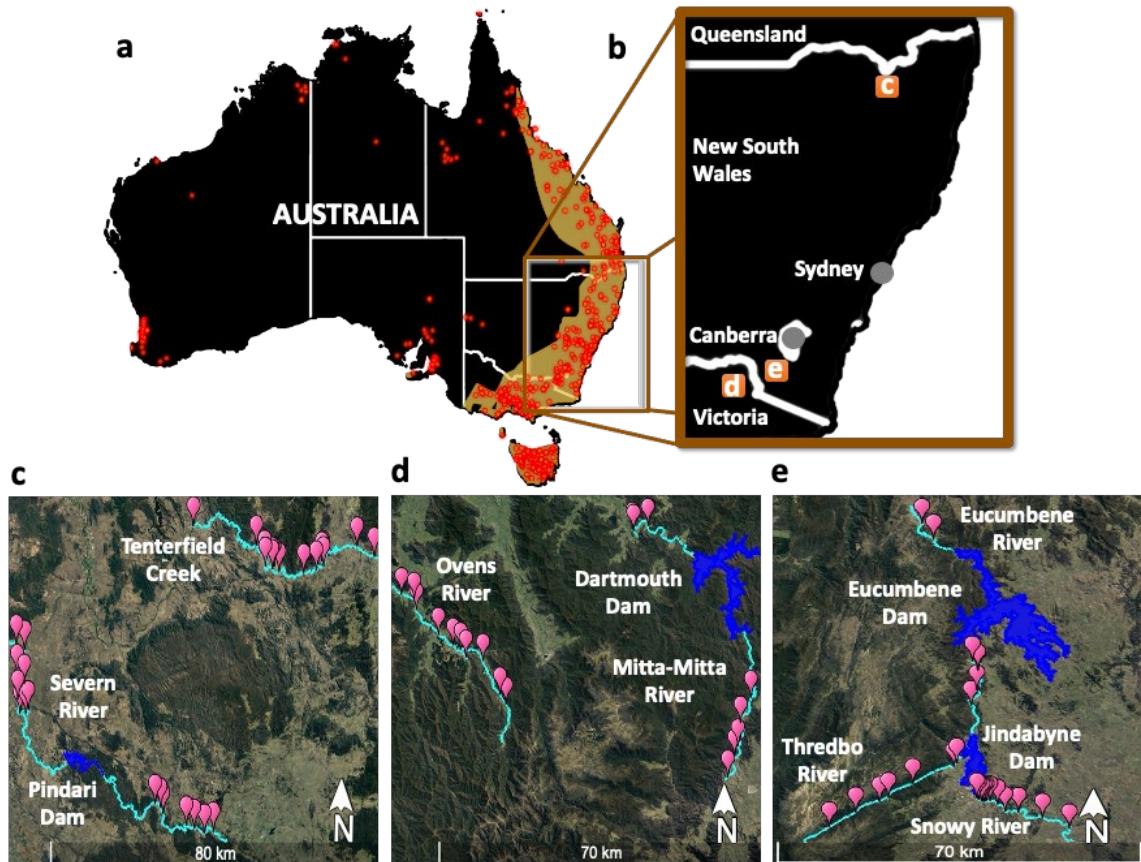


Figure 4.1 | a) Distribution of major dams (> 10 m; red points) built within the IUCN platypus distribution (yellow shade; Woinarski & Burbidge, 2016), and the focus regions for this study (brown inset). **b)** Location of rivers in south-east Australia where platypuses were sampled (orange squares) in regulated (with large dams) and unregulated (no large dams). **c)** Border Rivers: Tenterfield Creek (unregulated) and Severn River (regulated, upstream sections are to the east, confluence is out of the frame, in northwest). **d)** Upper Murray Rivers: Ovens (unregulated) and Mitta-Mitta Rivers (regulated upstream sections are in the south, confluence is out of the frame, in northwest). **e)** Snowy Rivers (do not follow paired experimental design, due to geographic constraints; see methods): Eucumbene (regulated), Thredbo (unregulated), and Snowy Rivers (regulated Snowy flows downstream to the southeast). Sampling sites ($n = 75$) are represented by pink balloons; rivers are coloured in light blue and reservoirs behind dams in dark blue.

The dispersal of individuals between populations or groups that then reproduce leads to gene flow, is a fundamental process in ecology and evolution. Dispersal enables individuals to recolonise available habitat, migrate to areas with more suitable conditions or find other mates (Baguette *et al.*, 2013). Gene flow replenishes the gene pool of populations, reducing the probability of inbreeding within populations (Garant *et al.*, 2007; Tigano & Friesen, 2016). Additionally, barriers to dispersal can prevent individuals from repopulating areas where populations have become extinct, or impede the recovery of populations from the consequences of small population sizes, such as lower survival and lower reproduction output due to inbreeding depression or catastrophic stochastic events (Frankham, 2015). Population viability analyses suggest significant impacts by large dams, particularly in synergy with lower habitat quality and droughts, which are projected to increase (Bino *et al.*, 2020). However, the extent to which dams restrict platypus dispersal remains unclear because landscape connectivity varies by both the species' life history and landscape features (Baguette *et al.*, 2013). Gene flow may occur between rivers (*e.g.*, Shoalhaven and adjacent Hawkesbury-Nepean River; Kolomyjec *et al.*, 2009), with evidence of shared haplotypes between river systems potentially indicating gene flow over longer distances than previously thought. Furthermore, it is uncertain whether platypuses could climb over a dam wall or move overland around a dam.

Genetic-based methods used to infer patterns of dispersal and gene flow (Balkenhol *et al.*, 2015) commonly examine the positive relationship between the amount of genetic differentiation between populations or individuals and the geographic distance separating them (Ramachandran *et al.*, 2005). The presence of a dispersal barrier could be inferred by testing whether populations or individuals, separated by potential barriers, are more genetically differentiated than populations or individuals in landscapes lacking such barriers but separated by the same distance. Genetic differentiation can increase due to dispersal barriers within one to 15 generations during simulations (Landguth *et al.*, 2010), but is unlikely to arise if population size is large (> 50 individuals) or if the species lifespan is long (> 22 years; Hoffman *et al.*, 2017).

To determine whether dams have reduced dispersal and gene flow between platypus groups, I analysed genetic data from platypuses sampled in seven rivers, four rivers were regulated by high dams and three were unregulated (Figure 4.1). If dams adversely affected gene flow between platypus groups, I predicted the following: a) individuals and groups separated by a dam in a river should be more differentiated than in an unregulated river, and; b) genetic differentiation across dams should correlate with the time since the dam was built. The latter investigation included dams not sampled in this study (Figure 4.1).

4.3 METHODS

4.3.1 Study areas and field work

Samples from platypuses were collected from seven different rivers (four regulated by dams and three unregulated) across three different regions in south-east Australia (see Figure 4.1 and Table 4.1), also described in Hawke *et al.*, (2021). Trapping and handling of platypuses were carried out in accordance with guidelines and approved by the NSW Office of Environment and Heritage (SL101655), NSW Department of Primary Industries (P15/0096-1.0 & OUT15/26392), and UNSW's Animal Care and Ethics Committee (16/14A). River flows upstream of dams were minimally regulated, contrasting with heavily regulated downstream flows (Hawke *et al.*, 2021).

Table 4.1 The three study systems and the major dams. See Figure 4.1 for details of geography. The letters c, d, e refer to panels in Figure 4.1.

Region	River/Creek	Dam name	Year of construction	Dam height (m)	Dam volume (GL)
Border Rivers ^c	Tenterfield	-	-	-	-
	Severn	Pindari	1969	85	312
Upper Murray Rivers ^d	Ovens	-	-	-	-
	Mitta-Mitta	Dartmouth	1979	180	3,856
Snowy Rivers ^e	Snowy	Jindabyne	1967	72	688
	Thredbo	-	-	-	-
	Eucumbene	Eucumbene	1958	116	4,798

GL - Gigalitres.

Platypuses were surveyed for 172 nights (1,550 net hours) across 108 sites (Jan 2016–May 2018, Fig. 1), aiming to cover a minimum of 40 km of each unregulated river and 20 km of river above and below dams on regulated rivers. The procedure of trapping and sampling platypuses, including details of anaesthesia, used in this study are described in Hawke, Bino and Kingsford (2021) and Bino, Kingsford, Grant, Taylor and Vogelnest (2018). Briefly, platypuses were captured using fyke nets or unweighted mesh (gill) and implanted with a Passive Integrated Transponder (PIT) tag (Trovan) to identify recaptured individuals. Platypuses were then weighed, measured, sexed, aged, and blood collected (~2 ml) and stored in Qiagen RNAProtect® animal blood tubes (Qiagen,

Hilden, Germany). My proxy of abundance for each river was the following metric: unique number of captures / number of sampling nights x length of the river surveyed (see Hawke *et al.*, 2021).

4.3.2 Laboratory work

Genomic DNA was extracted from whole blood using a Qiagen DNeasy Blood and Tissue kit (Qiagen, Hilden, Germany). DNA quality and concentration were visualized using agarose gel electrophoresis and quantified fluorimetrically with a Qubit 2.0 (Thermo Fisher Scientific). Samples were genotyped using DArTseq™ (DArT Pty Ltd, Canberra, ACT, Australia). DArT's procedure uses a combination of genome complexity reduction methods using restriction enzymes, implicit fragment size selection and next-generation sequencing to produce thousands of SNPs (single nucleotide polymorphisms), randomly distributed throughout the genome (Kilian *et al.*, 2012). Read sequences were processed using proprietary DArT analytical pipelines (Kilian *et al.*, 2012) and mapped to the representative platypus genome (mOrnAna1.p.v1, GenBank assembly accession: GCA_004115215.2; total sequence length of 1.8 Gbp, 305 scaffolds with a N50 of 83 Mbp). Refer to Georges *et al.* (2018) for details of DArT sequencing, genotyping and filtering processes. DArT's genotyping has various advantages such as limiting the potential for ascertainment bias (Steane *et al.*, 2011), providing metadata for each locus with various quality and BLAST alignment measures including the proportion of replicates for which the marker score is consistent (RepAvg) and the average of the polymorphism information content of the reference and SNP allele (AvgPIC).

4.3.3 SNP filtering

The criterion for SNP filtering used to analyse variation between populations (*e.g.*, F_{ST}) can bias estimates of variation within populations (*e.g.*, heterozygosity; Schmidt *et al.*, 2020). Therefore, I used different filtering settings for each type of analysis (Table 4.2).

Table 4.2 Filters and their thresholds used to remove loci for the analyses based on variation between populations and variation within populations.

Filter	Variation between populations	Variation within populations
Reproducibility (RepAvg)	> 100%	> 100%
Retain only one SNP per read	Used	Used
Departure from Hardy-Weinberg proportions	< 0.05	< 0.05
Mapped to chromosome	Used	Used
BLAST alignment E-value	< 1e-20	< 1e-20
Missing data by loci	> 90%	> 100%
Missing data by individual	> 90%	> 100%
Minor allele count (MAC)	> 3	Not used
Linkage disequilibrium (r^2)	< 0.2	Not used
Remove loci located within coding regions	Used	Not used
Remove loci located within sex chromosomes	Not used	Used
Monomorphic loci	Used	Not used
Total SNPs after filtering	2,252	4,790

I first describe the filters, applied to the entire dataset, and then those applied for analyses, based on variation between populations, and finalise with the description of filters applied to analyses, based on variation within populations. The filtering was performed in R (R Core Team, 2021) package *dartR* (Gruber *et al.*, 2018).

Filters applied to entire dataset. Because filtering for Hardy-Weinberg proportions requires the delimitation of populations or groups, I assigned individuals *a priori* into groups, based on the river that individuals were sampled and whether they were sampled below or above dams. To reduce genotyping errors that might have arisen during library preparation and SNP calling (O'Leary *et al.*, 2018), I discarded loci that had a reproducibility of less than 100% (RepAvg). I also discarded loci if the read had more than one SNP; a significant departure from Hardy-Weinberg proportions, within any one group after Bonferroni correction with a p-value of less than 0.05 and; was not mapped to a chromosome and had a BLAST alignment E-value of less than 1×10^{-20} .

Filters applied to analyses based on genetic variation between populations. I discarded loci with more than 10% of missing data and then discarded individuals with more than 10% of missing data. By discarding loci, based on minor allele frequency (MAF), there is the potential to alter subsequent analyses (Linck & Battey, 2019). I discarded loci that had a minor allele count (MAC) of less than three. Because physical linkage between loci can affect analyses of genetic structure (Abdellaoui *et al.*, 2013), I removed one of two loci if they were in linkage disequilibrium (LD). For this, I used a threshold of the LD statistic r^2 (Hill & Robertson, 1968) of > 0.2 and the clumping algorithm in the R package *bigsnpr* (Privé *et al.*, 2018). In the latter filter, the SNP with lower polymorphic information content (AvgPIC) was discarded. Selectively neutral loci are better suited to infer population dynamics, such as dispersal, than loci under selection because they allow separation of potential confounding factors arising from natural selection (Holderegger & Wagner, 2008). Consequently, I removed loci located within coding regions using the GFF (General Feature Format; ref_mOrnAna1.p.v1_top_level.gff3) file retrieved from *Ensembl* (www.ensembl.org; Yates *et al.*, 2020). Note that despite 3rd position SNPs usually being silent (*i.e.*, synonymous mutations), I also filtered out these SNPs for two reasons. Firstly, recent research suggests that synonymous mutations have also effects on fitness (Lebeuf-Taylor *et al.*, 2019). Secondly, natural selection can affect not only genetic variation with direct consequences on fitness, but can also affect adjacent neutral genetic variation due to genetic linkage (Smith & Haigh, 1974). I also discarded loci monomorphic within each group.

Filters applied to analyses based on genetic variation within populations. When loci with missing data are not removed, observed and expected heterozygosity estimates diverge (Schmidt *et al.*, 2020), accordingly I removed loci with missing data. Heterozygosity (He) is generally lower for SNPs with rare alleles than for SNPs with common alleles (Schmidt *et al.*, 2020), therefore I did not filter out loci using minor allele count (MAC). Because filtering out loci based on linkage disequilibrium results in a decrease of rare and monomorphic alleles and excess of the common alleles and therefore biases estimates of He (Dementieva *et al.*, 2021), I did not remove loci that were in linkage disequilibrium. Estimates of He should reflect as much as possible genome wide patterns (Miller *et al.*, 2014), therefore I kept loci located outside and

inside coding regions. I removed loci in sex chromosomes for two reasons. Firstly, He in sex chromosomes differs from He in autosomes because sex chromosomes occur in different proportions in males and females (Schaffner, 2004). Secondly, the platypus has five pair of chromosomes (Veyrunes *et al.*, 2008), which will bias the overall estimation of He . Including monomorphic sites loci to estimate He reduces the bias from sample size (Schmidt *et al.*, 2020), consequently, I retained all monomorphic loci. Finally, relatedness analyses using the R package *related* (Pew *et al.*, 2015) were performed to identify any recaptures not identified or mislabelling in the field or the laboratory.

A total of 235 platypuses were captured and blood sampled across three river regions in south-east Australia between January 2016 and May 2018. DNA extraction and DArTTM sequencing were successful in 218 blood samples from individuals. Two samples, each collected in a different river (V30 in Ovens and V32 in Mitta Mitta), showed contrasting genetic patterns relative to samples collected in the same river (Extended Data Figure 4.1). Relatedness analyses performed in the R package *related* (Pew *et al.*, 2015), revealed these two samples had closer relatives in the opposite river (Extended Data Table 4.1). Additionally, the locations of these two samples were separated by 46 Km, steep mountainous terrain, and a river system. Under these conditions, I considered that dispersal events were unlikely and concluded that samples were mislabelled and therefore assigned them to the presumed correct river and site. Relatedness analyses also identified two pairs of samples in which each pair was collected from the same individual (*i.e.*, recaptures; samples T3-T5 and T28-T42; Extended Data Table 4.1), consequently, I removed one sample from each pair. In the unlikely event that these were pairs of identical twins, it would still be appropriate to remove one of each pair.

Sequencing provider DArTTM (Canberra) successfully genotyped 17,631 single nucleotide polymorphism (SNP) loci. After stringent filtering, my dataset for analysing genetic variation between populations comprised 2,252 SNPs genotyped in 214 platypus samples (108 females, 106 males). After filtering, my dataset for analysing genetic variation within populations comprised 4,790 SNPs genotyped in 214 platypus samples (108 females, 106 males).

4.3.4 Data analyses

4.3.4.1 Genetic variation within groups

To measure genetic variation within rivers, I calculated observed heterozygosity (H_o) and expected heterozygosity (H_e) and allelic richness using the R package *Hierfstat* (Goudet, 2005). After identifying the data did not conform to a normal distribution, using a Shapiro-Wilk test of normality (R function *shapiro.test*), I tested whether H_e was significantly different between groups using a non-parametric Mann-Whitney U test (R function *wilcox.test* with option paired = FALSE). Additionally, I calculated the inbreeding coefficient (F_{IS}) of each river group using *Hierfstat*. To understand the degree to which genetic drift influences the rate of loss of genetic variation in the surveyed groups, I estimated the effective population size (N_e ; England *et al.*, 2006) of each river, using the linkage disequilibrium method of Waples and Do (2008) as implemented in the software *NeEstimator* V2.1 (Do *et al.*, 2014). To identify inbred individuals (*i.e.*, offspring of parents that were close relatives), I estimated inbreeding for each individual using two different statistics (see Keller *et al.*, 2011): a) F_{alt} where homozygous loci are weighted with the inverse of their allele frequency using the software GCTA (Yang *et al.*, 2011) using the *-ibc* option and; b) F_h which is a deviation in homozygosity from its Hardy-Weinberg expectation using the software PLINK (Purcell *et al.*, 2007) using the *-het* option.

4.3.4.2 Investigating whether dams affect connectivity between platypus groups

I used multiple approaches to investigate whether dams affect gene flow between platypus groups. Firstly, to test whether groups separated by dams are more genetically different than otherwise, I first divided the sampling sites of each pair of rivers into comparable upstream and downstream groups. For regulated rivers (Severn and Mitta-Mitta), the dam was used as reference point for the division, and for unregulated rivers (Tenterfield and Ovens) the division point was chosen at a comparable position to the dam in the paired regulated river. I then calculated the genetic differentiation using Weir & Cockerham's F_{ST} estimator (1984) in each river. I tested the significance of the difference of F_{ST} values between dammed and unregulated rivers using a Mann-Whitney U test.

Secondly, to test whether the number of platypus generations since the building of the dams can predict the genetic differentiation between populations (F_{ST}), I used univariate linear regression models (R function *lm*). For this analysis, I also included previous F_{ST} estimates of platypus groups, separated by dams from Kolomyjec (2010; Nepean dam; height 82 m; built in 1935; $F_{ST} = 0.077$) and Furlan *et al.* (2013; Upper Yarra dam; height 89 m; built in 1957; $F_{ST} = 0.046$). These two studies used microsatellite data. I considered one platypus generation to be 7.9 years based on Pacifici *et al.* (2013).

Thirdly, to visualise the spatial distribution of genetic variation of the sampled individuals, I performed principal coordinates analysis (PCA) using the R package *dartR* (Gruber *et al.*, 2018). PCA is a statistical method that summarises the variance in the data and projects the top principal components onto a series of orthogonal axes (McVean, 2009). PCA is a method that does not rely on any genetic model or principle, but spatial patterns revealed by PCA are representative of evolutionary processes such as genetic structure, gene flow and founder effects (McVean, 2009).

As another way of presenting whether dams act as barriers for gene flow, I generated maps depicting the spatial patterns of genetic differentiation between individuals within each river. I considered that a dam acts as a barrier if areas of high genetic differentiation coincided with the location of the dam in each pair of rivers. Maps were generated with the program Alleles In Space (AIS; Miller, 2005). AIS produces a three-dimensional map where the X and Y axes correspond to geographical locations and the Z axis (height) represents genetic differentiation. For this, AIS first constructs a network connecting sampling locations. Next, the program calculates the genetic differentiation between the connected locations. The program considers that this differentiation occurs at the midpoint of each connection. In the final step, calculated values are used to predict genetic differentiation values at unsampled locations by using an inverse distance-weighted interpolation (Watson & Philip, 1985; Watson, 2013).

The linear structure of rivers represents a challenge for spatial analyses, such as AIS, because distances within rivers differed from straight-line distances. Furthermore, the spatial patterns that might emerge from the linearity of rivers cannot be captured properly by methods designed to analyse spatial data in two-dimensions. To address this, I first calculated the distances between sampling locations along the river network using the R package *riverdist* (Tyers, 2017). Secondly, to convert the spatial data (*i.e.*, the sampling locations and the distance between them) from one-dimension into two-dimensional data, I placed the sampling locations on the circumference of a 180° semicircle with a diameter equal to the distance between the locations at the extremes of the river. As a result, the sampling locations on the semicircle were separated from each other by about the same distance as they are separated within the actual river. These converted coordinates were then provided to AIS as input. Such approach is suggested in the user manual of the program *Geneland* (Guillot *et al.*, 2005, p. 54).

To account for the positive correlation between genetic differentiation and geographical distance (Ramachandran *et al.*, 2005) in the AIS analyses, I chose the option to use the residual values from a regression of geographic distance and genetic differentiation as the values to be plotted in the Z axis (height) of the maps, which represents genetic differentiation. Because my approach of converting the data from one-dimension into two-dimensional data created many unsampled areas in the maps, I developed a network that connected only neighbouring sampling locations using the Delaunay triangulation (Brassel & Reif, 1979) option. AIS failed to construct the connectivity network using Delaunay triangulation in rivers with a small number of sample sites or where sampling sites were clumped together (Snowy, Thredbo, Eucumbene, Severn, and Mitta-Mitta). In these rivers the connectivity network was constructed by connecting all pairwise locations. Delaunay triangulation option was used in the remaining rivers (Eucumbene/Snowy, Thredbo/Snowy, Tenterfield and Ovens).

4.4 RESULTS

4.4.1 Genetic variation within groups

Mean genetic variation across all rivers (expected heterozygosity; $He = 0.141$) was low. He was significantly different between all groups (except for Severn above the dam/Severn below the dam; p-value > 0.05; Table 4.3). He was also significantly different between regions (except for Snowy Rivers/Upper Murray Rivers; p-value > 0.05; Table 4.3). Border Rivers, located in the north, had the lowest He (range: 0.132-0.138), followed by the Snowy Rivers (0.136-0.143) and the Upper Murray Rivers (0.143-0.153), river regions in the south (Figure 4.1). Estimates of allelic richness follow the same trend as heterozygosity estimates. Inbreeding estimates (F_h and F_{alt}) were moderately high (Extended Data Figure 4.2 and Extended Data Table 4.2) and showed a similar pattern as He . Border Rivers had the highest inbreeding estimates followed by Snowy Rivers and Upper Murray Rivers. Ne estimates were uncorrelated with sample size as shown by simple linear regression (Data not shown), so the Ne estimates were not artefacts of sample size. Samples collected below the dam of the Severn River had the lowest He but the highest Ne among all rivers surveyed (Table 4.3). Similarly, samples collected in the Eucumbene River below the dam had the lowest He among the Snowy Rivers. In contrast, samples collected below the dam in the Mitta-Mitta River had the highest He , one of the lowest inbreeding estimates (F_{IS} , F_h and F_{alt}) among all the rivers and the lowest in both sample size ($n = 4$) and my proxy of abundance ($n = 4$). Likewise, samples collected in Eucumbene above the dam had the highest He among the Snowy Rivers, low inbreeding estimates and low sample size ($n = 4$).

Table 4.3 Summary genetic statistics across the three rivers regions, the number of samples and a proxy of abundance calculated as (unique number of captures / number of sampling nights) x length of the river surveyed based on Hawke, Bino and Kingsford (2021); H_o - observed heterozygosity; H_e - expected heterozygosity; F_{IS} - inbreeding coefficient; F_h - inbreeding estimates with program PLINK (Purcell *et al.*, 2007); F_{alt} inbreeding estimates with program GCTA (Yang *et al.*, 2011); N_e - effective population size; N_e CI - confidence intervals; NA not calculated due to small sample size.

Region	River	Sample size	Proxy of abundance	Allelic richness	H_o	H_e	F_{IS}	F_h	F_{alt}	N_e	N_e CI
Border Rivers	Tenterfield	39	207	1.492	0.137	0.138	0.011	0.121	0.104	147	143-151
	Severn above dam	23	115	1.453	0.135	0.134	-0.008	0.134	0.103	39	38-39
	Severn below dam	17	83	1.440	0.133	0.132	0.000	0.149	0.109	320	277-378
Upper Murray Rivers	Ovens	19	27	1.295	0.145	0.146	0.005	0.072	0.121	66	64-68
	Mitta-Mitta above dam	13	19	1.296	0.141	0.143	0.011	0.099	0.101	79	75-84
	Mitta-Mitta below dam	4	4	1.304	0.153	0.153	-0.021	0.020	0.087	NA	NA
Snowy Rivers	Snowy	56	46	1.282	0.141	0.141	0.007	0.098	0.047	113	112-115
	Thredbo	19	37	1.281	0.142	0.141	-0.005	0.092	0.043	95	92-99
	Eucumbene above dam	4	36	1.289	0.145	0.143	-0.028	0.074	0.067	NA	NA
	Eucumbene below dam	20	50	1.267	0.137	0.136	-0.004	0.120	0.061	32	32-33
Mean across rivers		21	62	1.492	0.141	0.141	-0.003	0.098	0.084	111	-
Note that small sample size in Mitta-Mitta below the dam and Eucumbene above the dam (both n=4) is likely to result in unreliable estimates of diversity.											

4.4.2 Connectivity between platypus groups – effects of dams

Genetic differentiation between groups within the same region estimated by F_{ST} (Table 4.4) were low. For unregulated and regulated river comparisons, the river with the dam showed higher genetic differentiation: Severn below versus above dam had $F_{ST} = 0.051$, whereas Tenterfield above versus below had $F_{ST} = 0.009$; and Mitta-Mitta above versus below dam had $F_{ST} = 0.021$, whereas Ovens above versus below had $F_{ST} = 0.001$ (Table 4.4). F_{ST} values from Mitta-Mitta above and below were not significantly different from F_{ST} values from Ovens above and below. Other F_{ST} values for more distantly separated stretches of rivers were comparable or higher to the F_{ST} values of groups separated by dams. In the more complex Snowy Rivers system (Figure 4.1), this simple F_{ST} analysis was not easy to interpret, so I relied upon the AIS analysis, the PCA analysis, and a regression of F_{ST} on dam age. Over all three river systems plus the two other dams from the literature, I observed a positive and significant relationship ($R^2 = 0.71$; p -value = 0.004) between F_{ST} and the number of platypus generations since the building of the dam (Figure 4.2).

Table 4.4 Genetic differentiation (F_{ST}) between rivers in different connectivity scenarios.

Region	River 1	River 2	F_{ST}	SD	SE	Connectivity scenario
Border Rivers	Tenterfield	Severn above dam	0.059	0.072	0.002	Separated by a river system
	Tenterfield	Severn below dam	0.063	0.080	0.002	Separated by a river system
	Severn below dam	Severn above dam	0.051	0.078	0.002	Separated by dam for 47 years (<i>Circa</i> 1969)
	Tenterfield above	Tenterfield below	0.009	0.044	0.001	No dam
Upper Murray Rivers	Ovens	Mitta-Mitta above dam	0.045	0.079	0.002	Contiguous river systems
	Ovens	Mitta-Mitta below dam	0.035	0.105	0.003	Contiguous river systems
	Mitta-Mitta above dam	Mitta-Mitta below dam	0.021	0.104	0.003	Separated by dam for 39 years (<i>Circa</i> 1979)
	Ovens above	Ovens below	0.001	0.070	0.002	No dam
Snowy Rivers	Snowy	Thredbo	0.024	0.044	0.001	Separated by dam for 50 years (<i>Circa</i> 1967)
	Snowy	Eucumbene above dam	0.042	0.101	0.003	Separated by dam for 59 years (<i>Circa</i> 1958)
	Snowy	Eucumbene below dam	0.040	0.059	0.001	Separated by dam for 50 years (<i>Circa</i> 1967)
	Thredbo	Eucumbene above dam	0.043	0.111	0.003	Separated by dam for 59 years (<i>Circa</i> 1958)
	Thredbo	Eucumbene below dam	0.030	0.060	0.002	Separated by lake for 50 years (<i>Circa</i> 1967)
	Eucumbene above dam	Eucumbene below dam	0.053	0.115	0.003	Separated by dam for 59 years (<i>Circa</i> 1958)

SD - standard deviation; SE - standard error.

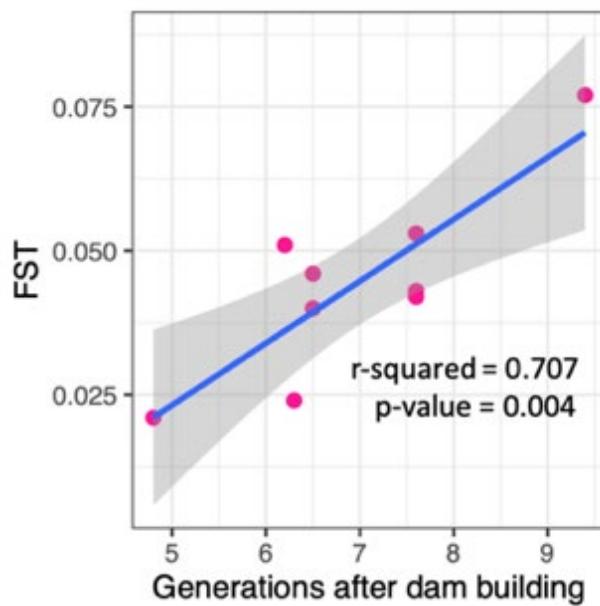


Figure 4.2 | Relationship between genetic differentiation (F_{ST}) between platypus groups separated by dams and the number of platypus generations (7.9 years, Pacifici *et al.* (2013) since the building of the dam. Genetic differentiation increased at a rate of 0.012 per generation. For this analysis, I also included previous F_{ST} estimates of platypus groups, separated by dams from Kolomyjec (2010; Nepean dam; height 82 m; built in 1935; $FST = 0.077$) and Furlan *et al.* (2013; Upper Yarra dam; height 89 m; built in 1957; $FST = 0.046$). These two studies used microsatellite data. I considered one platypus generation to be 7.9 years based on Pacifici *et al.* (2013).

Display of the first two principal components (PC) of the PCA analyses of the Border Rivers (Tenterfield Creek and Severn River) indicated three well separated clusters (Figure 4.3 a), with platypuses collected below and above the dam in the Severn River forming different groups. When the three first PCs are displayed (Figure 4.3 b), platypuses from the Severn River below the dam appeared to be more clustered than platypuses from either Tenterfield Creek or the Severn River above the dam. When comparing the AIS maps of this pair of rivers (Figure 4.3 c and Figure 4.3 d), the presence of peaks of high genetic differentiation near the dam in the Severn River and the relative absence of large peaks in Tenterfield Creek, is consistent with the expectation that dams are barriers to gene flow.

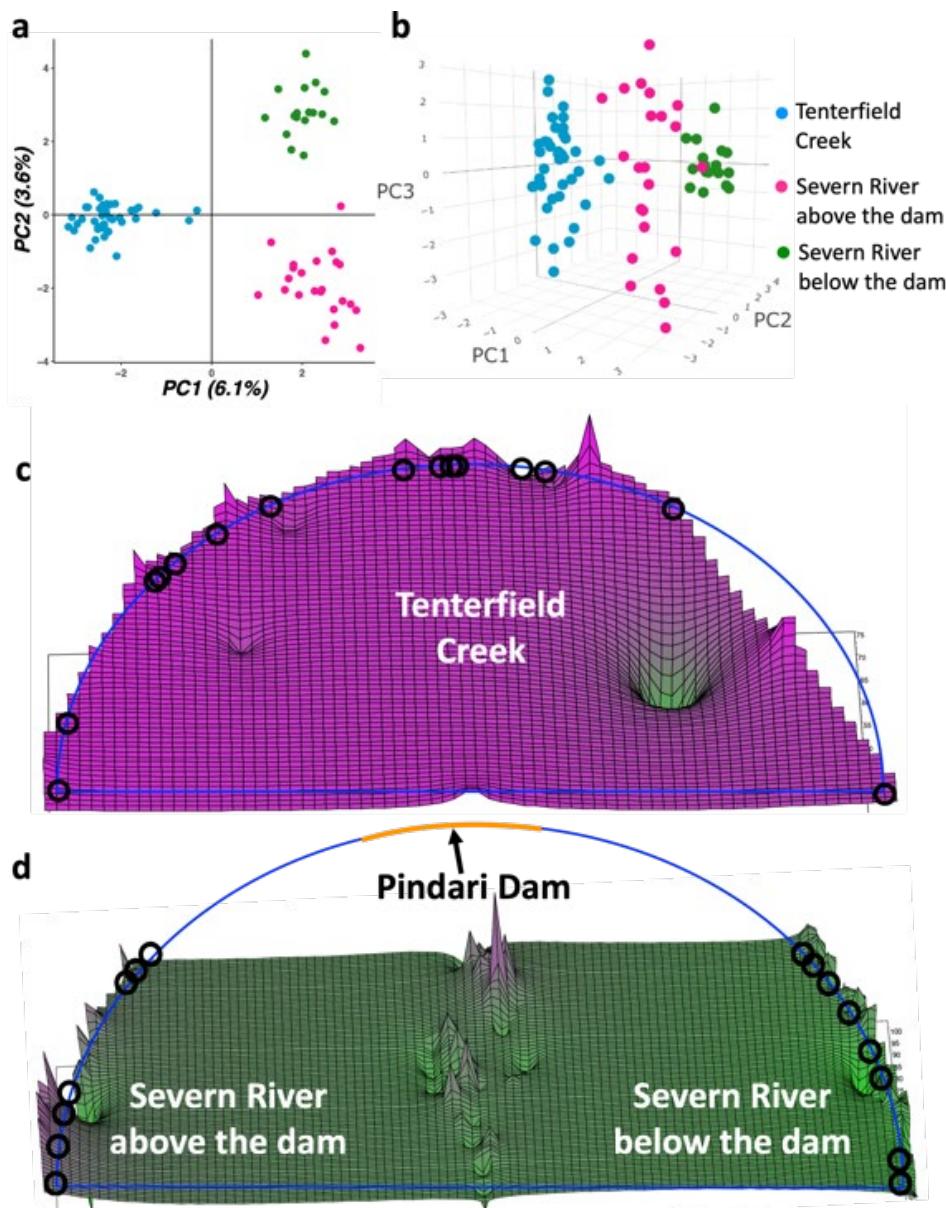


Figure 4.3 | Border Rivers: unregulated (no dam) Tenterfield Creek and regulated (dam) Severn River. **a)** Plot of the first two principal components (PCs) of PCA analyses (points represent platypus individuals) showing three well defined clusters. **b)** Plot of first three PCs where samples from Severn River below the dam are more clumped together than the other two groups. **c)** Maps of genetic differentiation between individuals, produced by the program Alleles in Space (Miller, 2005) for Tenterfield Creek; and **d)** Alleles in Space for Severn River, showing high genetic differentiation, represented by peaks, near the dam and relative absence of large peaks in Tenterfield Creek. The circumference of the semicircle representing sampling sites is in blue, with their locations represented by black circles and location of the Pindari Dam represented by an orange line.

In contrast to the Border Rivers, PCA analyses of the Upper Murray Rivers (Mitta Mitta and Ovens Rivers) did not show a clear clustering between platypuses below and above the dam in the Mitta-Mitta River (Figure 4.4 a). However, platypuses below the dam in the Mitta-Mitta River were more tightly clustered than individuals in the other groups (Figure 4.4 b). Peaks of high genetic differentiation in the AIS maps coincided with the location of the town of Bright in the Ovens River (Figure 4.4 c) and with the location of the dam in the Mitta-Mitta River (Figure 4.4 d).

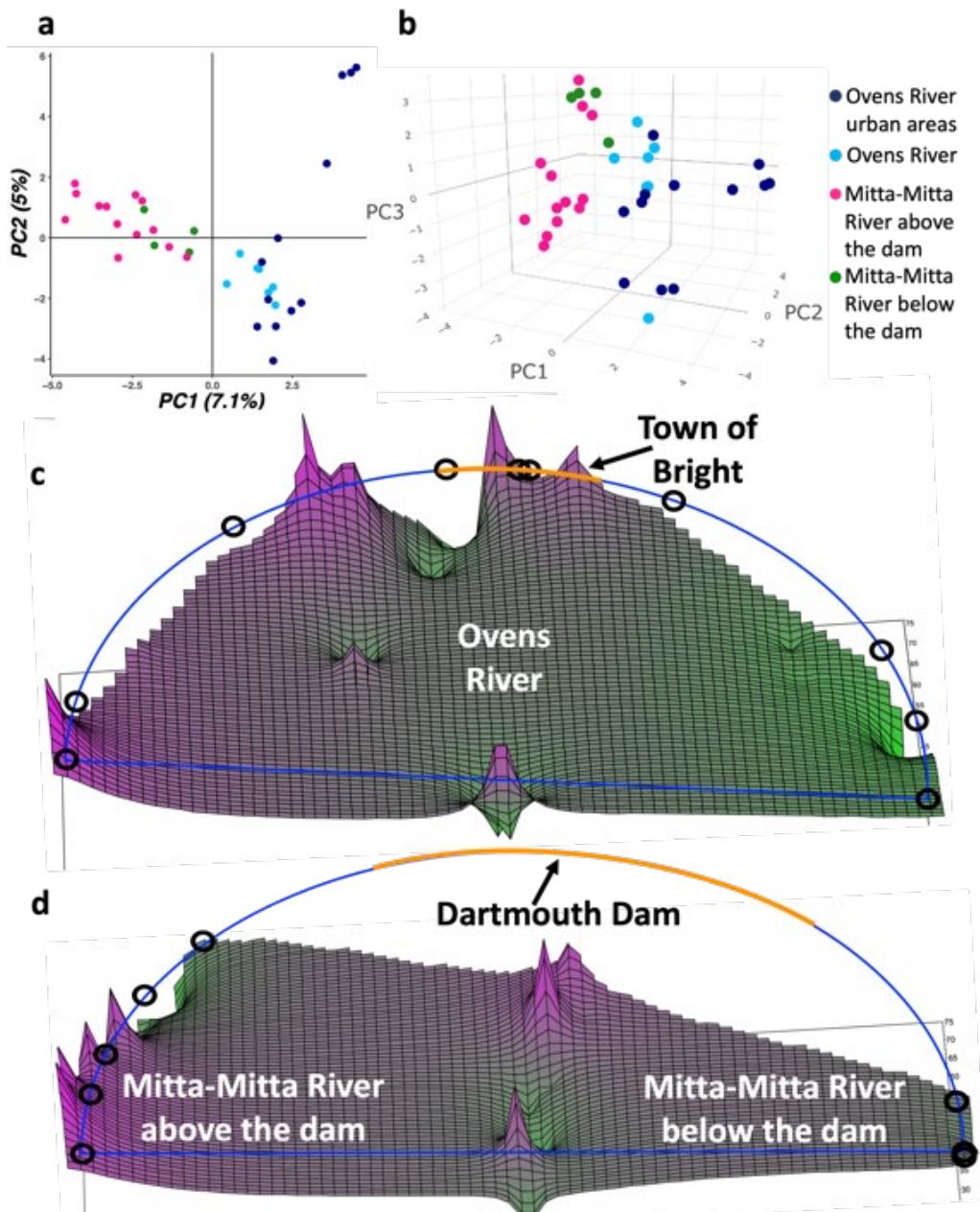


Figure 4.4 | Upper Murray Rivers: unregulated (no dam) Ovens and regulated (dam) Mitta-Mitta Rivers. **a)** Plot of the first two principal components (PCs) of PCA analyses (points represent platypus individuals) showing that samples below and above the dam do not form different clusters. **b)** Plot of first three PCs. **c)** Maps of genetic differentiation between individuals produced by the program Alleles in Space (Miller, 2005) in Ovens River; and **d)** Alleles in Space result for Mitta-Mitta River. Areas of high genetic differentiation are represented by peaks. The circumference of the semicircle that was used to place the sampling sites is shown in blue. The location of sampling sites is represented by black circles and location of the Dartmouth Dam and the town of Bright are represented by orange lines.

Snowy Rivers (Snowy, Thredbo and Eucumbene Rivers) did not follow the paired experimental design due to geographic constraints. PCA analyses showed that platypuses from the Snowy River formed a separated cluster to that of the Thredbo and Eucumbene Rivers (Figure 4.5 a), whereas platypuses from the two former rivers overlapped on the PCA plot. Notably platypuses from the Eucumbene River above the dam were closer to platypuses from Thredbo River compared to platypuses from Eucumbene River below the dam. Peaks of high genetic differentiation were near Jindabyne dam, separating Snowy and Eucumbene Rivers (Figure 4.5 c) and Thredbo and Snowy Rivers (Figure 4.5 d). Similarly, peaks of high genetic differentiation coincided with the location of a train station (Ski Tube) in the Thredbo River (Figure 4.5 d). Peaks of high genetic differentiation were also evident near Eucumbene dam (Figure 4.5 c).

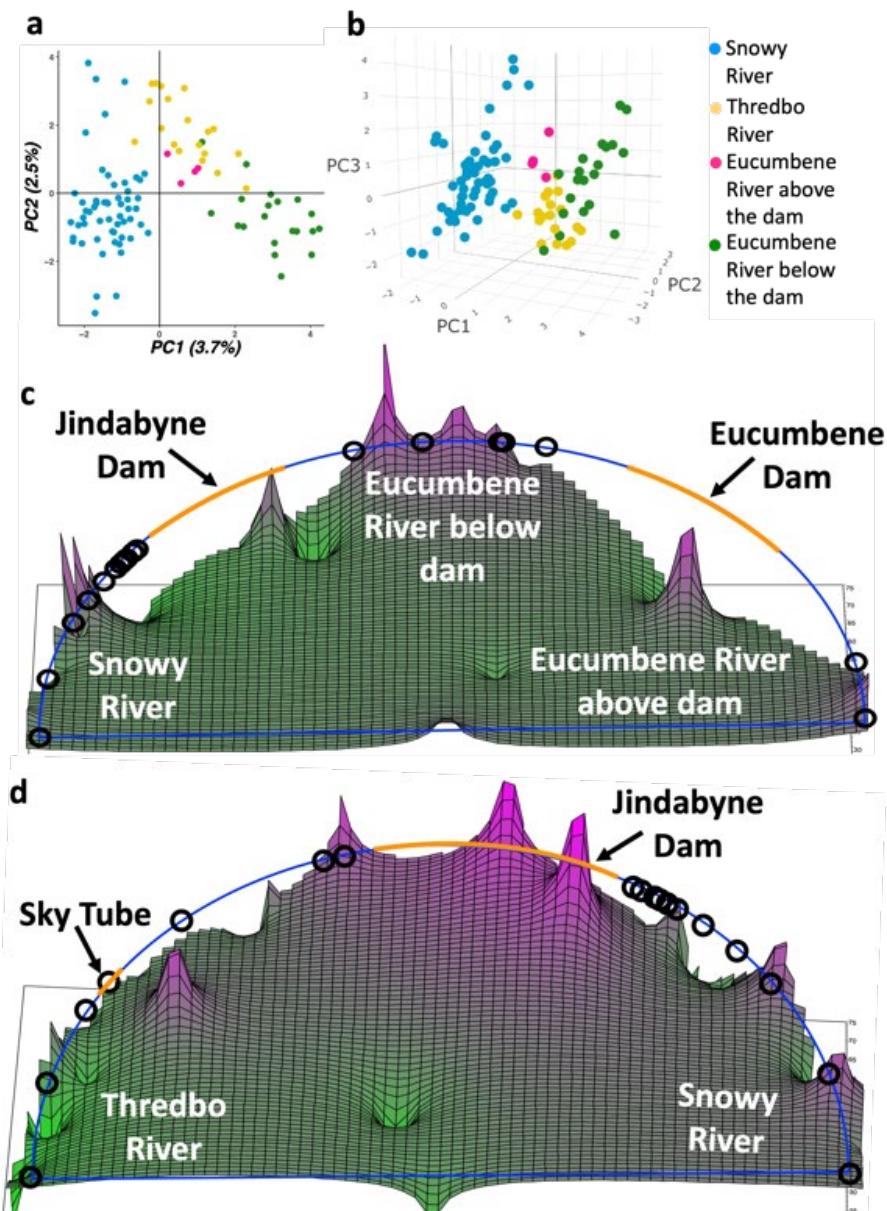


Figure 4.5 | Snowy Rivers: regulated (dam) Snowy, unregulated (no dam) Thredbo and regulated (dam) Eucumbene Rivers. These rivers do not follow the paired experimental design due to geographic constraints. **a)** Plot of the first two principal components (PCs) of PCA analyses (points represent platypus individuals) showing that samples from Snowy River form a separated cluster from the other locations. **b)** Plot of first three PCs. **c)** Maps of genetic differentiation between individuals produced by the program Alleles in Space (Miller, 2005) in Eucumbene and Snowy Rivers; and **d)** Alleles in Space result for Thredbo and Snowy Rivers. Areas of high genetic differentiation are represented by peaks. The circumference of the semicircle that was used to place the sampling sites is shown in blue. The location of sampling sites is represented by black circles and locations of the Jindabyne Dam, Eucumbene Dam and Sky Tube are represented by orange lines.

4.5 DISCUSSION

Dispersal and gene flow are key for the viability of natural populations, critical for ecological and evolutionary processes such as recolonisation, dispersal to suitable habitats, increased genetic diversity to avoid inbreeding depression, and adaptation (Garant *et al.*, 2007; Baguette *et al.*, 2013; Tigano & Friesen, 2016). My analyses suggest that dams are barriers to platypus dispersal and gene flow. Reflecting this, genetic differentiation increased proportionally with time after the building of a dam. There were implications for within river variation, reflecting whether dams were present, affecting dispersal.

Levels of genetic variation for sampled platypus, measured by expected heterozygosity (He), were significantly different between river regions: lower in northern (Border Rivers; mean $He = 0.135$) than in southern regions (Snowy and Upper Murray Rivers; mean $He = 0.143$; Table 4.3). This was consistent with 57 individuals that were genotyped across the platypus range, possibly reflecting dramatic climate change during the Pliocene and Pleistocene epochs (Martin *et al.*, 2018). This, reduced rainfall and transformed rainforests to drier, warmer, and more open ecosystems across eastern Australia, severely affecting east-northern and east-central regions of Australia (present Queensland; Bryant & Krosch, 2016). Throughout their range, the platypus comprise four major geographically defined genetic clusters: north Queensland, central Queensland, New South Wales and Tasmania (Martin *et al.*, 2018). The samples used in this study belong to the New South Wales cluster.

Note that estimates of heterozygosity (He) comparisons between studies should be regarded with caution, given He estimates vary with filtering protocols, sequencing technologies or estimation methods and there is no standardised methodology for the estimation of He using SNPs (but see Schmidt *et al.*, 2020). However, low genetic diversity within populations is an important factor reducing the ability of populations to respond to environmental change and thus increasing extinction risk (Frankham *et al.*, 2017). Genetic variation across the platypus samples (mean He across rivers = 0.141), was relatively low compared to other mammal species with larger population sizes using SNP data. The population size of the platypus has been estimated to be between 30,000-

300,000 mature individuals over the entire range (Woinarski & Burbidge, 2016). For instance, North American wolves (*Canis lupus*; $He = 0.17\text{--}0.28$; Cronin *et al.*, 2015). In contrast, platypus He was similar to He in other species with small population sizes, such as the threatened southern pygmy perch (*Nannoperca australis*; $He = 0.161$; Brauer *et al.*, 2016).

Similarly, estimates of inbreeding were generally high for platypus (Extended Data Figure 4.2 and Extended Data Table 4.2), with low estimates of effective population size (Ne ; range 32–320; Table 4.3). Platypus inbreeding estimates were high, consistent with previous analyses (Martin *et al.*, 2018), and even higher than other species with small population sizes, as measured by runs of homozygosity (F_{ROH}), such as bonobos (*Pan paniscus*; $F_{ROH} = \sim 0.05$) and western low-land gorillas (*Gorilla gorilla*; $F_{ROH} = \sim 0.04$; Prado-Martinez *et al.*, 2013). Platypus probably experienced a major population bottleneck that might have occurred around 10,000 years ago (Martin *et al.*, 2018), probably caused by climate fluctuations during the Pleistocene (Bryant & Krosch, 2016). Limited dispersal capabilities of the species probably also contribute to low estimates of genetic diversity and Ne and high estimates of inbreeding (Martin *et al.*, 2018). Accumulating evidence points out that inbreeding avoidance behaviour is less common, and even rare, in natural populations than previously realised (Szulkin *et al.*, 2013; de Boer *et al.*, 2021) and that inbreeding may have costs and benefits (Kokko & Ots, 2006).

Patterns of genetic diversity in groups sampled below dams were contrasting rather than uniform (Table 4.3). The group below the dam in the Severn River had significantly lower genetic variation than the group above the dam or in the unregulated Tenterfield Creek, while platypuses below the dam in the Snowy River had similar diversity to groups in the unregulated Thredbo River. Furthermore, the platypus group in the Eucumbene River below the dam had the lowest He and Ne among the rivers in this region, indicative of inbreeding, despite relatively high capture rates and abundance compared to other sampled parts of the Snowy River (Hawke *et al.*, 2021). Possibly this population is isolated by the dam upstream on the Eucumbene River and Jindabyne Dam downstream (Figure 4.1). Small samples sizes on the Mitta-Mitta River below the dam and

Eucumbene above the dam were not sufficient to understand potential population effects ($n = 4$ in both rivers).

There was no clear indication that populations separated by dams had lower genetic variation than those on unregulated rivers. However, we could not discard the possibility that the isolation caused by dams has contributed to loss of adaptive genetic variation. The small random sample of the total genetic variation across the genome used in this study might (Portanier *et al.*, 2019) or might not (Yıldırım *et al.*, 2018) reflect the same patterns as adaptive genetic variation.

In relation to whether dams affect the connectivity between platypus groups, F_{ST} values were higher when there was a dam, and some F_{ST} values between groups, separated by a dam, were as high as F_{ST} values between groups in different rivers (Table 4.4). Additionally, I found a significant association between F_{ST} and the number of platypus generations since dam construction ($R^2 = 0.71$, p-value < 0.05; Figure 4.2), suggesting that F_{ST} increases at a rate of 0.012 by generation. Note that the two oldest dams (Nepean dam built in 1935; Kolomyjec 2010 and Upper Yarra dam built in 1957; Furlan *et al.*, 2013) were analysed with a different type of molecular markers (microsatellites not SNPs), however, recent research indicates that estimates of F_{ST} using SNPs and microsatellites are comparable (Lemopoulos *et al.*, 2019; Sunde *et al.*, 2020). F_{ST} values in the Snowy Rivers were higher between groups separated by the Jindabyne Dam (Eucumbene below dam / Snowy; $F_{ST} = 0.04$) than between groups separated by the Jindabyne reservoir but not a dam (Eucumbene below dam / Thredbo; $F_{ST} = 0.03$). This observation suggests that gene flow might have occurred across the Jindabyne reservoir.

Further, areas of high genetic differentiation coincided with the location of dams (Figure 4.3 to Figure 4.5). Together with my F_{ST} results, these observations provided strong evidence that dams altered the connectivity between platypus groups. In addition, areas of high differentiation also coincided with human disturbances, such as the town of Bright in the Ovens River (Figure 4.4 c) and a train station in the Thredbo River (Figure 4.5 d). This might also indicate that urbanised areas are dispersal barriers, consistent

with previous work showing that platypuses are highly susceptible to disturbances associated with human activities (Grant, 1998; Serena & Williams, 1998).

Overall, this builds on evidence that high dams and their associated waterbodies may be considerable barriers for platypuses. Even though platypuses can move substantial distances (e.g., male juveniles can move > 40 km; Serena & Williams, 2013), the effect of dams on genetic differentiation was considerable. Some of the long-term effects of dams might be reduced by rare dispersal events between rivers (Kolomyjec *et al.*, 2014). Contrastingly, dams did not increase genetic differentiation in the blackfish (*Gadopsis marmoratus*), a non-migratory and low-mobility freshwater fish species in eastern Australia, possibly because population sizes were not small enough to observe this effect (Coleman *et al.*, 2018).

Dams represent dispersal barriers for most freshwater species (Søndergaard & Jeppesen, 2007; Nislow *et al.*, 2011), requiring mitigation strategies to offset negative demographic impacts. For instance, human mediated relocation of individuals between populations has been implemented successfully to limit the effects of population isolation and small population size (Hoffmann *et al.*, 2020). A common rule of thumb in conservation suggests that one dispersing individual per generation would minimise the effects of population isolation (Mills & Allendorf, 1996). Another strategy to improve connectivity between populations, despite some limitations and caveats, is the construction of dam passages that increase dispersal of freshwater species including platypuses (Broadhurst *et al.*, 2013; Brown *et al.*, 2013; Silva *et al.*, 2018), although there are adverse consequences of connectivity, such as disease risks (Sainsbury & Vaughan-Higgins, 2012).

I highlight that careful management is required if a type of balancing selection, called associative overdominance (AOD; Chapter 2), is responsible for the maintenance of higher-than-expected levels of genetic diversity in platypuses in the Mitta-Mitta River. In particular, strategies involving induced dispersal, could potentially have adverse outcomes if this evolutionary process is not properly accounted for (Chapter 2). Introducing new individuals from other populations in an attempt to restore gene flow, could perversely cause heterozygosity to crash if haplotypes maintained by AOD are

broken (Chapter 2). This breakage of haplotypes under AOD could cause a decline in fitness and population growth (Chapter 2). Note, however, that the low sample size of this group ($n = 4$) might be the most plausible explanation for this observation.

To minimize detrimental effects of river regulation on the platypus and other species, new dams within the distribution of platypus need to be avoided in favour of other water saving strategies. Further research is needed to investigate the effects of dams on adaptive genetic variation to determine whether platypuses are able to respond to environmental change caused by dams, as other freshwater species including cichlid fishes (Gilbert, Akama, *et al.*, 2020), Chinook salmon (*Oncorhynchus tshawytscha*; Waples *et al.*, 2017) and Amazon river prawn (*Macrobrachium amazonicum*; Silva *et al.*, 2020).

4.6 CONCLUSION

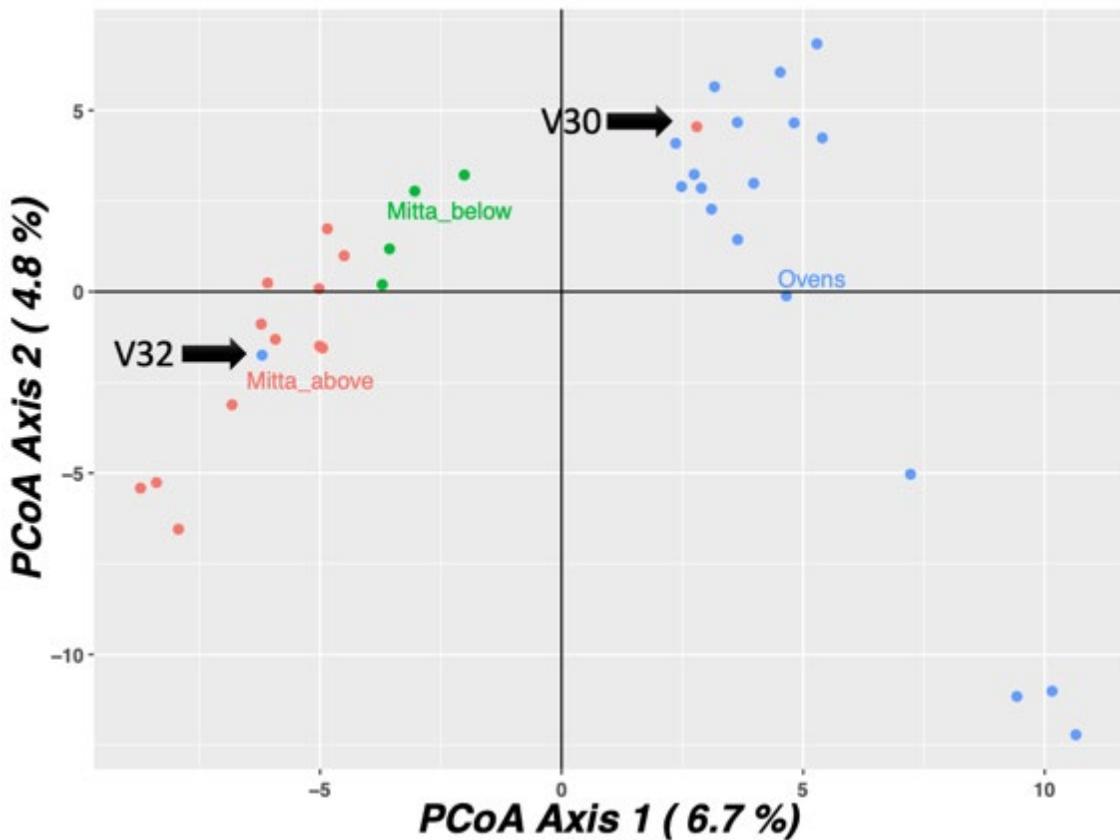
I compared regulated rivers, with dams, to adjacent unregulated rivers with no dams and identified that dams were barriers to movement of platypuses within a river system, reflected in genetic variation. Dams restricted dispersal and gene flow between groups and therefore increased the possibility of inbreeding depression, loss of adaptive genetic variation, failure to recolonise areas where local extinctions have occurred and failure to disperse to areas with more suitable conditions. These are all expected to lower the long-term viability of the platypus (Bino *et al.*, 2020). My analyses reinforce the growing evidence on the negative impacts of dams on platypus populations. These studies are relevant to inform the decision-making process of conservation managers and could be used in viability analysis and decision analysis (Drechsler & Burgman, 2004) to develop strategies that ensure the long-term persistence of the unique platypus.

4.7 AUTHORS CONTRIBUTIONS

Bill Sherwin, Gilad Bino, Richard Kingsford, and Jaime Gongora conceived the project and acquired the research funds; Harvinder Sidhu and Bill Sherwin supervised Luis Mijangos; Jenna Day, Kimberly Noel Dias and Jaime Gongora performed DNA extraction; Gilad Bino and Tahneal Hawke carried out field work and collected samples; Luis Mijangos analysed the data; Luis Mijangos wrote the manuscript with support from Bill

Sherwin; Tom Grant initiated the query about what effect dams may be having on the genetics of the platypus, which eventually resulted in one of the main thrusts of this platypus project. All authors discussed the results and contributed to the final manuscript.

4.8 EXTENDED DATA

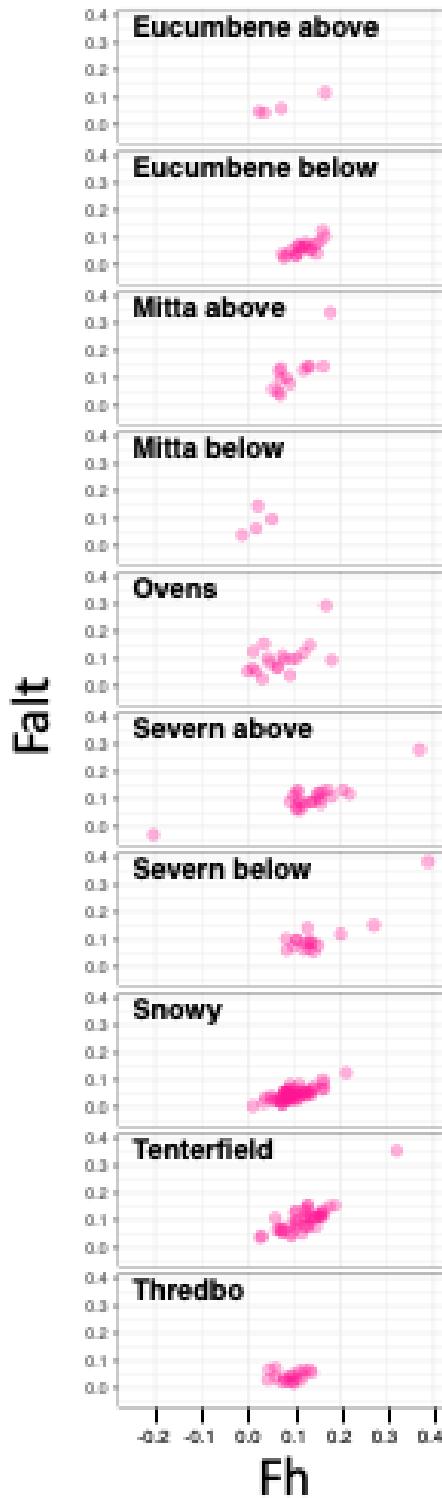


Extended Data Figure 4.1 | Principal coordinates analysis (PCA) based on SNP data.

Figure shows that the individuals V30 and V32 (indicated by arrows), each collected in a different river, showed contrasting genetic patterns relative to individuals collected in the same river. Relatedness analyses revealed these two samples had closer relatives in the opposite river (Extended Data Table 4.1). I also noted that the sampling locations of these individuals were separated by 46 Km, steep mountainous terrain, and a river system. Given these circumstances, I considered that two dispersal events were unlikely and concluded that these two samples were mislabelled and therefore I assigned them to the presumed correct river and site.

Extended Data Table 4.1 Results of relatedness analyses, based on SNP data, performed in the R package *related* (Pew *et al.*, 2015) to identify any recapture not identified or mislabelling either in the field or the laboratory. Individual codes are as in Extended Data Table 4.2.

Ind 1	Population	Ind 2	Population	Coefficient of relatedness	Probable relationship
V30	Ovens	V1	Mitta above	0.1741	First cousins
V30	Ovens	V4	Mitta above	0.0862	Second cousins
V30	Ovens	V5	Mitta above	0.0519	Second cousins
V30	Ovens	V31	Mitta above	0.0239	Second cousins
V30	Ovens	V3	Mitta above	0.0185	Second cousins
V30	Ovens	V33	Mitta above	0.0133	Second cousins
V32	Mitta above	V29	Ovens	0.0146	Second cousins
T3	Tenterfield	T5	Tenterfield	0.9949	Recapture
T28	Tenterfield	T42	Tenterfield	0.9739	Recapture



Extended Data Figure 4.2 | Summary of inbreeding estimates. Inbreeding was estimated for each individual using two different statistics: F_h which is a deviation in homozygosity from its Hardy–Weinberg expectation using the software PLINK (Purcell *et al.*, 2007); and, F_{alt} where homozygous loci are weighted with the inverse of their allele frequency using the software GCTA (Yang *et al.*, 2011). Refer to Extended Data Table 4.2 for inbreeding estimates for each individual.

Extended Data Table 4.2 General information per individual. H_o - observed heterozygosity; F_h - inbreeding estimates with program PLINK (Purcell *et al.*, 2007); F_{alt} inbreeding estimates with program GCTA (Yang *et al.*, 2011); NA - not available.

ID	Group	Sex	Age class	Weight	Latitude	Longitude	H_o	F_h	F_{alt}
E65	Eucumbene above	Female	Juvenile	0.58	-35.918995	148.540494	0.150	0.035	0.044
E68	Eucumbene above	Female	Adult	0.93	-35.887646	148.51589	0.130	0.166	0.117
E69	Eucumbene above	Female	Juvenile	0.56	-35.887646	148.51589	0.152	0.024	0.048
E70	Eucumbene above	Male	Adult	1.36	-35.887646	148.51589	0.145	0.071	0.060
E1	Eucumbene below	Male	Sub-adult	1.12	-36.164286	148.622989	0.136	0.128	0.068
E10	Eucumbene below	Female	Adult	0.94	-36.222028	148.632604	0.140	0.103	0.037
E11	Eucumbene below	Female	Adult	0.84	-36.253799	148.620099	0.130	0.165	0.104
E12	Eucumbene below	Female	Adult	0.82	-36.253799	148.620099	0.133	0.148	0.041
E18	Eucumbene below	Male	Adult	1.27	-36.181475	148.633916	0.134	0.138	0.057
E19	Eucumbene below	Male	Adult	1.30	-36.183226	148.634856	0.136	0.125	0.081
E2	Eucumbene below	Female	Adult	NA	-36.164286	148.622989	0.139	0.110	0.071
E20	Eucumbene below	Male	Sub-adult	1.25	-36.183226	148.634856	0.132	0.153	0.084
E28	Eucumbene below	Female	Juvenile	0.75	-36.180689	148.634867	0.140	0.103	0.031
E29	Eucumbene below	Female	Adult	0.91	-36.180689	148.634867	0.143	0.083	0.036
E3	Eucumbene below	Female	Adult	0.87	-36.164286	148.622989	0.131	0.161	0.126
E4	Eucumbene below	Female	Adult	0.83	-36.181475	148.633916	0.137	0.121	0.067
E5	Eucumbene below	Male	Adult	1.58	-36.181475	148.633916	0.141	0.093	0.052
E6	Eucumbene below	Male	Adult	1.48	-36.180689	148.634867	0.144	0.073	0.039
E63	Eucumbene below	Female	Adult	0.78	-36.180689	148.634867	0.139	0.110	0.043
E64	Eucumbene below	Female	Juvenile	0.74	-36.181475	148.633916	0.138	0.116	0.064
E7	Eucumbene below	Male	Adult	1.56	-36.181475	148.633916	0.139	0.106	0.069
E74	Eucumbene below	Male	Adult	1.47	-36.180689	148.634867	0.133	0.145	0.076

E8	Eucumbene below	Female	Adult	0.77	-36.181475	148.633916	0.135	0.134	0.051
E9	Eucumbene below	Male	Adult	1.09	-36.222028	148.632604	0.144	0.077	0.025
V1	Mitta above	Female	Juvenile	NA	-36.908784	147.622772	0.143	0.083	0.100
V2	Mitta above	Female	Juvenile	0.59	-36.908784	147.622772	0.142	0.090	0.078
V21	Mitta above	Male	Adult	0.93	-36.946087	147.607521	0.146	0.066	0.126
V22	Mitta above	Male	Adult	1.14	-36.946087	147.607521	0.128	0.177	0.338
V23	Mitta above	Male	Adult	1.18	-36.891481	147.631409	0.147	0.054	0.059
V24	Mitta above	Male	Adult	1.28	-36.891481	147.631409	0.135	0.134	0.140
V25	Mitta above	Female	Juvenile	0.51	-36.891481	147.631409	0.145	0.069	0.095
V3	Mitta above	Male	Juvenile	NA	-36.806872	147.66193	0.136	0.129	0.142
V30	Mitta above	Male	Adult	1.68	-36.699684	146.9096	0.145	0.071	0.135
V31	Mitta above	Female	Juvenile	0.46	-36.891481	147.631409	0.137	0.120	0.128
V33	Mitta above	Female	Adult	0.76	-36.891481	147.631409	0.146	0.065	0.049
V4	Mitta above	Male	Juvenile	NA	-36.851574	147.637379	0.130	0.162	0.142
V5	Mitta above	Male	Juvenile	NA	-36.891481	147.631409	0.145	0.069	0.039
V20	Mitta below	Male	Juvenile	0.82	-36.508674	147.411724	0.148	0.051	0.098
V35	Mitta below	Male	Adult	1.58	-36.508228	147.412276	0.153	0.018	0.064
V36	Mitta below	Male	Adult	1.64	-36.514876	147.374988	0.152	0.022	0.145
V37	Mitta below	Male	Adult	1.36	-36.51879	147.377085	0.158	-0.013	0.039
V10	Ovens	Female	Adult	0.93	-36.725589	146.963083	0.144	0.074	0.113
V11	Ovens	Female	Adult	0.88	-36.725589	146.963083	0.137	0.120	0.122
V12	Ovens	Female	Adult	0.99	-36.725589	146.963083	0.148	0.047	0.083
V13	Ovens	Male	Juvenile	0.90	-36.820253	147.072079	0.130	0.168	0.293
V14	Ovens	Female	Adult	1.02	-36.717882	146.945064	0.135	0.134	0.151
V15_1	Ovens	Female	Adult	1.27	-36.717882	146.945064	0.128	0.180	0.095

V15_2	Ovens	Male	Adult	1.46	-36.803133	147.057653	0.146	0.063	0.071
V17	Ovens	Female	Juvenile	0.78	-36.629973	146.821417	0.153	0.018	0.052
V18	Ovens	Male	Juvenile	0.76	-36.699684	146.9096	0.154	0.012	0.127
V19	Ovens	Female	Adult	0.87	-36.699684	146.9096	0.142	0.090	0.038
V26	Ovens	Male	Juvenile	1.02	-36.638183	146.851507	0.146	0.062	0.066
V27	Ovens	Female	Juvenile	0.82	-36.638183	146.851507	0.141	0.094	0.098
V28	Ovens	Female	Adult	1.03	-36.656708	146.856924	0.156	-0.001	0.054
V29	Ovens	Male	Adult	1.34	-36.738527	146.975783	0.139	0.105	0.102
V32	Ovens	Male	Adult	1.36	-36.891481	147.631409	0.151	0.031	0.025
V6	Ovens	Female	Adult	NA	-36.699684	146.9096	0.144	0.077	0.097
V7	Ovens	Female	Adult	NA	-36.699684	146.9096	0.154	0.010	0.065
V8	Ovens	Female	Adult	NA	-36.699684	146.9096	0.149	0.043	0.103
V9	Ovens	Male	Adult	1.56	-36.745153	147.016953	0.151	0.034	0.156
SUS19	Severn above	Male	Adult	0.93	-29.474119	151.485041	0.128	0.180	0.108
SUS20	Severn above	Female	Adult	1.34	-29.474119	151.485041	0.132	0.154	0.098
SUS21	Severn above	Female	Adult	1.13	-29.474119	151.485041	0.139	0.109	0.128
SUS22	Severn above	Male	Adult	1.84	-29.49832	151.562924	0.129	0.172	0.132
SUS23	Severn above	Male	Adult	1.80	-29.493532	151.543319	0.131	0.160	0.126
SUS24	Severn above	Male	Juvenile	1.15	-29.493532	151.543319	0.139	0.105	0.063
SUS25	Severn above	Female	Adult	1.22	-29.506811	151.589001	0.139	0.109	0.090
SUS26	Severn above	Male	Adult	1.86	-29.506811	151.589001	0.141	0.098	0.113
SUS27	Severn above	Female	Adult	1.16	-29.506811	151.589001	0.139	0.105	0.073
SUS28	Severn above	Female	Adult	1.23	-29.506811	151.589001	0.139	0.105	0.128
SUS29	Severn above	Male	Adult	1.47	-29.506811	151.589001	0.133	0.149	0.116
SUS30	Severn above	Male	Adult	1.74	-29.506811	151.589001	0.136	0.126	0.084

SUS31	Severn above	Male	Adult	2.01	-29.502849	151.611012	0.122	0.219	0.118
SUS32	Severn above	Male	Adult	1.73	-29.463007	151.478227	0.134	0.137	0.088
SUS33	Severn above	Female	Adult	1.21	-29.463007	151.478227	0.131	0.157	0.083
SUS34	Severn above	Male	Adult	1.84	-29.463007	151.478227	0.138	0.113	0.062
SUS35	Severn above	Female	Adult	1.22	-29.463007	151.478227	0.098	0.371	0.280
SUS36	Severn above	Female	Adult	1.28	-29.456671	151.464321	0.134	0.141	0.089
SUS37	Severn above	Male	Adult	1.64	-29.456671	151.464321	0.187	-0.204	-0.030
SUS38	Severn above	Male	Adult	1.45	-29.456671	151.464321	0.142	0.090	0.089
SUS39	Severn above	Female	Juvenile	0.96	-29.456671	151.464321	0.132	0.154	0.120
SUS40	Severn above	Male	Adult	1.84	-29.456671	151.464321	0.138	0.117	0.082
SUS41	Severn above	Male	Adult	2.08	-29.456671	151.464321	0.124	0.205	0.132
SDS10	Severn below	Male	Adult	1.64	-29.179456	151.111784	0.143	0.082	0.104
SDS11	Severn below	Female	Adult	1.27	-29.179456	151.111784	0.141	0.098	0.074
SDS12	Severn below	Male	Adult	1.68	-29.179456	151.111784	0.139	0.105	0.101
SDS13	Severn below	Male	Adult	1.43	-29.303547	151.138189	0.134	0.142	0.056
SDS14	Severn below	Male	Adult	1.83	-29.302878	151.124002	0.113	0.272	0.152
SDS15	Severn below	Male	Juvenile	1.05	-29.302878	151.124002	0.134	0.140	0.091
SDS16	Severn below	Male	Adult	2.11	-29.249669	151.133357	0.135	0.133	0.088
SDS17	Severn below	Female	Adult	1.24	-29.249669	151.133357	0.143	0.083	0.062
SDS18	Severn below	Female	Adult	1.22	-29.249669	151.133357	0.095	0.389	0.383
SDS2	Severn below	Male	Adult	1.66	-29.296528	151.144227	0.137	0.121	0.063
SDS3	Severn below	Female	Adult	1.07	-29.296528	151.144227	0.136	0.129	0.141
SDS4	Severn below	Female	Adult	1.24	-29.296528	151.144227	0.137	0.122	0.086
SDS5	Severn below	Female	Adult	1.39	-29.189063	151.132304	0.135	0.133	0.096
SDS6	Severn below	Male	Adult	1.80	-29.189063	151.132304	0.125	0.200	0.120

SDS7	Severn below	Male	Adult	1.77	-29.232346	151.114818	0.132	0.150	0.079
SDS8	Severn below	Male	Adult	0.84	-29.232346	151.114818	0.135	0.133	0.066
SDS9	Severn below	Female	Adult	1.34	-29.27808	151.117899	0.139	0.105	0.099
E100	Snowy	Male	Adult	1.03	-36.501988	148.831949	0.136	0.128	0.056
E13	Snowy	Male	Adult	1.11	-36.467166	148.693223	0.139	0.109	0.036
E14	Snowy	Male	Adult	1.53	-36.467166	148.693223	0.144	0.073	0.011
E21	Snowy	Female	Adult	0.78	-36.447367	148.648261	0.142	0.090	0.061
E22	Snowy	Female	Adult	0.78	-36.447367	148.648261	0.131	0.158	0.064
E23	Snowy	Female	Adult	0.86	-36.480859	148.774219	0.136	0.126	0.041
E24	Snowy	Female	Adult	0.85	-36.480859	148.774219	0.141	0.094	0.027
E25	Snowy	Male	Adult	1.55	-36.480859	148.774219	0.151	0.034	0.016
E26	Snowy	Female	Adult	0.93	-36.447367	148.648261	0.140	0.101	0.067
E27	Snowy	Female	Adult	0.80	-36.447367	148.648261	0.146	0.065	0.029
E33	Snowy	Female	Adult	1.00	-36.480859	148.774219	0.138	0.114	0.055
E34	Snowy	Male	Adult	1.45	-36.480859	148.774219	0.144	0.074	0.012
E36	Snowy	Female	Juvenile	0.45	-36.447412	148.643942	0.144	0.078	0.052
E37	Snowy	Male	Adult	1.29	-36.447412	148.643942	0.141	0.093	0.036
E38	Snowy	Male	Adult	1.52	-36.447412	148.643942	0.134	0.140	0.069
E39	Snowy	Female	Adult	0.93	-36.447412	148.643942	0.139	0.106	0.024
E40	Snowy	Male	Adult	1.21	-36.439499	148.63292	0.154	0.011	0.004
E41	Snowy	Male	Juvenile	0.52	-36.439499	148.63292	0.141	0.094	0.050
E42	Snowy	Male	Juvenile	0.47	-36.443192	148.636579	0.137	0.120	0.046
E43	Snowy	Female	Adult	0.81	-36.443192	148.636579	0.135	0.130	0.052
E44	Snowy	Female	Juvenile	0.46	-36.443192	148.636579	0.143	0.081	0.043
E45	Snowy	Male	Adult	1.42	-36.443192	148.636579	0.134	0.137	0.054

E46	Snowy	Female	Adult	NA	-36.443192	148.636579	0.139	0.105	0.050
E47	Snowy	Female	Adult	0.85	-36.446414	148.652903	0.136	0.125	0.053
E48	Snowy	Female	Adult	0.89	-36.446267	148.661103	0.144	0.077	0.055
E49	Snowy	Male	Adult	1.40	-36.446267	148.661103	0.148	0.050	0.040
E50	Snowy	Male	Adult	1.33	-36.446267	148.661103	0.137	0.118	0.061
E51	Snowy	Female	Adult	0.89	-36.446267	148.661103	0.134	0.141	0.071
E52	Snowy	Female	Adult	0.72	-36.446414	148.652903	0.141	0.093	0.041
E53	Snowy	Female	Adult	0.89	-36.447367	148.648261	0.130	0.164	0.067
E54	Snowy	Male	Adult	1.48	-36.447367	148.648261	0.139	0.110	0.037
E57	Snowy	Female	Adult	0.71	-36.480859	148.774219	0.143	0.081	0.028
E58	Snowy	Male	Adult	1.76	-36.480859	148.774219	0.145	0.071	0.026
E59	Snowy	Female	Juvenile	0.48	-36.480859	148.774219	0.140	0.102	0.040
E67	Snowy	Male	Adult	1.47	-36.452387	148.677242	0.145	0.070	0.019
E71	Snowy	Male	Adult	1.59	-36.480859	148.774219	0.123	0.212	0.126
E72	Snowy	Female	Adult	0.96	-36.480859	148.774219	0.139	0.106	0.050
E73	Snowy	Female	Adult	0.94	-36.480859	148.774219	0.143	0.081	0.023
E79	Snowy	Male	Adult	NA	-36.501988	148.831949	0.141	0.094	0.028
E80	Snowy	Male	Adult	1.23	-36.501988	148.831949	0.143	0.083	0.026
E81	Snowy	Female	Adult	NA	-36.501988	148.831949	0.138	0.111	0.083
E82	Snowy	Male	Adult	1.37	-36.501988	148.831949	0.142	0.091	0.083
E83	Snowy	Female	Adult	1.03	-36.501988	148.831949	0.142	0.087	0.040
E84	Snowy	Male	Adult	1.53	-36.501988	148.831949	0.150	0.038	0.034
E85	Snowy	Male	Adult	1.67	-36.46894	148.722478	0.142	0.091	0.037
E86	Snowy	Female	Adult	0.72	-36.46894	148.722478	0.142	0.087	0.052
E87	Snowy	Female	Adult	0.81	-36.46894	148.722478	0.131	0.161	0.101

E88	Snowy	Male	Adult	NA	-36.46894	148.722478	0.146	0.065	0.024
E89	Snowy	Male	Adult	1.18	-36.46894	148.722478	0.134	0.137	0.043
E93	Snowy	Male	Sub-adult	0.64	-36.439499	148.63292	0.143	0.081	0.057
E94	Snowy	Male	Adult	NA	-36.480859	148.774219	0.142	0.086	0.066
E95	Snowy	Female	Adult	0.70	-36.501988	148.831949	0.144	0.077	0.054
E96	Snowy	Male	Sub-adult	1.10	-36.501988	148.831949	0.147	0.055	0.030
E97	Snowy	Male	Adult	NA	-36.447367	148.648261	0.147	0.054	0.021
E98	Snowy	Male	Adult	1.15	-36.447367	148.648261	0.143	0.083	0.057
E99	Snowy	Female	Adult	0.68	-36.480859	148.774219	0.130	0.162	0.086
T1	Tenterfield	Female	Juvenile	0.90	-29.036629	151.820534	0.137	0.122	0.103
T10	Tenterfield	Female	Juvenile	0.57	-29.035987	151.759721	0.137	0.122	0.076
T11	Tenterfield	Female	Juvenile	0.82	-29.035987	151.759721	0.135	0.134	0.110
T12	Tenterfield	Female	Adult	1.05	-29.035987	151.759721	0.133	0.146	0.078
T13	Tenterfield	Female	Adult	0.98	-28.984905	151.951568	0.138	0.117	0.053
T14	Tenterfield	Male	Adult	1.60	-28.984905	151.951568	0.127	0.187	0.153
T15	Tenterfield	Female	Juvenile	0.82	-28.984905	151.951568	0.129	0.174	0.149
T16	Tenterfield	Female	Adult	0.91	-28.984905	151.951568	0.133	0.146	0.110
T17	Tenterfield	Male	Juvenile	0.85	-28.984905	151.951568	0.145	0.069	0.068
T18	Tenterfield	Female	Juvenile	0.67	-29.022531	151.86925	0.135	0.130	0.077
T19	Tenterfield	Male	Adult	1.79	-29.022531	151.86925	0.136	0.126	0.110
T2	Tenterfield	Male	Adult	1.79	-29.030611	151.851523	0.131	0.160	0.111
T20	Tenterfield	Male	Adult	1.69	-29.002064	151.995101	0.144	0.073	0.058
T21	Tenterfield	Female	Adult	1.02	-29.002064	151.995101	0.132	0.156	0.118
T22	Tenterfield	Male	Juvenile	1.07	-29.002064	151.995101	0.140	0.102	0.073
T23	Tenterfield	Male	Adult	1.58	-29.030611	151.851523	0.131	0.160	0.135

T24	Tenterfield	Male	Adult	1.55	-29.014065	151.864932	0.134	0.141	0.106
T25	Tenterfield	Male	Adult	1.51	-29.014065	151.864932	0.141	0.093	0.053
T26	Tenterfield	Male	Adult	1.50	-28.989203	151.701152	0.147	0.058	0.107
T27	Tenterfield	Female	Adult	0.91	-28.989203	151.701152	0.131	0.161	0.125
T28	Tenterfield	Female	Juvenile	0.79	-29.026713	151.746975	0.152	0.027	0.038
T29	Tenterfield	Female	Adult	0.98	-29.026713	151.742656	0.140	0.103	0.133
T30	Tenterfield	Female	Adult	0.94	-29.031212	151.742656	0.140	0.101	0.126
T31	Tenterfield	Female	Adult	0.79	-29.031212	151.742656	0.151	0.030	0.043
T32	Tenterfield	Female	Adult	1.56	-29.031212	151.742656	0.135	0.130	0.155
T33	Tenterfield	Male	Adult	1.39	-29.027057	151.724683	0.146	0.062	0.065
T34	Tenterfield	Female	Adult	0.83	-29.027057	151.724683	0.136	0.129	0.151
T35	Tenterfield	Male	Adult	1.62	-29.027057	151.724683	0.144	0.074	0.068
T36	Tenterfield	Female	Adult	1.00	-29.027057	151.724683	0.132	0.154	0.116
T38	Tenterfield	Female	Adult	1.01	-29.014825	151.722949	0.141	0.093	0.043
T39	Tenterfield	Female	Adult	0.98	-29.014825	151.722949	0.106	0.321	0.353
T4	Tenterfield	Male	Juvenile	0.88	-29.011746	151.86722	0.135	0.132	0.084
T40	Tenterfield	Male	Juvenile	0.82	-28.959487	151.544385	0.144	0.074	0.058
T41	Tenterfield	Female	Adult	1.08	-28.959487	151.544385	0.139	0.105	0.099
T5	Tenterfield	Female	Juvenile	0.62	-29.011746	151.86722	0.137	0.122	0.138
T6	Tenterfield	Female	Juvenile	0.93	-29.011746	151.86722	0.132	0.152	0.114
T7	Tenterfield	Female	Juvenile	0.91	-29.035987	151.759721	0.141	0.094	0.085
T8	Tenterfield	Female	Adult	1.23	-29.035987	151.759721	0.139	0.105	0.087
T9	Tenterfield	Male	Juvenile	0.11	-29.035987	151.759721	0.135	0.130	0.129
E15	Thredbo	Female	Adult	0.61	-36.44681	148.424506	0.140	0.099	0.017
E16	Thredbo	Male	Adult	1.10	-36.44681	148.424506	0.141	0.097	0.024

E17	Thredbo	Female	Adult	0.70	-36.44681	148.424506	0.149	0.046	0.068
E30	Thredbo	Male	Adult	1.20	-36.410407	148.495869	0.137	0.120	0.065
E31	Thredbo	Female	Adult	0.91	-36.410407	148.495869	0.149	0.043	0.030
E35	Thredbo	Male	Adult	1.33	-36.44681	148.424506	0.142	0.091	0.043
E55	Thredbo	Female	Adult	0.63	-36.410407	148.495869	0.139	0.109	0.060
E56	Thredbo	Female	Adult	0.92	-36.410407	148.495869	0.137	0.120	0.048
E60	Thredbo	Female	Juvenile	0.50	-36.44681	148.424506	0.134	0.138	0.056
E61	Thredbo	Male	Adult	0.90	-36.44681	148.424506	0.144	0.077	0.038
E62	Thredbo	Male	Adult	1.60	-36.44681	148.424506	0.138	0.114	0.029
E66	Thredbo	Female	Juvenile	0.40	-36.467726	148.373502	0.135	0.134	0.067
E75	Thredbo	Male	Juvenile	0.66	-36.369731	148.589586	0.143	0.081	0.027
E76	Thredbo	Female	Adult	0.95	-36.369731	148.589586	0.142	0.089	0.023
E77	Thredbo	Female	Juvenile	0.44	-36.376912	148.583204	0.141	0.095	0.048
E78	Thredbo	Male	Juvenile	0.76	-36.376912	148.583204	0.140	0.102	0.035
E90	Thredbo	Female	Juvenile	NA	-36.438257	148.443454	0.144	0.073	0.021
E91	Thredbo	Male	Adult	1.50	-36.497628	148.316844	0.147	0.059	0.038
E92	Thredbo	Male	Adult	0.91	-36.497628	148.316844	0.147	0.059	0.073

5 Thesis discussion

In the PDF version of this thesis, each reference to a Figure, Table, Equation or Section contains a hyperlink that, when clicked, will take the reader to that element. For hyperlinks, typing **Alt+left arrow** (or ⌘+left in a Mac) and **Alt+right arrow** (or ⌘+right in a Mac) act like the “go back” and “go forward” buttons in a web browser.

5.1 Introduction

The amount of genetic differentiation between populations reflects the action of evolutionary forces, such as gene flow, genetic drift and natural selection. This thesis investigated a range of questions relating to genetic differentiation, from theoretical to practical and from ecological to evolutionary processes. To answer these questions, I analysed four datasets, each with different characteristics and attributes: experimental populations of vinegar flies (*Drosophila melanogaster*; Holleley, 2009), a feral population of cattle (*Bos taurus*; Williams *et al.*, 2016), computer-simulated populations (Chapter 3), and wild populations of the platypus (*Ornithorhynchus anatinus*).

In **Chapter 2**, I investigated associative overdominance (AOD), a relatively unrecognised genetic mechanism that appears to be a major driver of genetic variation in small and isolated populations, which is paradoxically triggered by deleterious alleles. Accumulating evidence (Latter, 1998; Schou *et al.*, 2017; Becher *et al.*, 2020; Gilbert, Pouyet, *et al.*, 2020) suggests that AOD might be the cause of higher genetic diversity than expected in small and isolated populations, and its possible causes have been modelled mathematically and by simulations (Latter, 1998; Zhao & Charlesworth, 2016; Becher *et al.*, 2020; Gilbert, Pouyet, *et al.*, 2020). However, it has been challenging to demonstrate the occurrence of AOD directly because we cannot unambiguously determine the coefficients of selection and dominance on particular segregating alleles. This uncertainty was expressed by Zhao and Charlesworth (2016): “*Current theory and data cannot convincingly answer the question of whether AOD due to deleterious mutations is a credible explanation for the presence of more than expected levels of*

variability in small populations". In this chapter, I examined the fly experiment and the Chillingham datasets to look for signals consistent with the effects of AOD.

In **Chapter 3**, I developed a computer simulation program that allowed me to explore linked selection scenarios that cannot be solved analytically by theoretical models, including the joint interaction of selection, recombination, genetic drift and gene flow in multiple loci. I explored how linked selection, such as AOD and background selection (BGS), affects genetic variation within and between populations, based on three simulation models: general simulations, fly simulations and Chillingham simulations.

In **Chapter 4**, I investigated how dams affect gene flow and dispersal between platypus groups. I used various approaches to compare the genetic differentiation between platypus groups separated by dams and the differentiation between platypus groups in unregulated rivers.

In this final chapter, I first present the answers to my main research questions. Following this, I discuss the main findings of this thesis and their significance and implications. I then point out the research gaps, future directions and challenges in the areas studied in this thesis and finish with concluding remarks.

5.2 Answers to research questions

What are the evolutionary, demographic and genomic conditions in which linked selection via deleterious alleles impacts genetic variation within and between populations?

Using my general simulation model, I explored 320 scenarios with different combinations of the main drivers of linked selection that I and others (Latter, 1998; Schou *et al.*, 2017; Becher *et al.*, 2020; Gilbert, Pouyet, *et al.*, 2020) identified: dominance (h), selection coefficient (s), effective population size (N_e) and the number of deleterious alleles per individual per centiMorgan (del/Ind/cM). These simulations (Extended Data Figure 2.3) showed that in most scenarios, the effects of associative overdominance (AOD) overcame the effects of background selection (BGS). I found that AOD depressed genetic differentiation between populations as measured by F_{ST} .

(Extended Data Figure 2.2) and decreased the rate of loss of genetic variation within populations as measured by heterozygosity (H_e ; Figure 2.1). AOD was more prevalent in those scenarios with lower dominance coefficients (h), higher selection coefficient (s), lower effective population size (N_e) and higher number of targets of selection as measured by the number of deleterious alleles per individual per centiMorgan (del/ind/cM).

Can computer simulations parameterised with realistic values explain patterns of genetic diversity observed in live populations?

My computer simulations designed to recreate the density of targets of selection and recombination rate in specific genomic regions in *D. melanogaster* and *B. taurus* replicated the patterns of genetic variation empirically observed in both the fly experiment and the Chillingham cattle. Specifically, fly simulations reproduced the variance and amount of H_e and F_{ST} observed in the loci genotyped in the fly experiment (Figure 2.2). Similarly, Chillingham simulations replicated the location of genomic regions of high polymorphism observed in the Chillingham cattle (Figure 2.4).

Can linked selection be detected in natural populations?

I hypothesised that AOD would be stronger in genomic regions with a high density of targets of selection (*i.e.*, deleterious alleles) and low recombination. To test this hypothesis, I needed first to determine the distance at which each locus is influenced by surrounding loci under selection and recombination. To infer this distance, which I called the “genomic neighbourhood”, I developed a method based on sets of multiple regression analyses using windows of different physical (bp) and genomic (cM) distances (Supplementary Methods 2.9.3.2). By using this method, I showed in the fly experiment that H_e was higher (Extended Data Figure 2.11) and F_{ST} lower (Figure 2.3) in loci located in genomic regions of low recombination and high density of targets of selection. Furthermore, my method produced the same results in the Chillingham cattle (in 12 out of 28 chromosomes; Appendix 1: Investigating AOD in each Chillingham chromosome) and in simulations of flies (Extended Data Figure 2.13 and Extended Data Figure 2.14) and Chillingham cattle (Extended Data Figure 2.15).

To further investigate my AOD hypotheses, I developed a method to identify and visualise haplotypes based on the linkage disequilibrium (LD) statistic r^2 (Hill & Robertson, 1968; Supplementary Methods 2.9.3.2). I applied this method to the Chillingham cattle dataset, which had sufficiently high-resolution data to characterise patterns of LD. Results showed that haplotypes with high polymorphism were located mostly in genomic regions with high or low recombination (Figure 2.5a). Haplotypes with high polymorphism located in regions of low recombination might be the result of the effects of AOD. One of the predictions of linked selection proposes that polymorphism in high recombination regions can escape from the effects of BGS (Cutter & Payseur, 2013). This prediction is congruent with the observation that haplotypes located in regions of high recombination have higher polymorphism. I also found that haplotypes with low polymorphism were located in areas with a high density of targets of selection (Figure 2.5 b). This low polymorphism could be due to the effects of BGS (Ellegren & Galtier, 2016). These findings collectively agree with recent research suggesting that AOD and BGS occur concurrently (Zhao & Charlesworth, 2016; Becher *et al.*, 2020).

Do dams affect dispersal and gene flow between platypus populations?

I found that F_{ST} was significantly higher between platypus groups separated by dams than groups in unregulated rivers (Table 4.4). I further found a significant relationship between F_{ST} and the number of platypus generations since the building of dams, in which F_{ST} increased at a rate of 0.012 per generation (Figure 4.2). Further evidence suggesting that dams impede gene flow between platypus came from maps generated to visualise areas of high genetic differentiation (Figure 4.3 to Figure 4.5). Maps of regulated rivers showed that areas of high genetic differentiation coincide with the location of dams, in contrast to maps of unregulated rivers where areas of high genetic differentiation were relatively absent.

5.3 Discussion of results

The concept of genetic differentiation was the predominant element present throughout the thesis, where I illustrated its potential to unravel the effects of evolutionary forces on genetic variation. For example, estimation of genetic differentiation was crucial to detect AOD and understand its mechanisms. Genetic differentiation was also my primary inference tool to investigate the impact of dams on platypus gene flow and dispersal.

In **Chapters 2 and 3**, I found strong evidence indicating that the occurrence, prevalence and intensity of AOD might be more widespread than previously appreciated. These conclusions were obtained from experimental fly populations that were highly replicated (36 population pairs in total). In addition, fly populations were maintained for an extended period in the laboratory (~60 generations) at large population sizes (~2,000 individuals) to avoid possible confounding effects of selective sweeps. These are essential aspects to consider to adequately assess the occurrence of AOD, as suggested by Zhao and Charlesworth (2016).

A significant contribution from my research is the development of a novel method to identify the likely distance at which each locus is influenced by surrounding deleterious alleles and recombination, which I called the “genomic neighbourhood”. This method produced consistent results across the fly experiment, the Chillingham cattle, and computer simulations. A further original contribution of my work is that I focused on the density of deleterious alleles per individual per centimorgan (del/ind/cM) as one of my main parameters to model the effects of AOD on genetic variation. The del/ind/cM parameter explicitly incorporates the main parameters driving AOD (*i.e.*, recombination and density of deleterious alleles), making it a suitable metric to link AOD models with real-life scenarios.

As discussed before, I found that AOD can occur across a wide range of parameter values and scenarios. These results raise the question of why the effects of AOD have not been more widely observed throughout nature. The explanation might be due to various reasons. Firstly, heterozygosity for a single locus in small populations has a very high

stochastic variance (Zhao & Charlesworth, 2016), which might obscure the signals of AOD. A further reason explaining the lack of a larger number of observations congruent with AOD could be that genomic resources required to detect the effects of linked selection, such as recombination maps and reference genomes, are not yet widely available in most species. Also, it might be possible that signals of linked selection could sometimes be obscured by the opposing effects of AOD and BGS, which could cancel each other out. This interpretation agrees with studies showing that AOD and BGS can occur simultaneously in chromosomal regions of low recombination (Zhao & Charlesworth, 2016; Becher *et al.*, 2020).

In **Chapter 2**, I found a positive relationship between the strength of AOD and the density of targets of selection in a given genomic region. These results support the findings of previous studies suggesting that inbreeding depression might be caused by several deleterious mutations of small effect rather than caused by few deleterious mutations with large effect (Charlesworth & Willis, 2009).

In **Chapter 4**, I showed, by using various approaches, that dams probably represent impassable barriers for platypuses. My results also corroborate critical genetic aspects of the platypus, such as high inbreeding estimates, low genetic variation, low N_e estimates and lower genetic variation in northern than in southern populations (Martin *et al.*, 2018). I identified that the platypus group with probably the smallest population size (below the dam in the Mitta-Mitta River) had the highest genetic diversity among all the groups sampled. The low sample size of this group ($n = 4$) might be the most plausible explanation for this observation. However, we cannot discard the possibility that this high genetic diversity might be due to AOD because the demographics of this population are consistent with the conditions required for AOD to occur (*i.e.*, small and isolated populations). Indeed, further investigation is necessary to determine whether AOD can explain the high genetic diversity observed in this river, preferably using a larger sample size than the current one.

5.4 Research gaps, future directions and challenges

In **Chapters 2 and 3**, I explored the role that deleterious alleles play in developing AOD at relatively large distances (*i.e.*, millions of base pairs). This is equivalent to assuming that the deleterious alleles involved in the emergence of AOD are located at different genes. However, it is possible that AOD could also arise from deleterious alleles located at shorter physical distances (*i.e.*, within genes). Further investigation of this possibility could determine whether AOD might be a potential explanation for the high genetic variability in fitness observed in *Drosophila* that cannot be explained by current population genetics theory (Charlesworth, 2015). Another critical topic deserving further work is investigating whether adaptive genetic variation maintained by AOD can, in fact, confer to populations a higher evolutionary potential (Frankham *et al.*, 2017) and whether this potential is retained after an increase in population size. This topic could be explored, for example, by using my simulation model.

My framework to analyse and detect AOD could potentially have broader applications in fields beyond population genetics and conservation. For example, my framework could be used to identify haplotypes with deleterious alleles that are the underlying cause of common diseases with a genetic component, such as cardiovascular diseases, preeclampsia, asthma, schizophrenia and cancers. For example, this framework could be used to identify chromosomal regions under the effects of AOD (as shown with the Chillingham cattle) in datasets of small and isolated populations in which diseases, as the mentioned above, have been diagnosed. This approach could be used to prioritise candidate genes for further study. Potential datasets with these characteristics are held by the National Centre for Indigenous Genomics and The John Curtin School of Medical Research at the Australian National University and the Centre for Genetic Origins of Health and Disease at the University of Western Australia.

My research will help to gain a better understanding of AOD, which might derive into various potential applications, such as:

- Determining whether estimates of genetic variation, upon which management decisions are based, are affected by linked selection, and interpret them accordingly.

- Determining whether management actions are required despite a small population size.
- Establishing the breeding conditions that originate AOD.

An essential next step in my research is to extend and improve my simulation models and make them more accessible to potential users. This goal could be achieved by incorporating my models into an already established software, such as *dartR* (Gruber *et al.*, 2018).

I envision that the platypus dataset used in **Chapter 4**, obtained from sampling several sites within each river, has enough resolution to investigate more complex dispersal patterns and at different temporal scales using other genetic resources, such as relatedness analyses, spatial autocorrelation, and coalescence. I also anticipate that a critical avenue of research would be using a simulation approach to explore the predictive ability of the relationship between the number of generations and the length of time since dams were built. Further investigation of this relationship could potentially allow us to use F_{ST} to predict in which situations interruption of gene flow is severe enough to trigger active management actions, such as animal translocations (Seddon *et al.*, 2007) or construction of ecological corridors (Paetkau *et al.*, 2009). This investigation would need first to establish a relationship between F_{ST} , Ne , levels of gene flow, loss of adaptive genetic variation and levels of inbreeding. Determining these relationships in realistic scenarios would allow us to establish F_{ST} thresholds that could be included in management policies of natural populations. Additionally, F_{ST} could also be implemented as an essential measure to monitor conservation and restoration efforts.

5.5 Concluding remarks

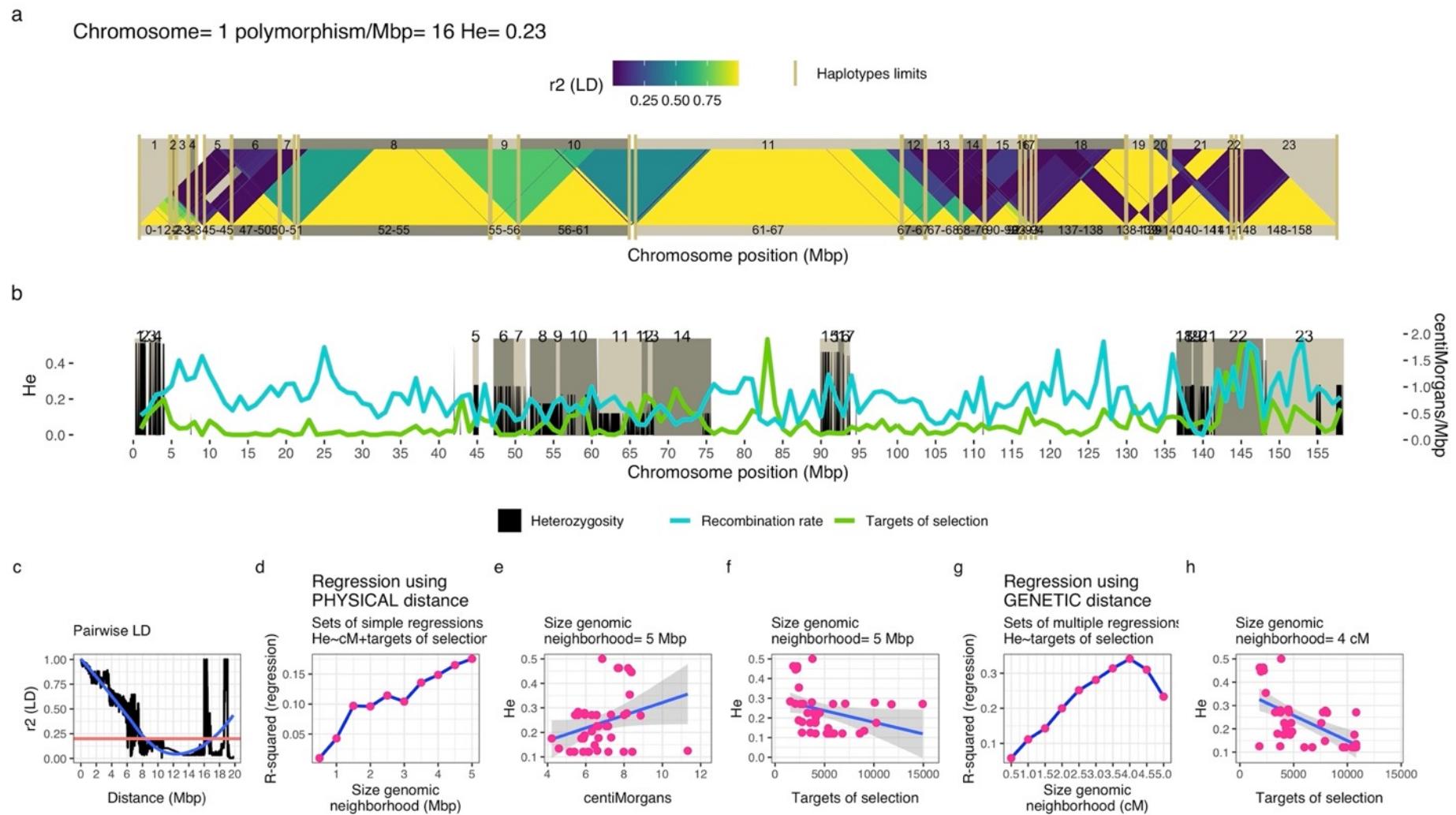
Collectively my findings constitute one of the most comprehensive studies about the prevalence, intensity and occurrence of AOD, an underappreciated driver of genetic variation in small and isolated populations. My results add to the list of recent research indicating that linked selection plays a vital role in determining genetic variation at the species and genome levels (Ellegren & Galtier, 2016; Charlesworth & Charlesworth, 2018). The prevalence of linked selection, particularly AOD, has significant implications for various applied and theoretical fields in genetics. This is especially significant for many approaches and methods that assume unlinked and selectively neutral loci, many of which are used to inform the restoration and conservation of biodiversity. I highlighted that AOD should be considered when doing translocations to avoid problems that could arise from breaking haplotypes maintained by AOD.

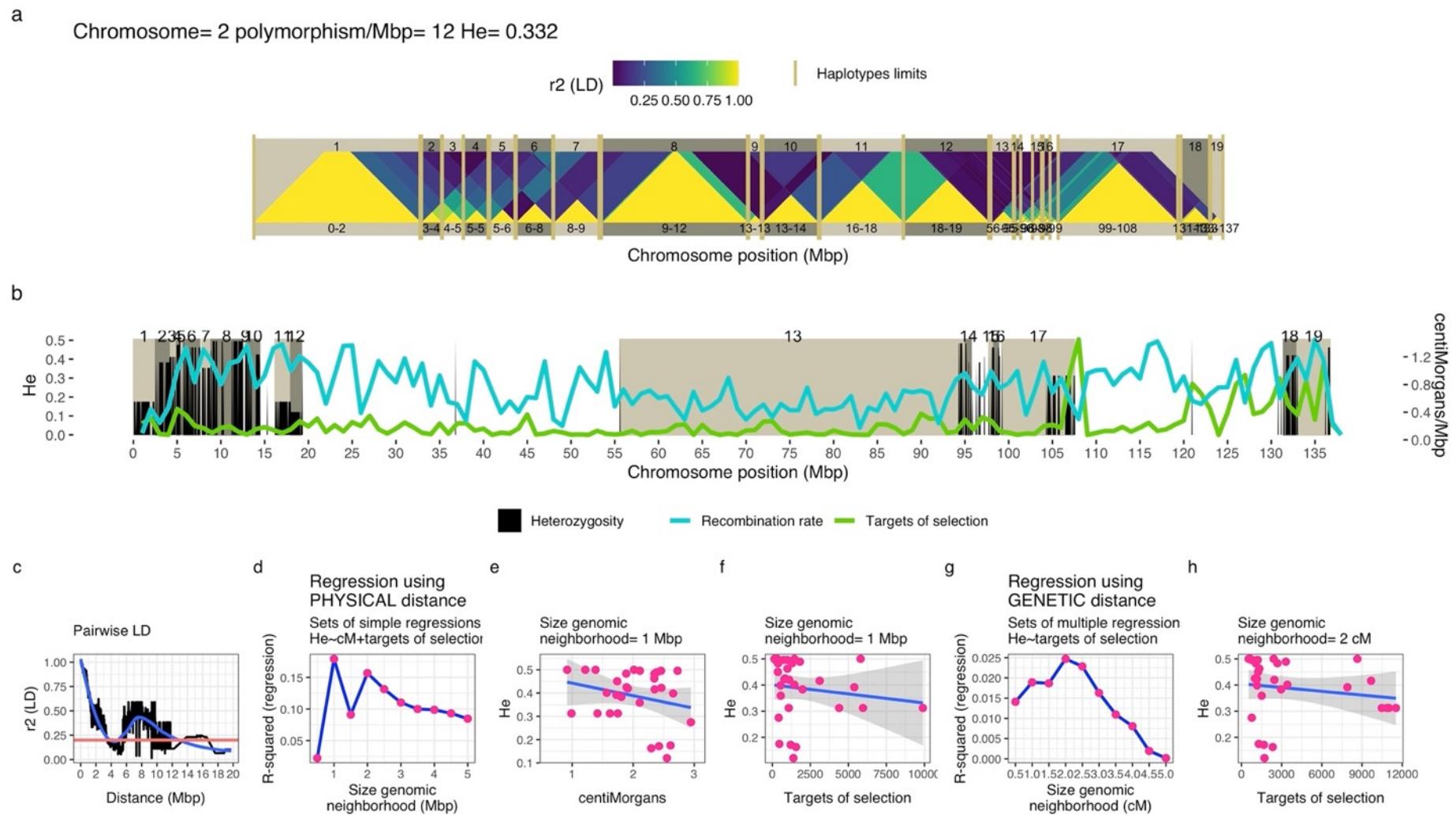
6 Appendix 1: Investigating AOD in each Chillingham chromosome

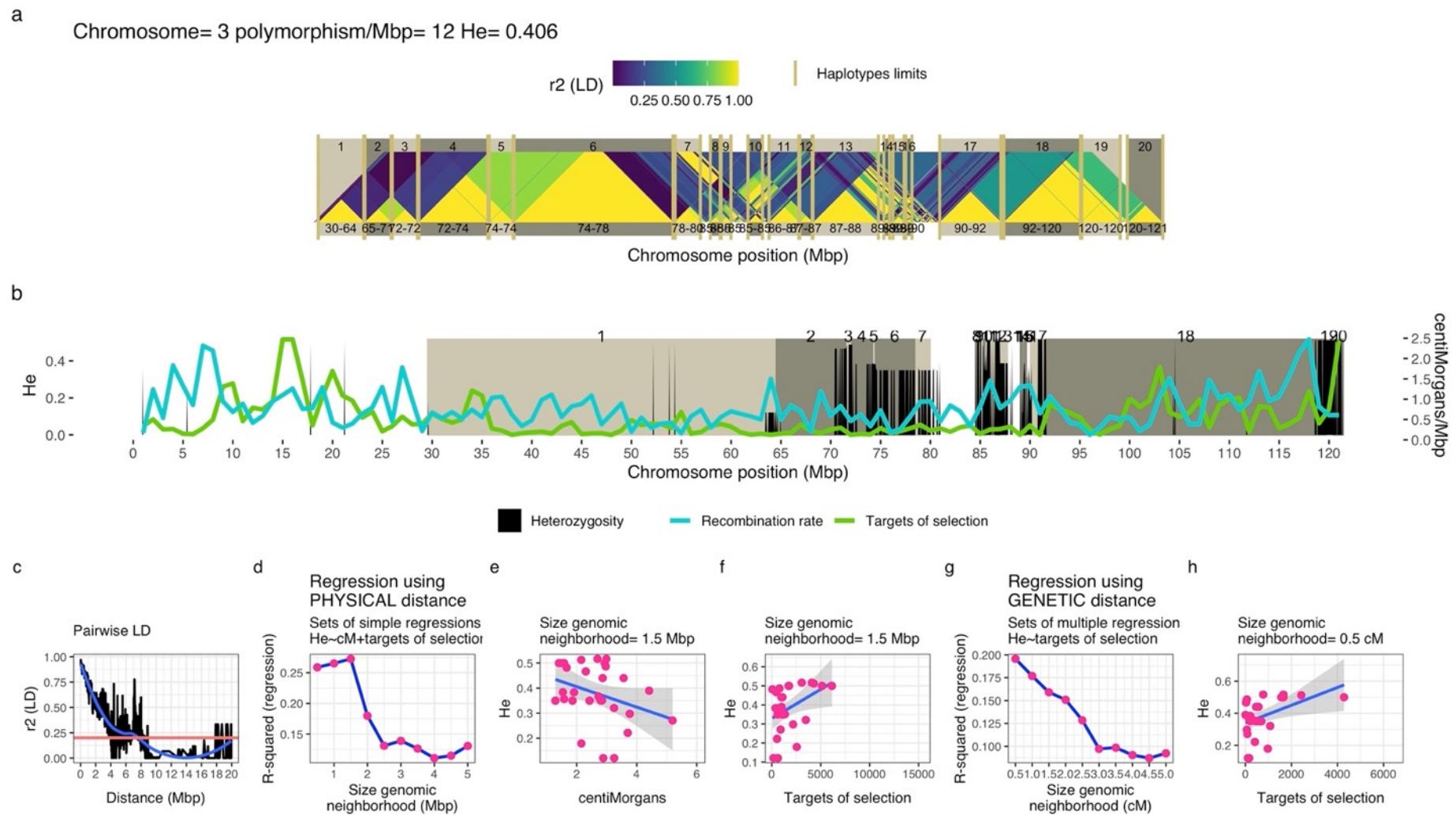
In this document, I present the results of my two methods to investigate AOD in each chromosome of the Chillingham cattle: the multiple regression method to infer the size of the genomic neighbourhood (*i.e.*, the likely distance at which each locus is influenced by surrounding deleterious alleles and recombination; Supplementary Methods 2.9.3.2), and the method to identify haplotypes (Supplementary Methods 2.9.3.2). 28 chromosomes were analysed; Chromosome X was not included because the recombination map for this chromosome was not available (Ma *et al.*, 2015); Chromosome 28 was not included because it had a level of polymorphism that was too low to identify haplotypes. In these analyses, I found patterns consistent with AOD, *i.e.*, a negative association between H_e and recombination and a positive association between H_e and my proxy for targets of selection (see Section 2.3.4), in 43% of the chromosomes (12 of 28 chromosomes). I considered a haplotype to be those contiguous SNPs in complete linkage disequilibrium (*i.e.*, with an r^2 statistic (Hill & Robertson, 1968) = 1) and containing more than ten polymorphic loci.

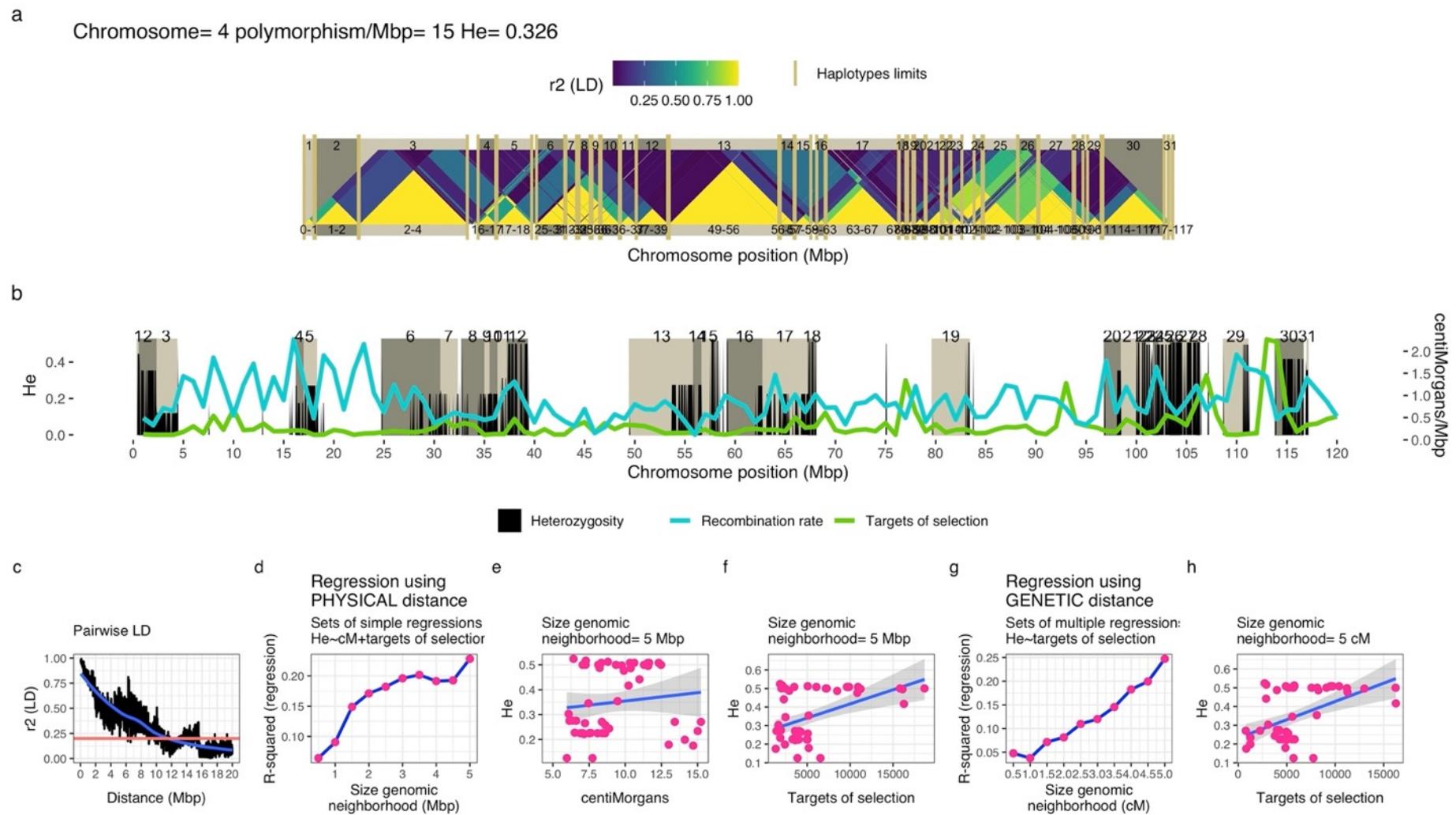
The following caption applies to all the figures in this Appendix. The first title of each figure includes the analysed chromosome, the number of polymorphic sites per Mbp, and the mean H_e across all polymorphic sites in the chromosome.

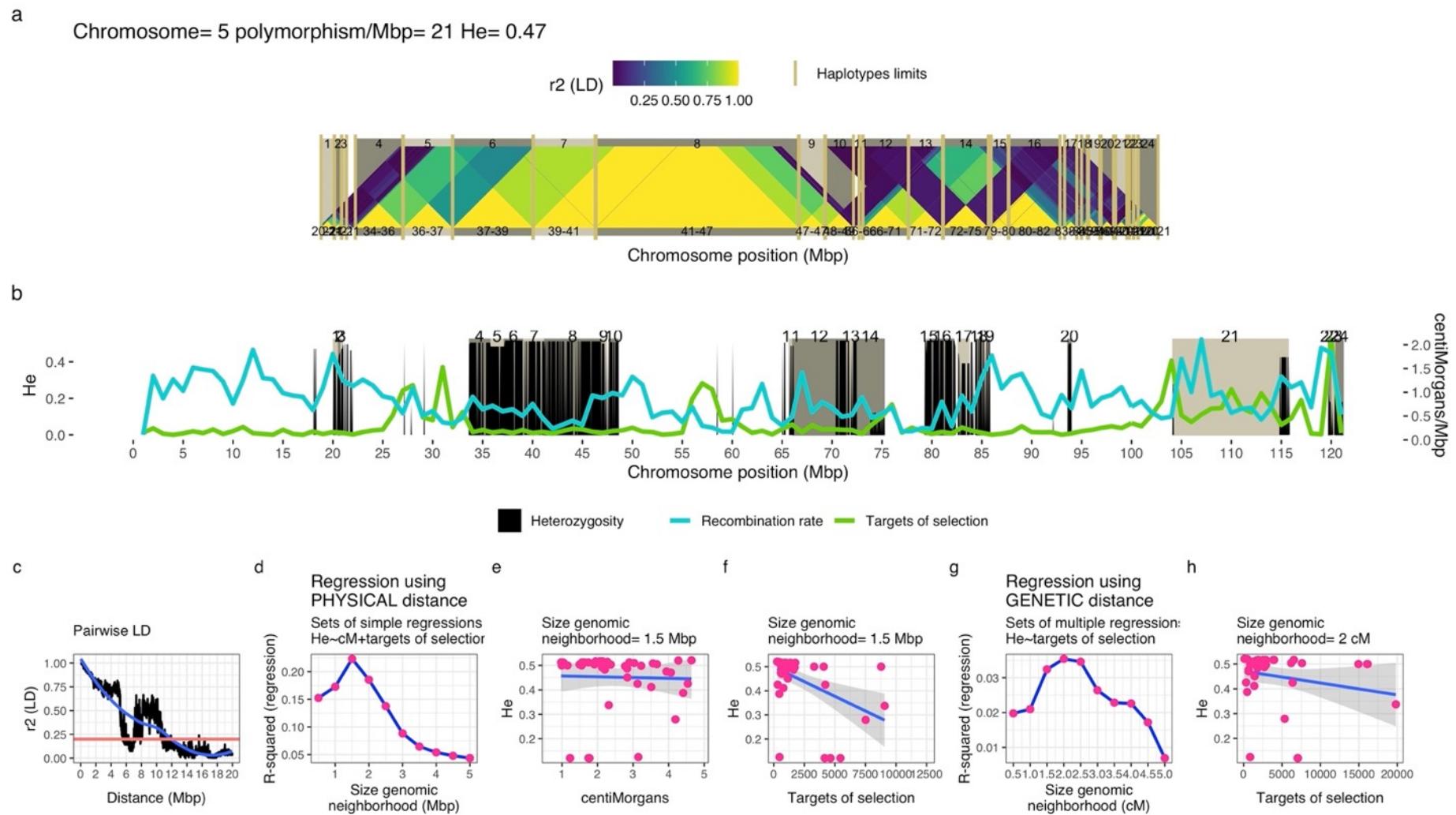
a) Linkage disequilibrium (LD) heatmap, where the colour shading indicates the strength of LD, using the statistic r^2 (Hill & Robertson, 1968). Chromosome positions (Mbp) are shown on the horizontal axis, and haplotypes appear as bright yellow triangles and delimited by dark yellow vertical lines. Numbers identifying each haplotype are shown in the upper part of panels of **a** and **b**. **b)** Visualisation of recombination rate (cM/Mbp; blue line), proxy for targets of selection (number of non-synonymous mutations; green line; scale is the same as in Figure 2.4) and polymorphic loci (black columns) across chromosomes. **c)** Decay of the pairwise LD as a function of physical distance (bp), using the statistic r^2 . Horizontal red line indicates the LD threshold ($r^2 = 0.2$), which is commonly used to imply that two loci are unlinked (Delourme *et al.*, 2013; Li *et al.*, 2014). **d)** Inference of the size of the genomic neighbourhood using physical distance, where each point is the R^2 of each multiple regression (H_e on recombination [cM] and proxy for targets of selection). The distance at which the regression signal reaches its peak of significance is considered the size of the genomic neighborhood. Note that this distance generally coincides with the physical distance at which pairwise LD (r^2) decayed below 0.2. **e)** Regression line between H_e and the recombination rate within the inferred genomic neighborhood. **f)** Regression line between H_e and proxy for targets of selection within the inferred genomic neighbourhood. **g)** Inference of the size of the genomic neighbourhood using genetic distance (cM), where each point is the R^2 of each simple regression (H_e on proxy for targets of selection). The distance at which the regression signal reaches its peak of significance is considered the size of the genomic neighborhood. **h)** Regression line between H_e and proxy for targets of selection within the inferred genomic neighbourhood. Shaded areas in **e**, **f**, and **h** are 95% confidence intervals of the regression line. Each point in **e**, **f**, and **h** is the H_e calculated in bins of 1 Mbp.

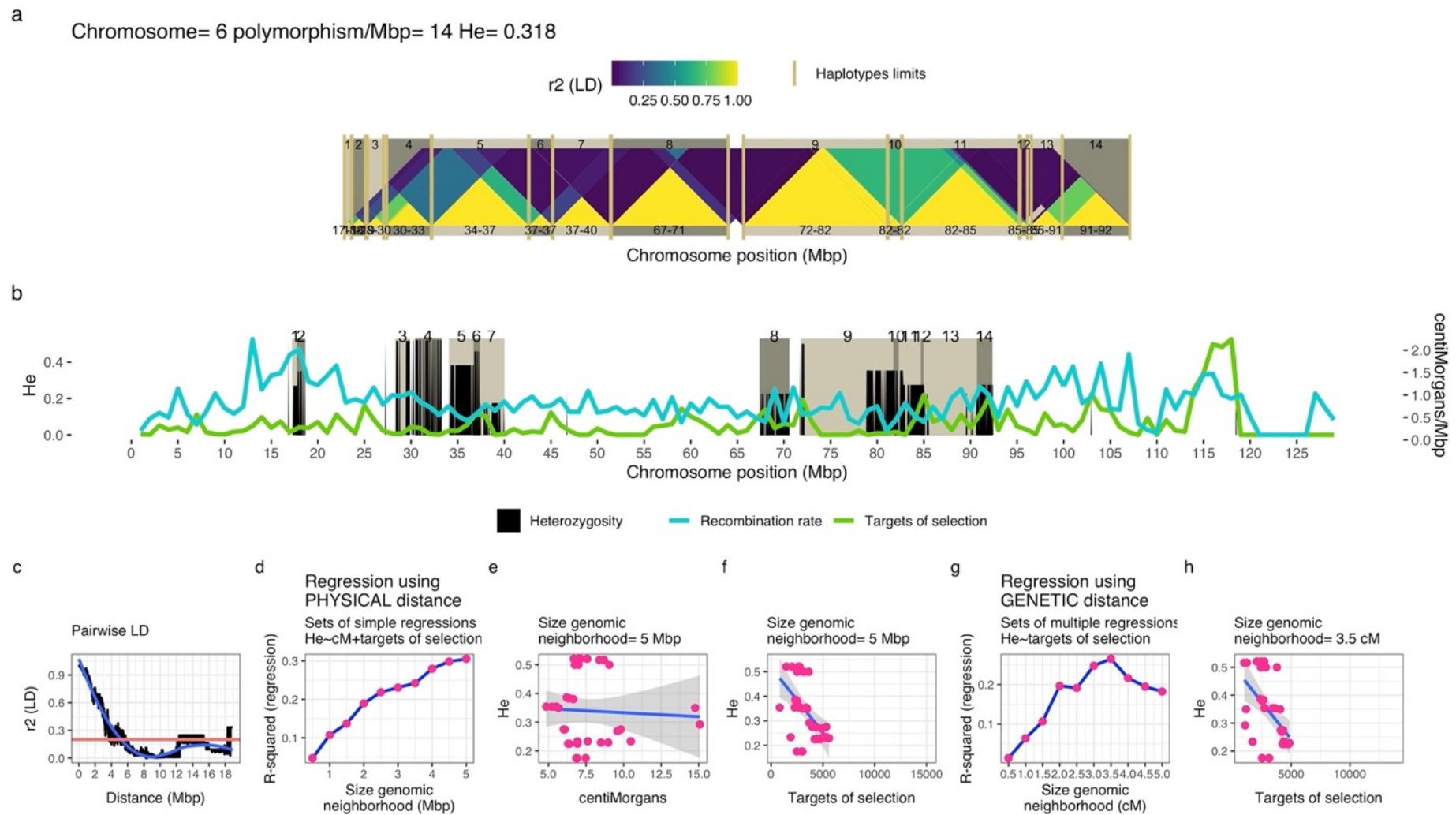


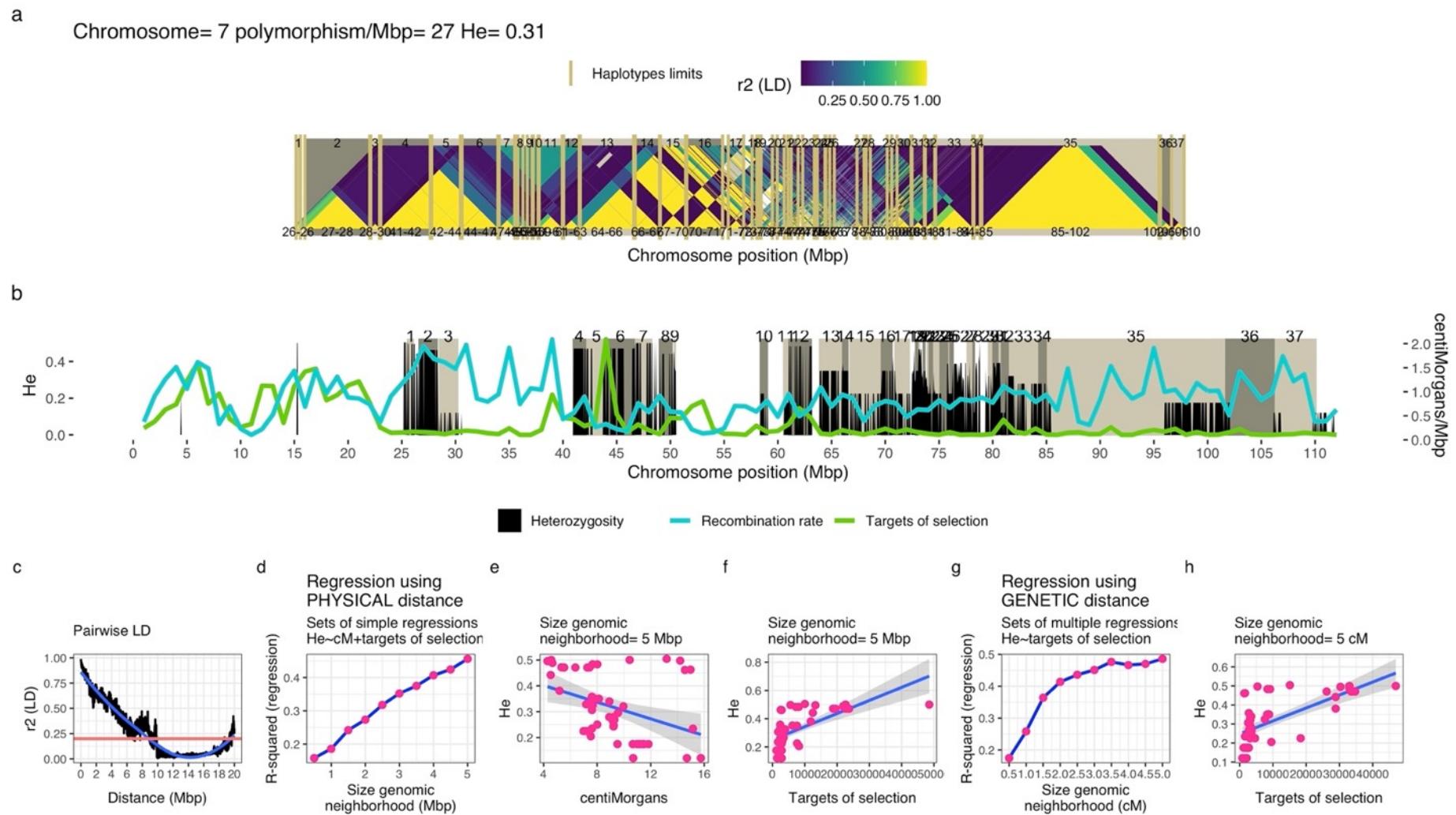


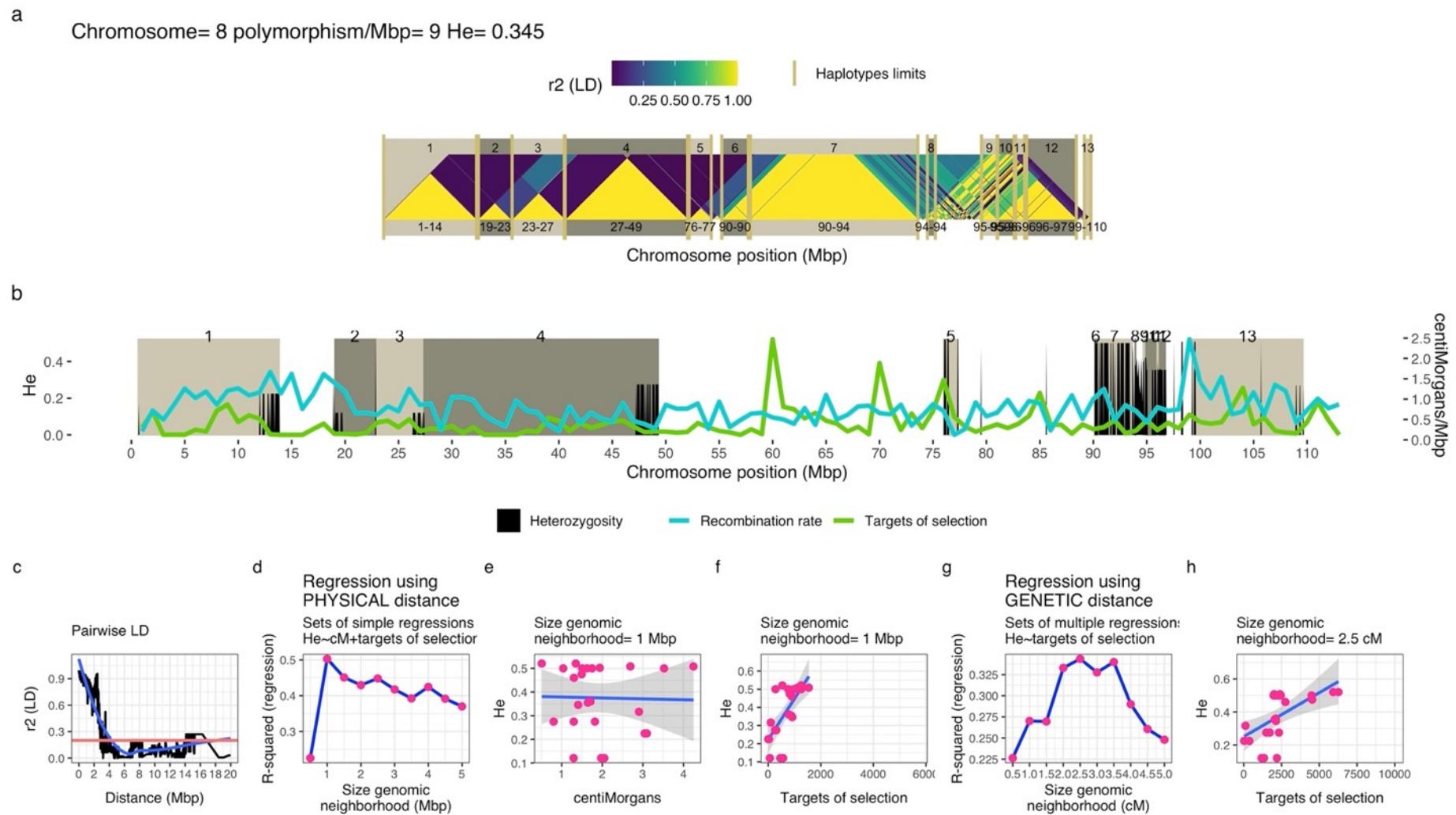


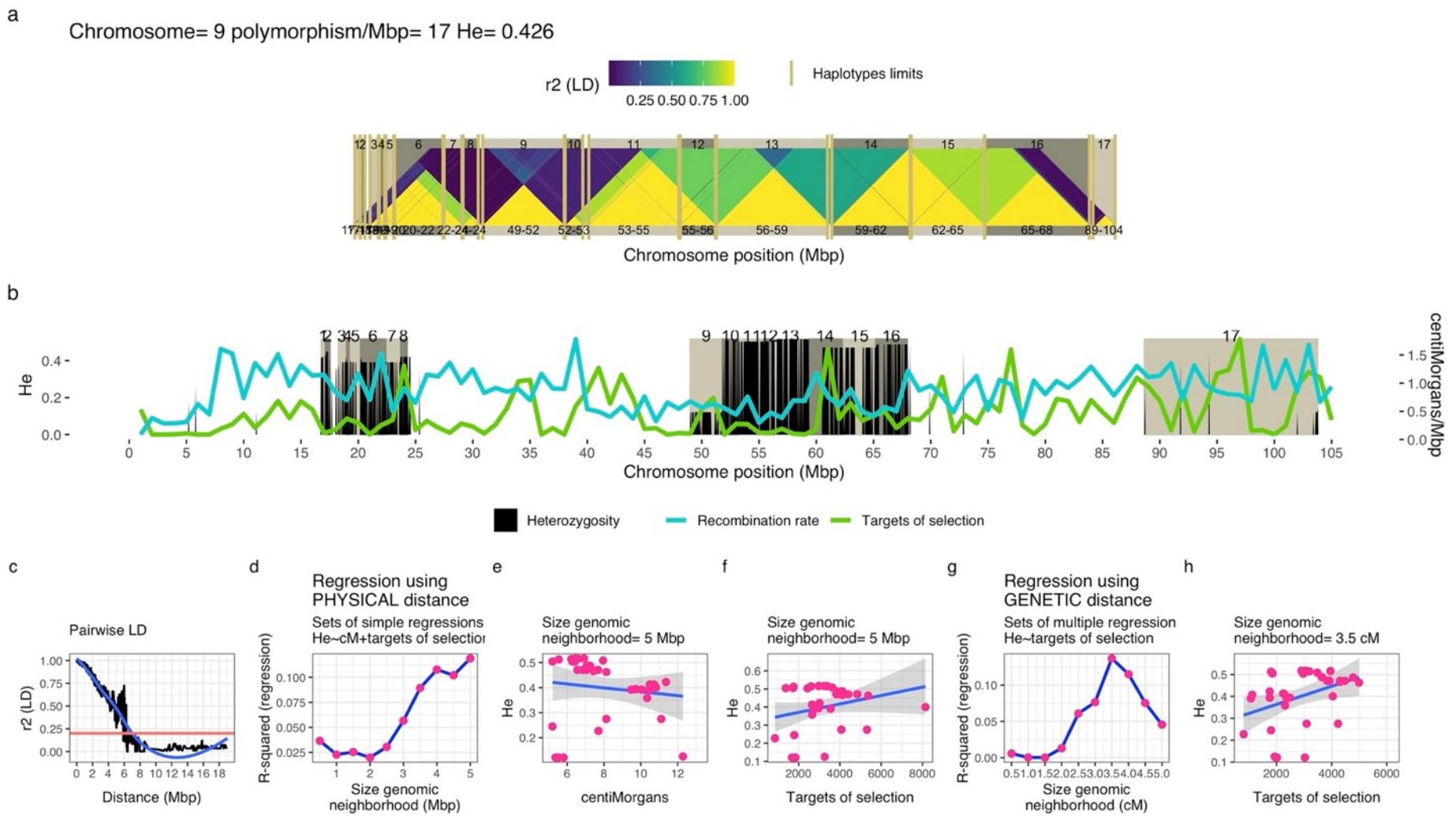


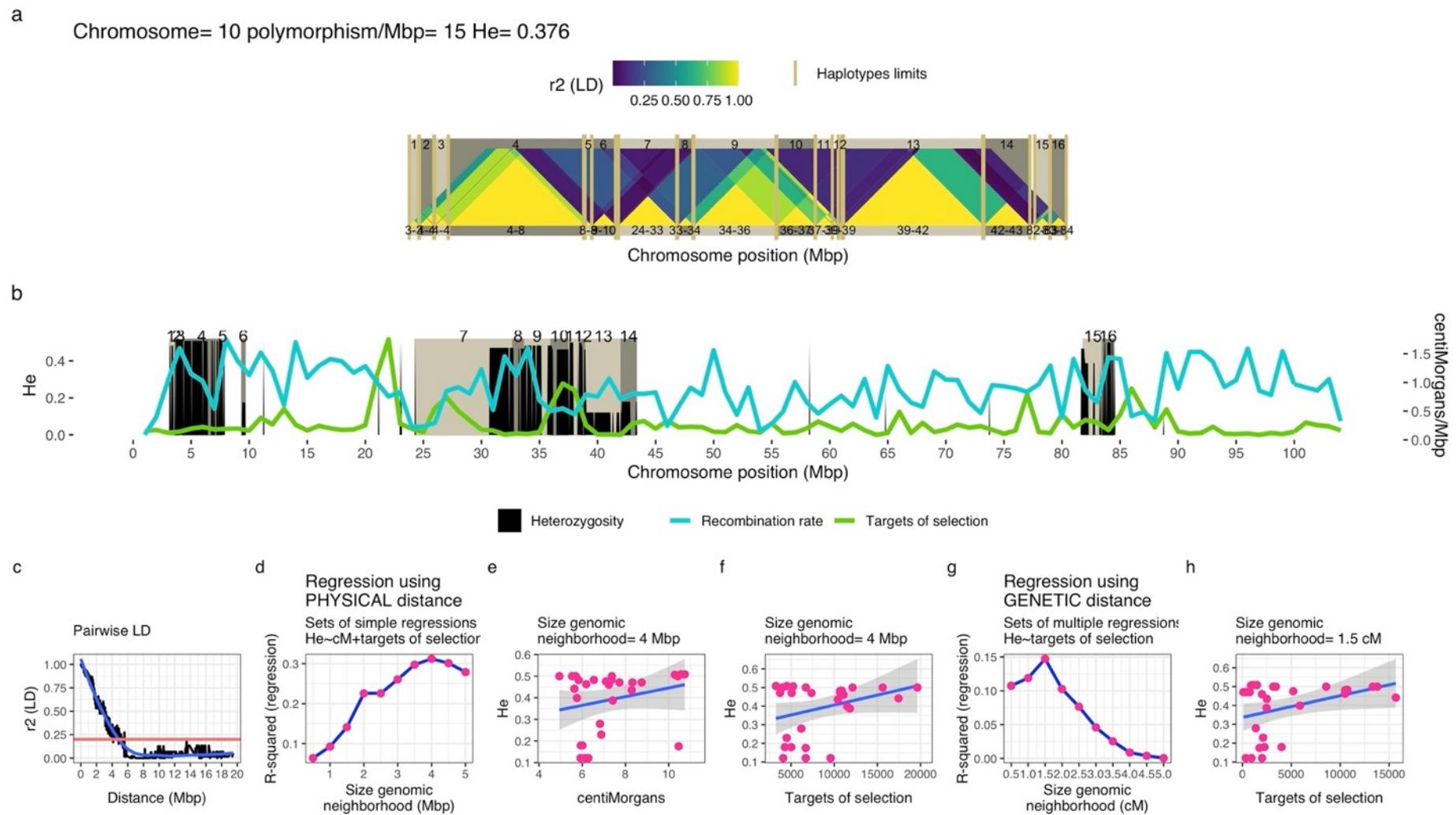


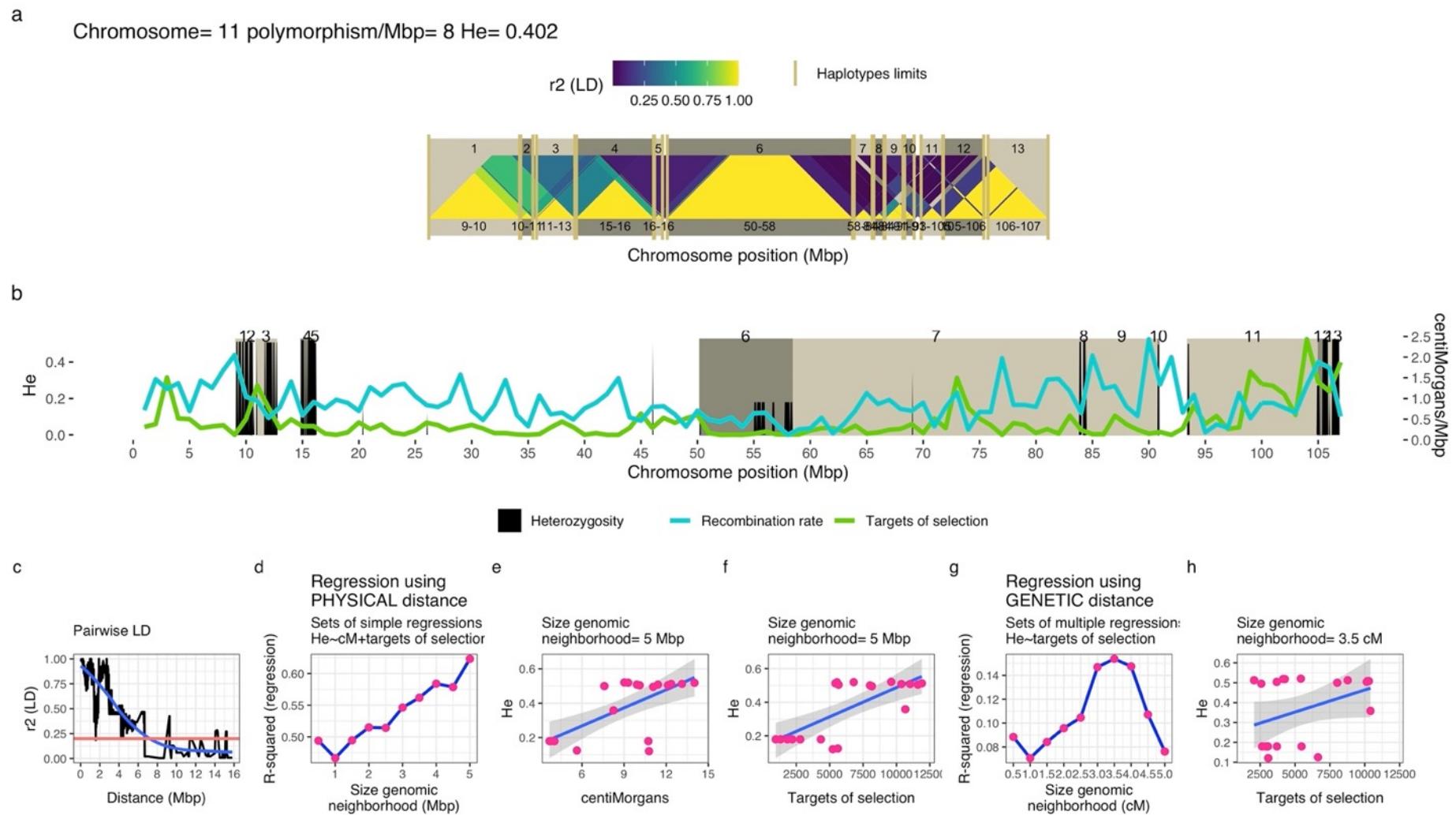


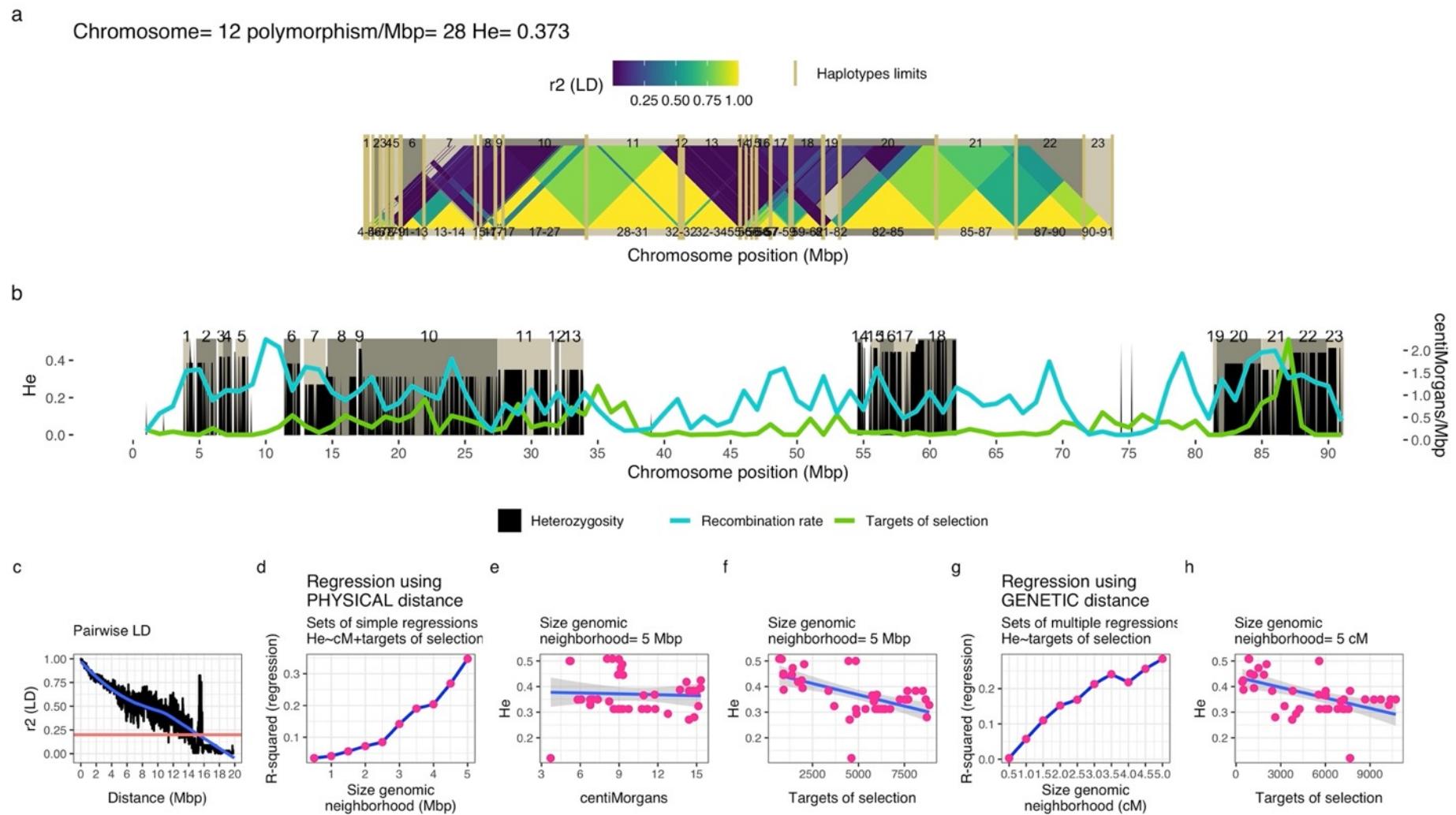


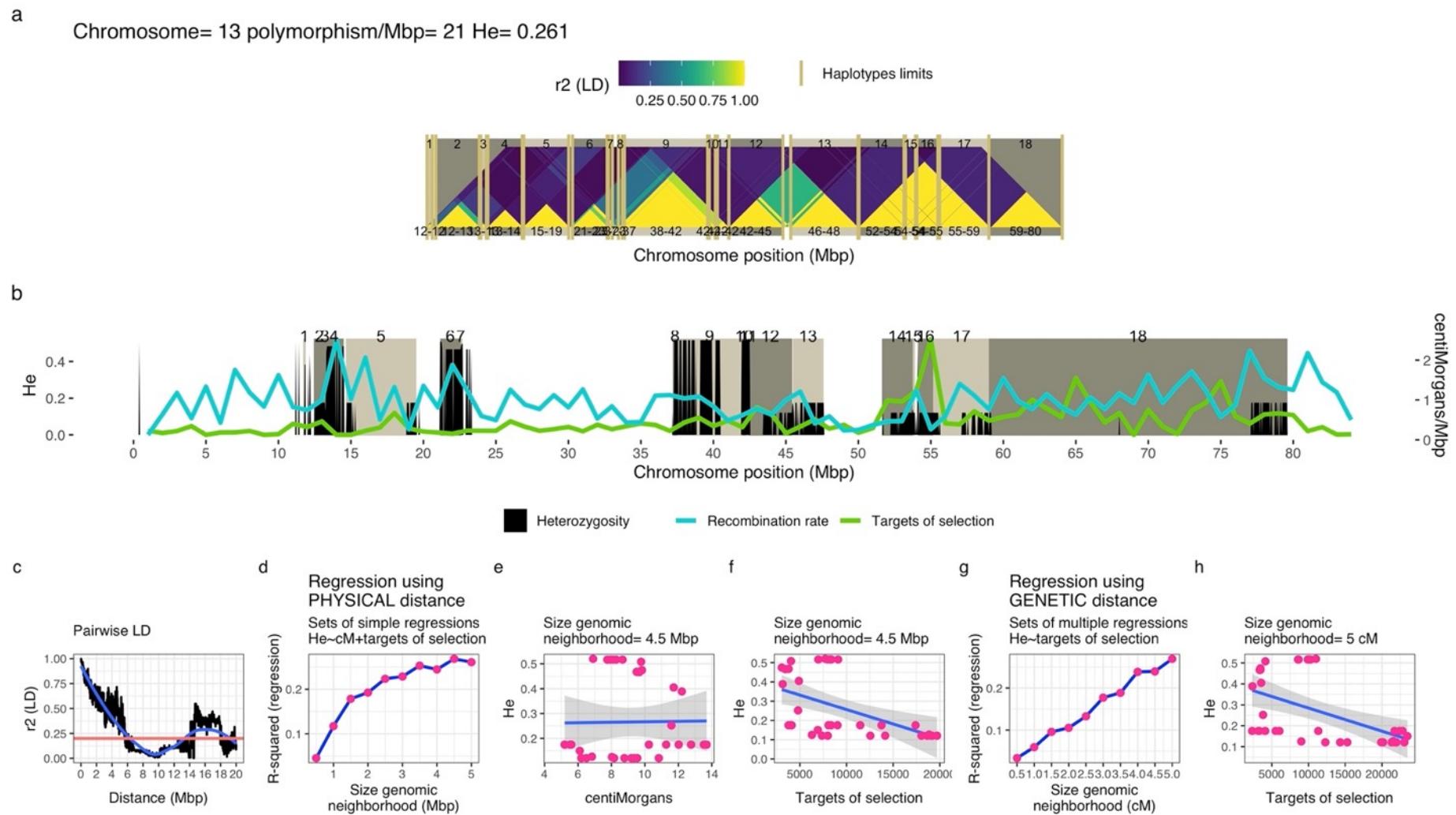


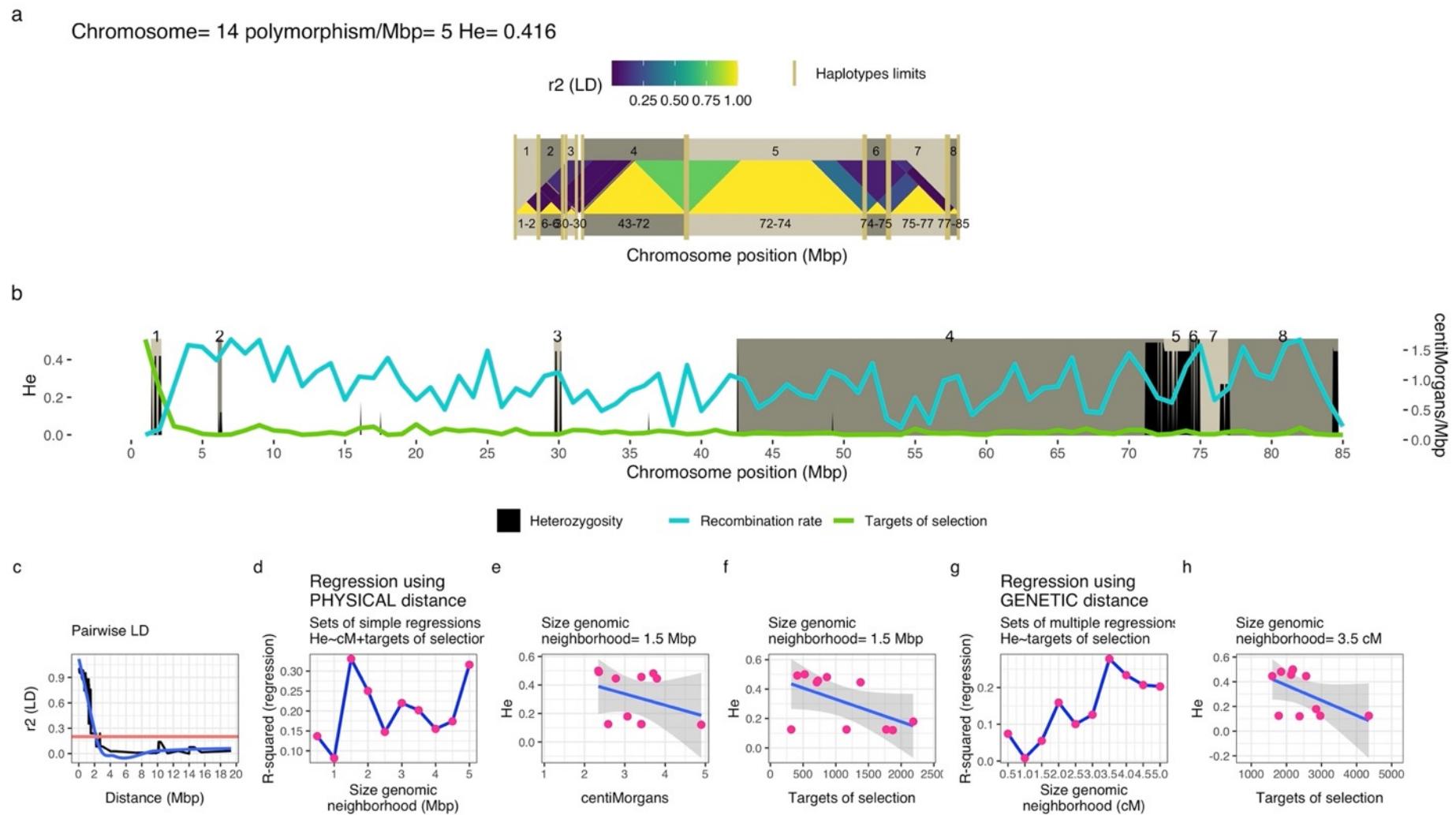


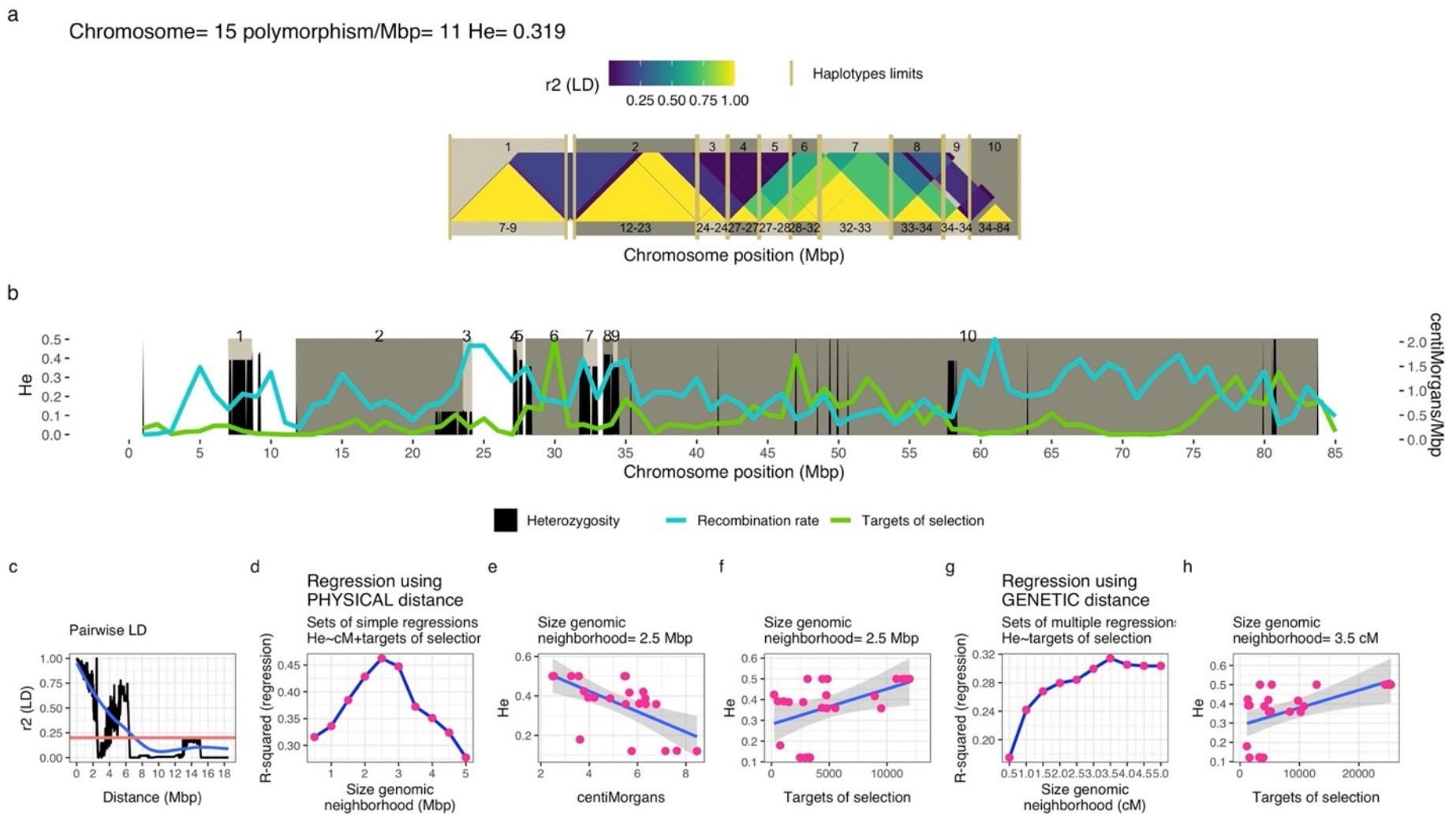


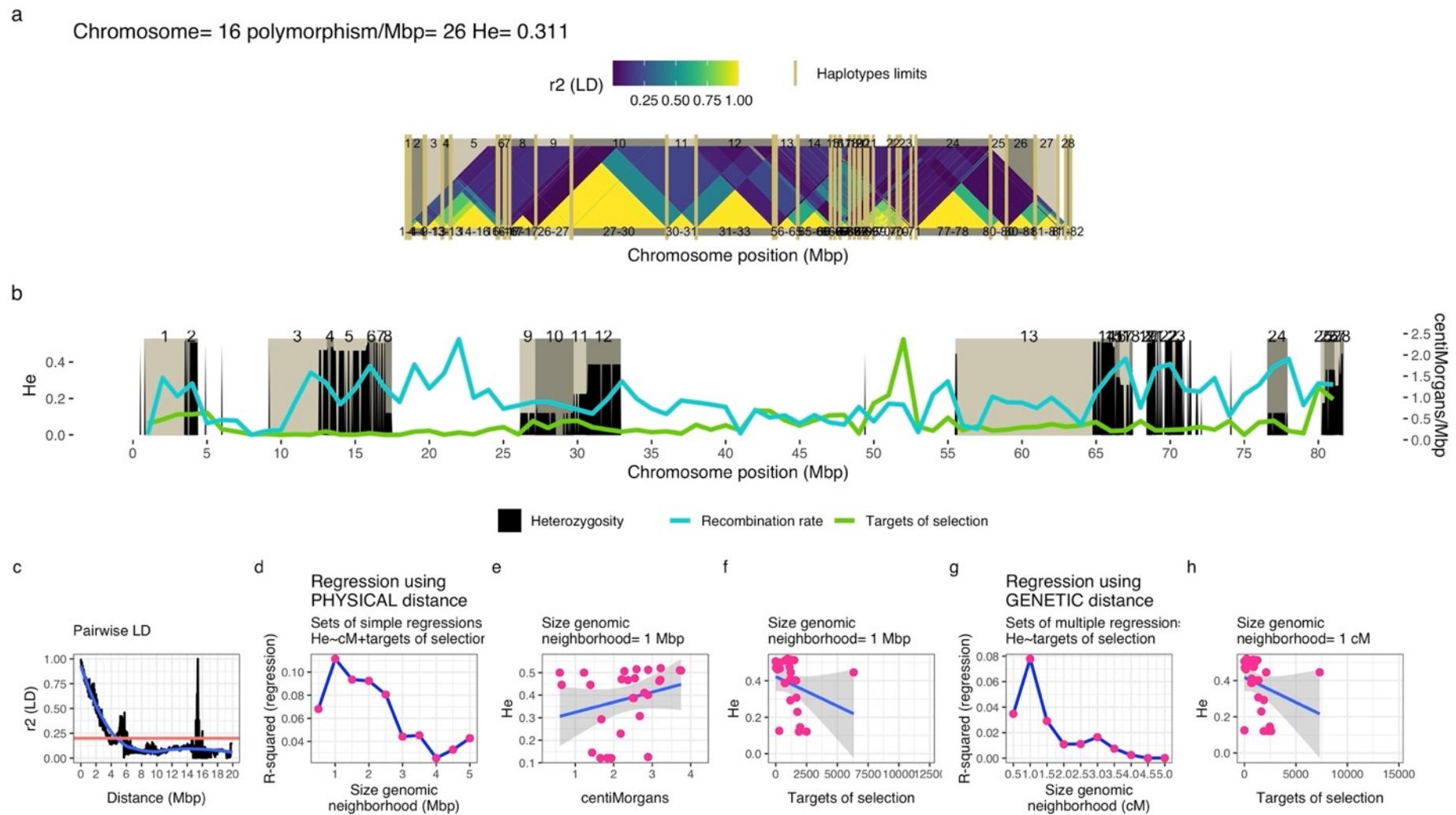


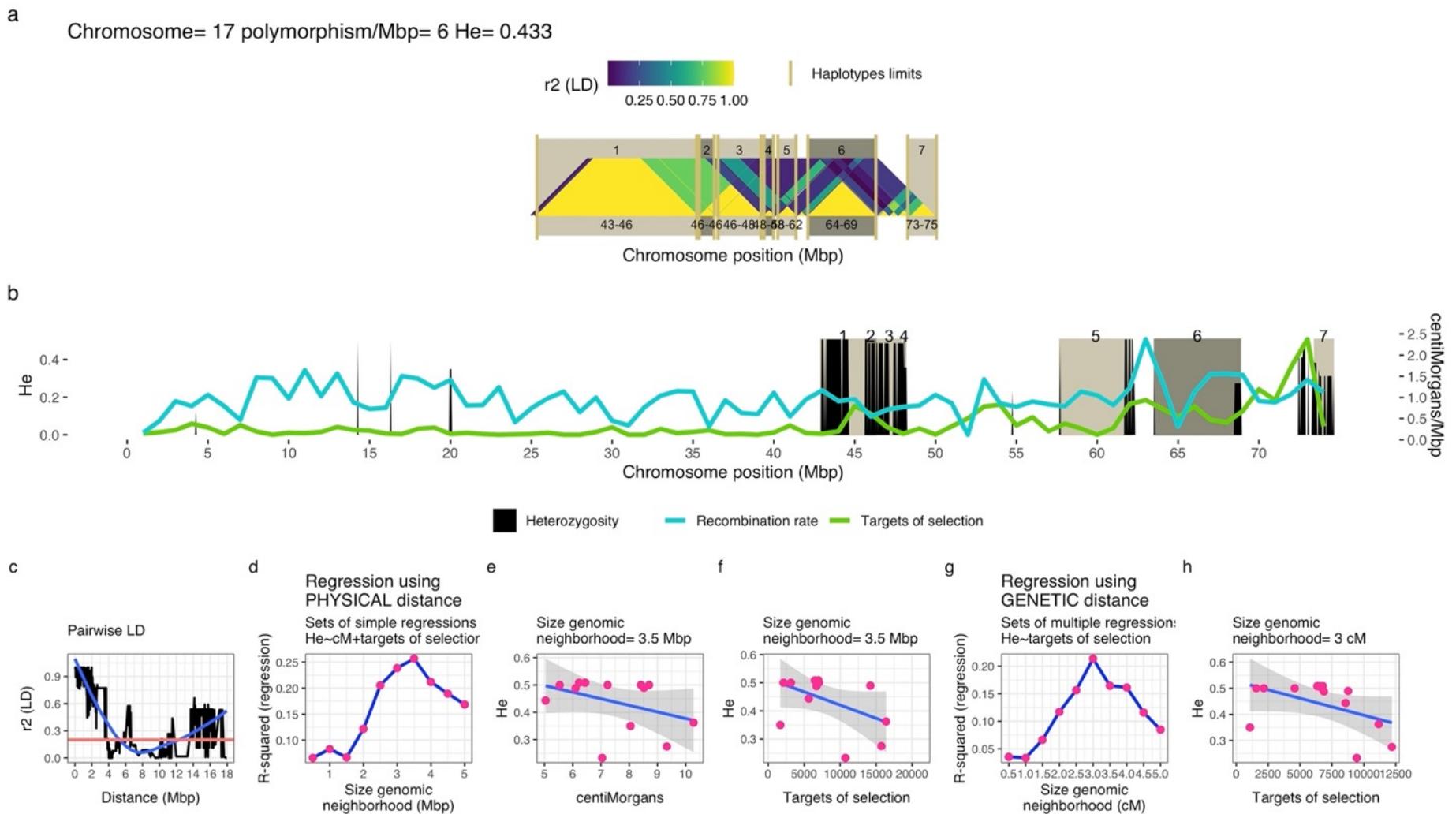


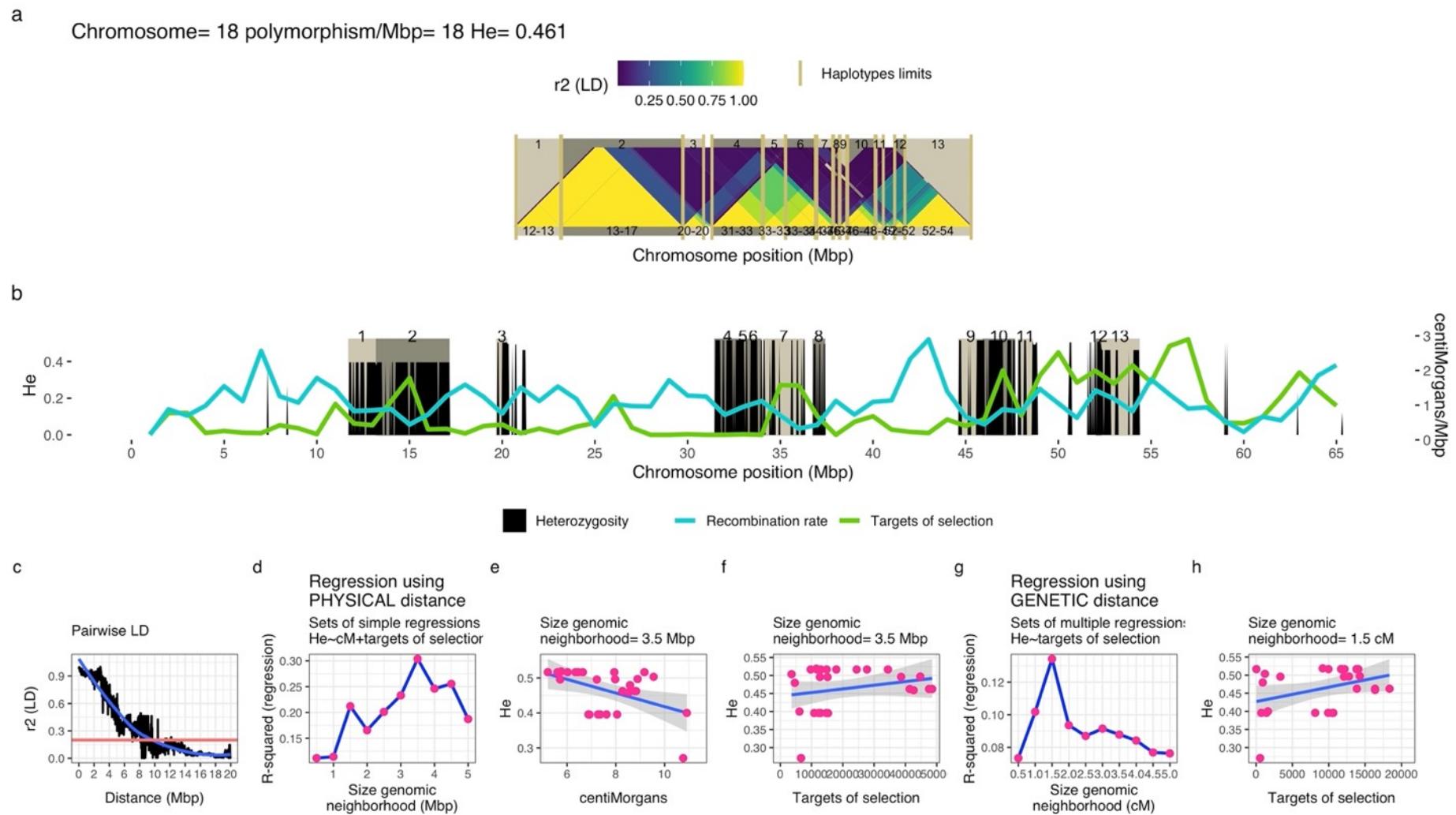


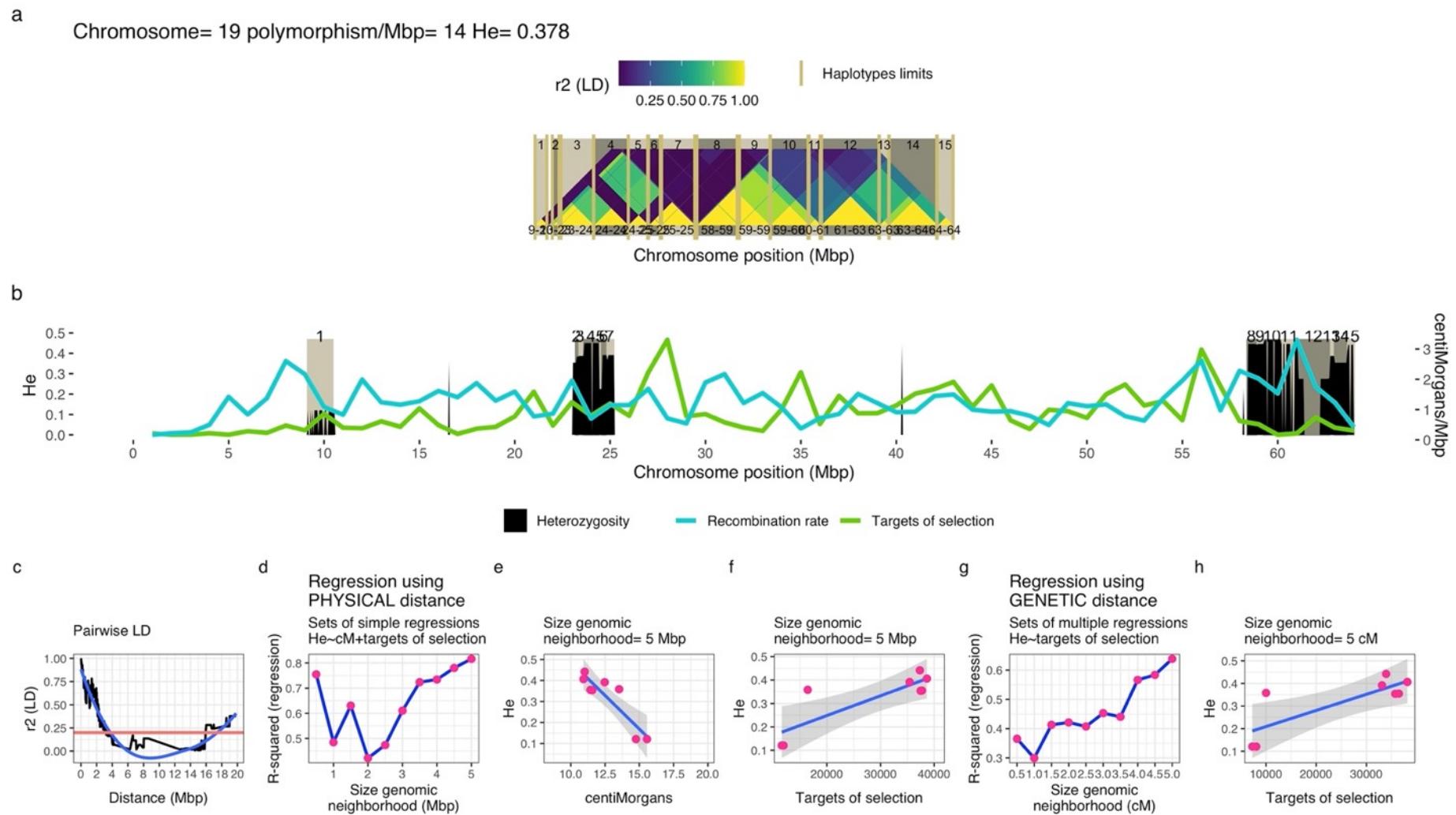


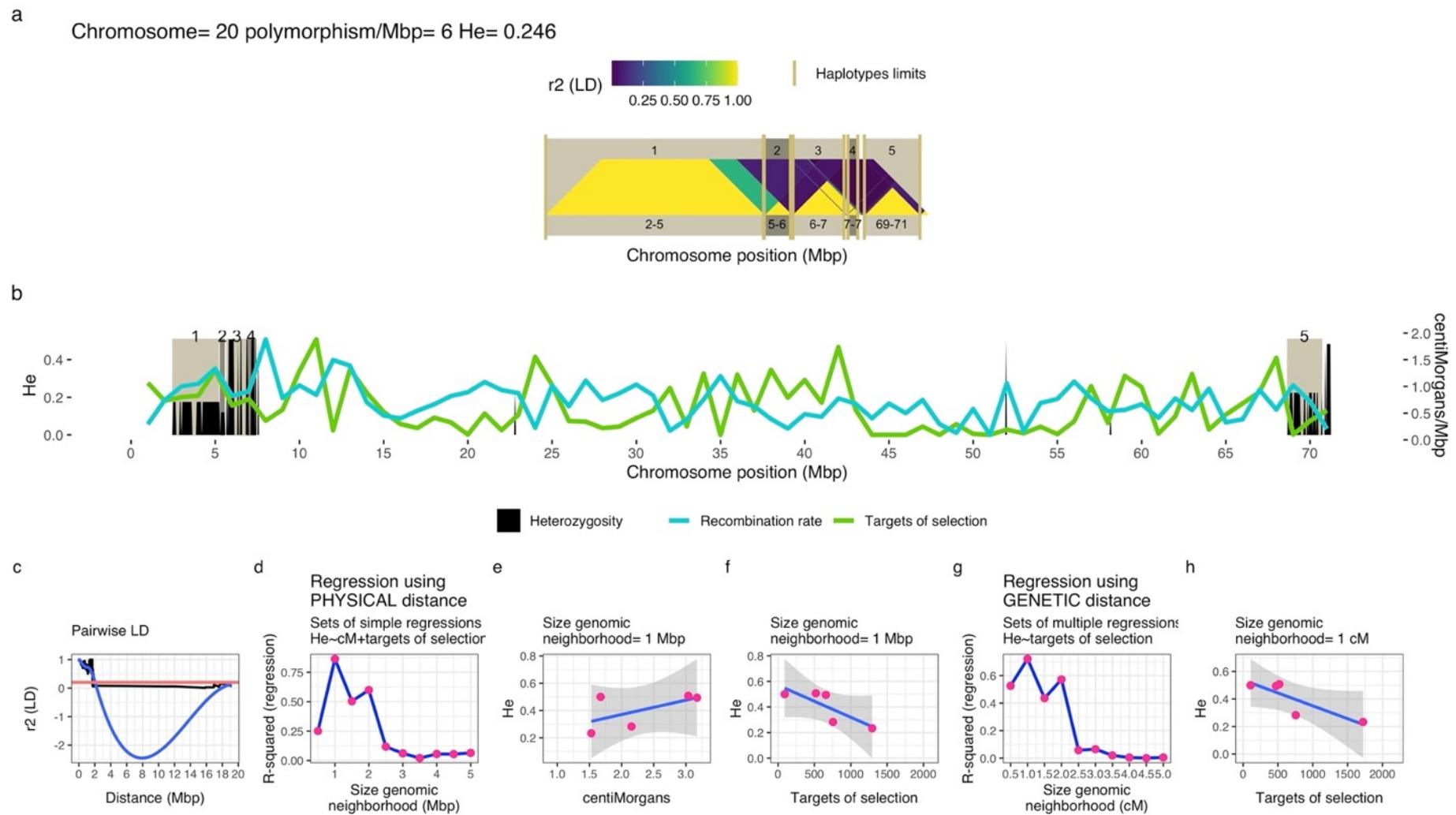


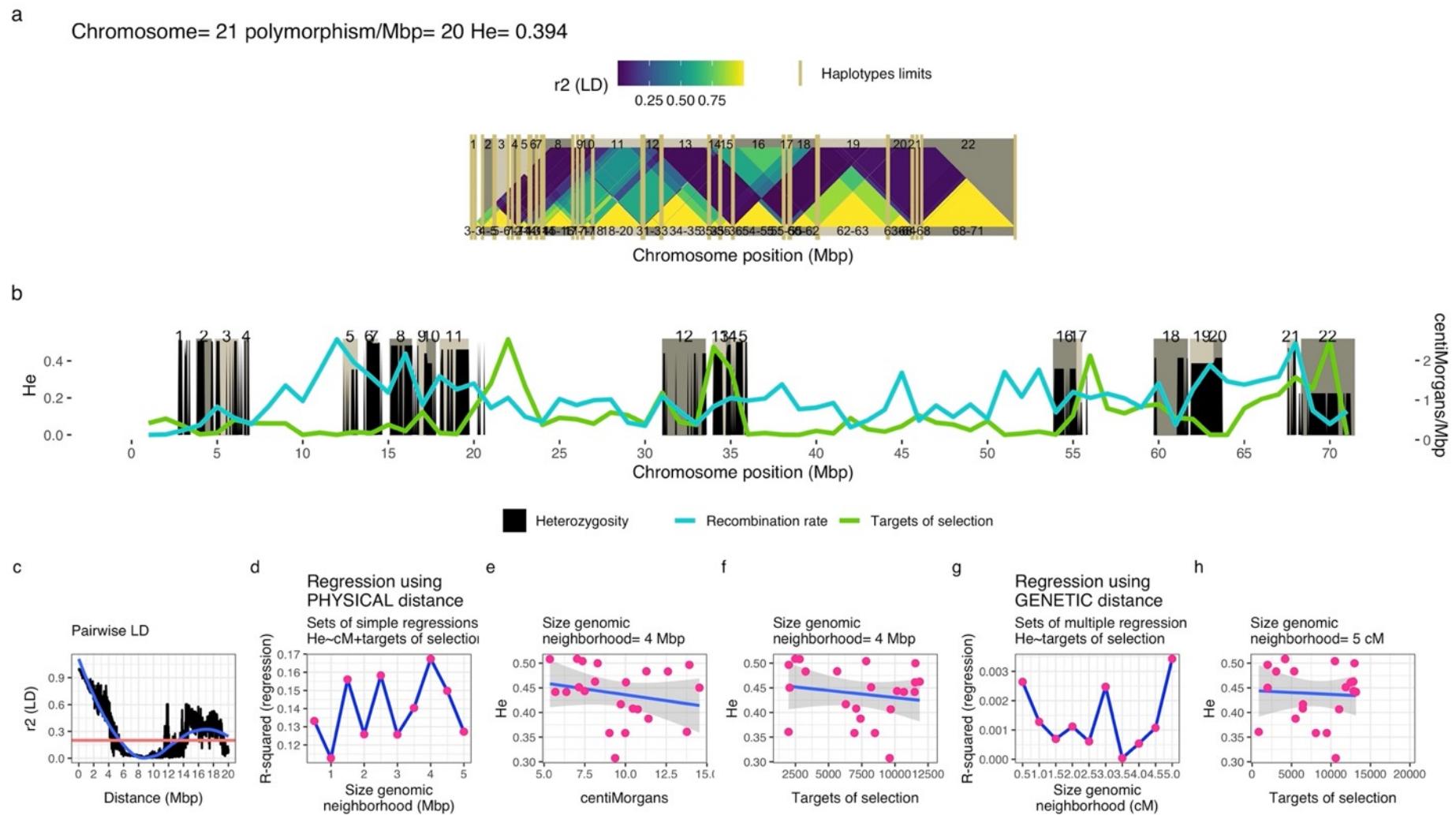


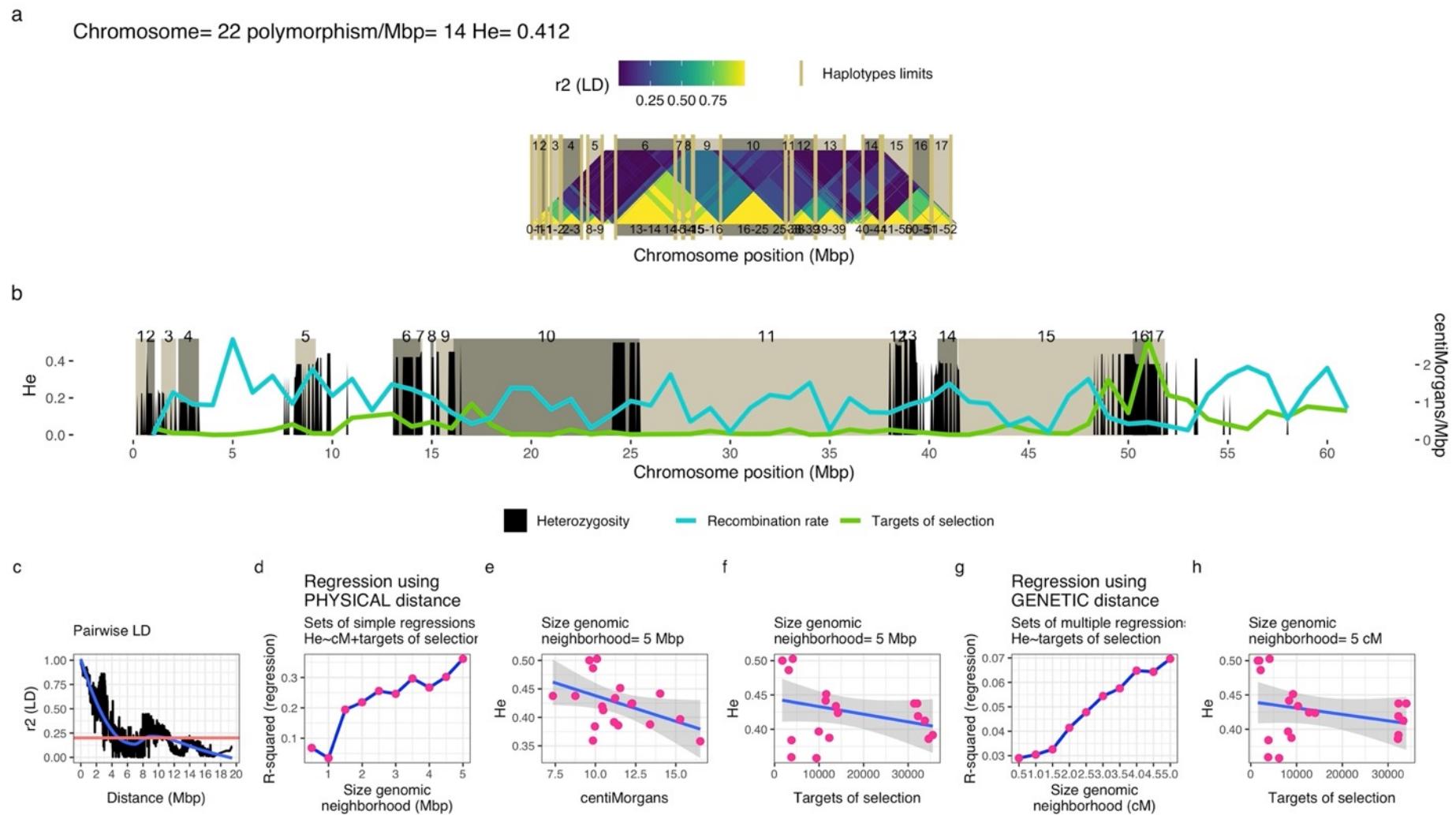


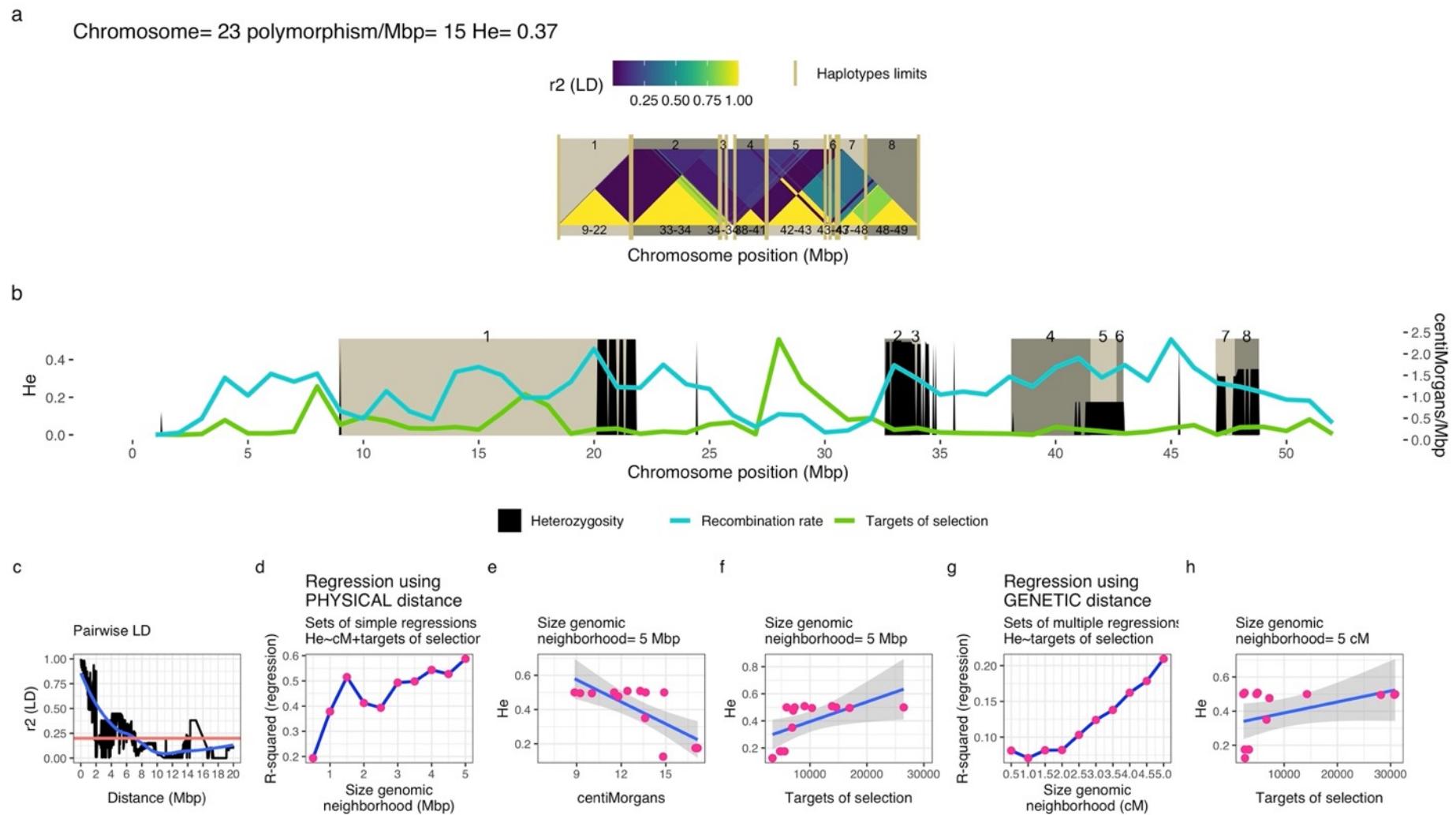


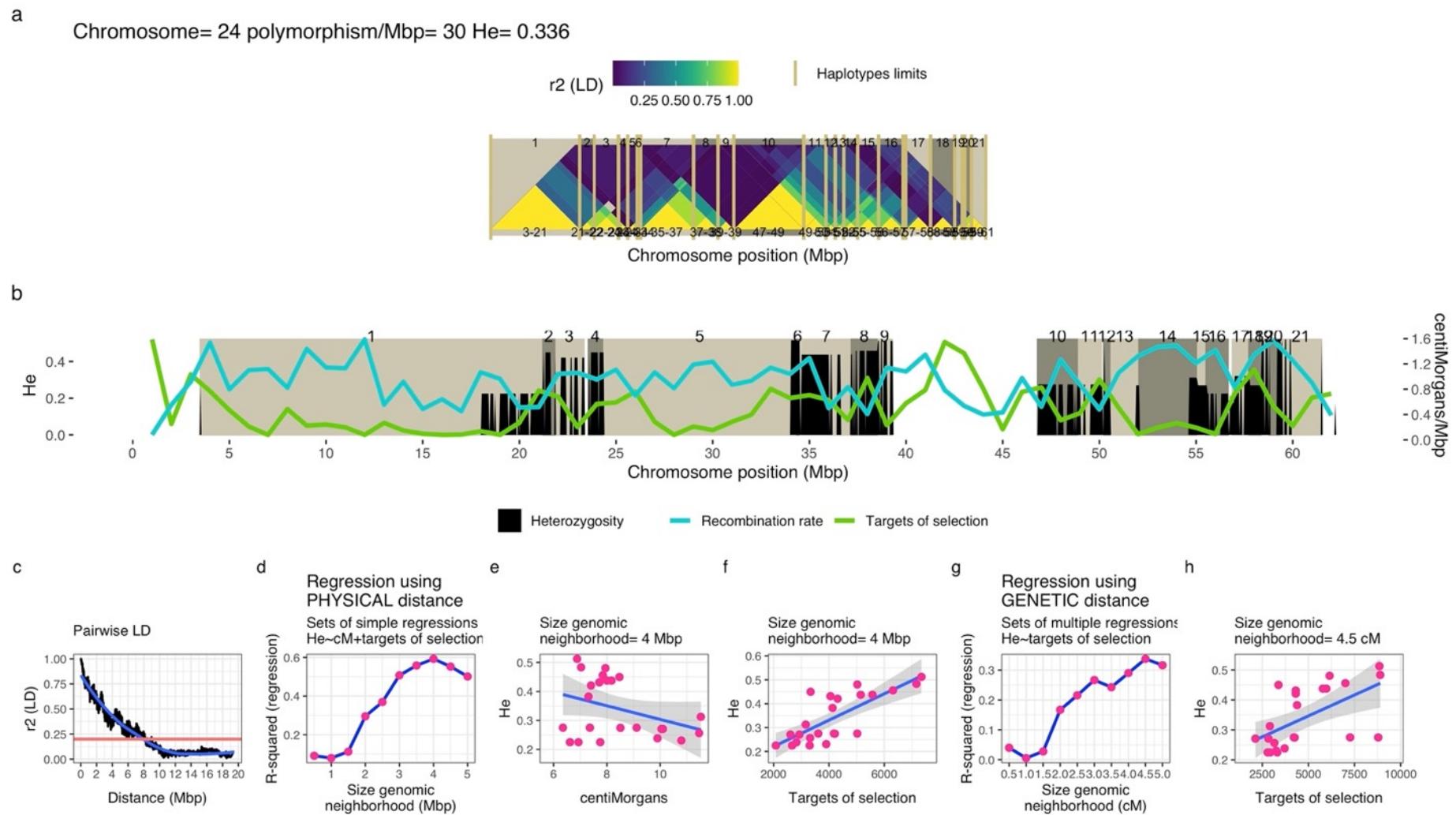


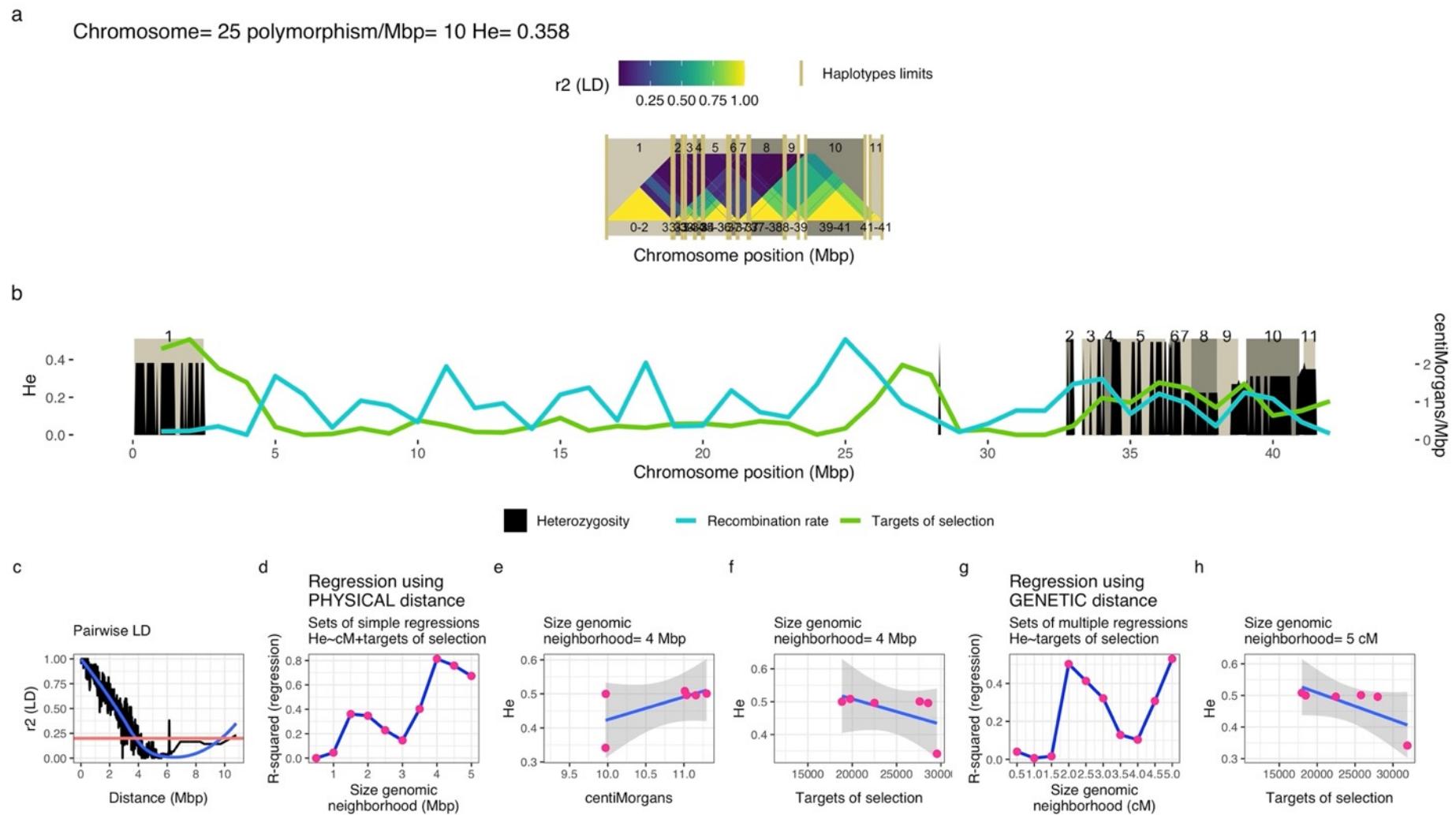


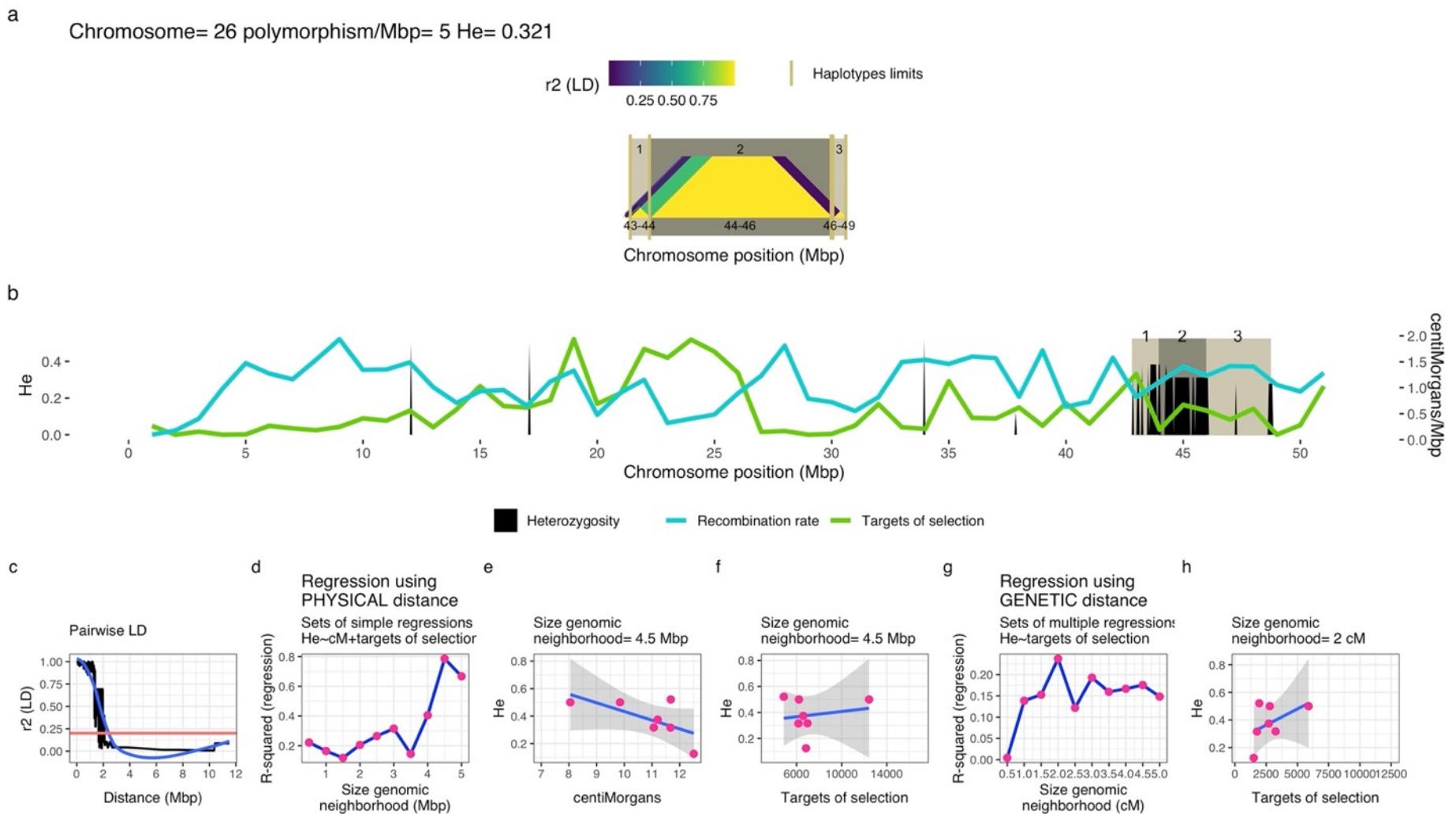


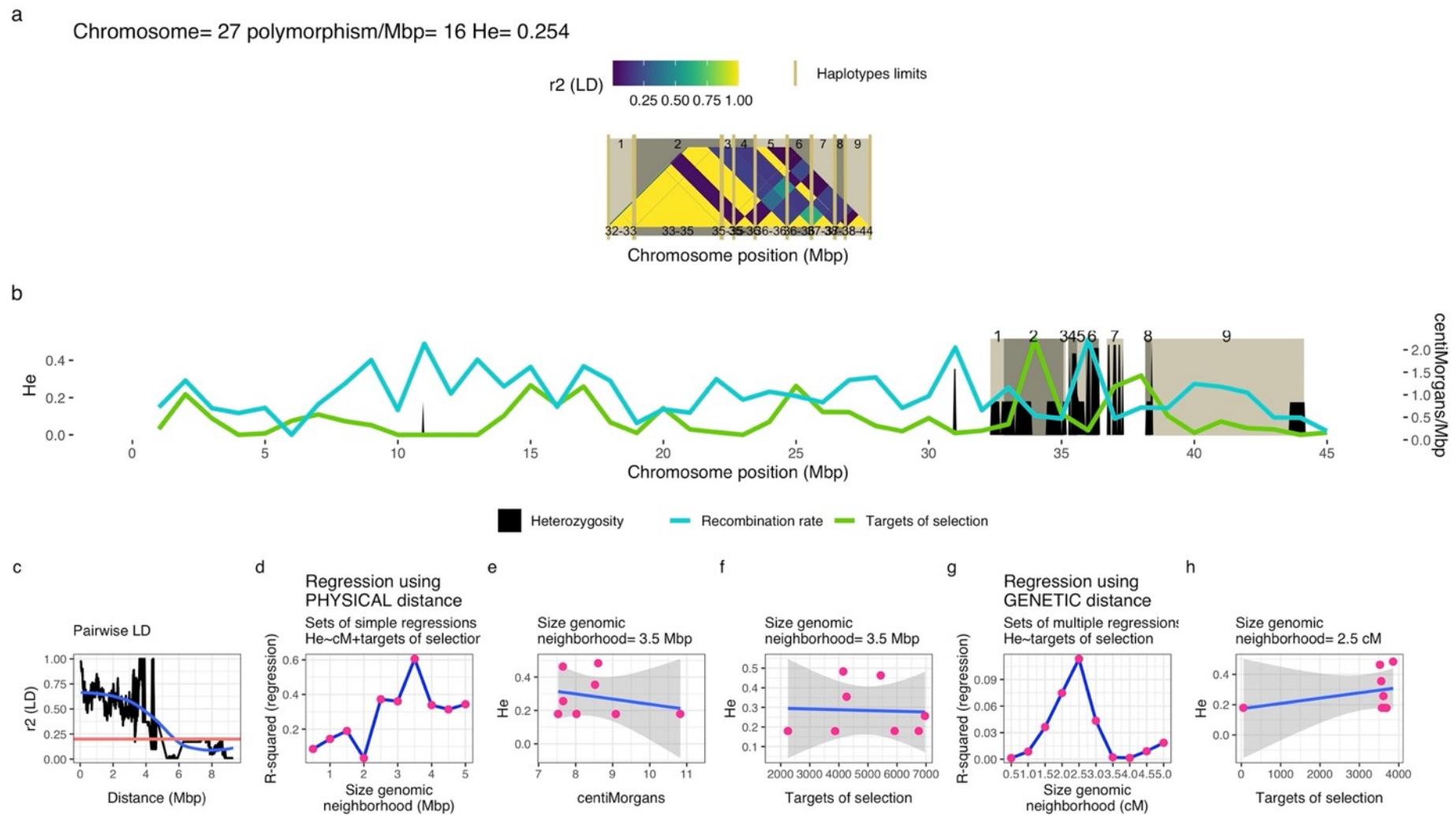


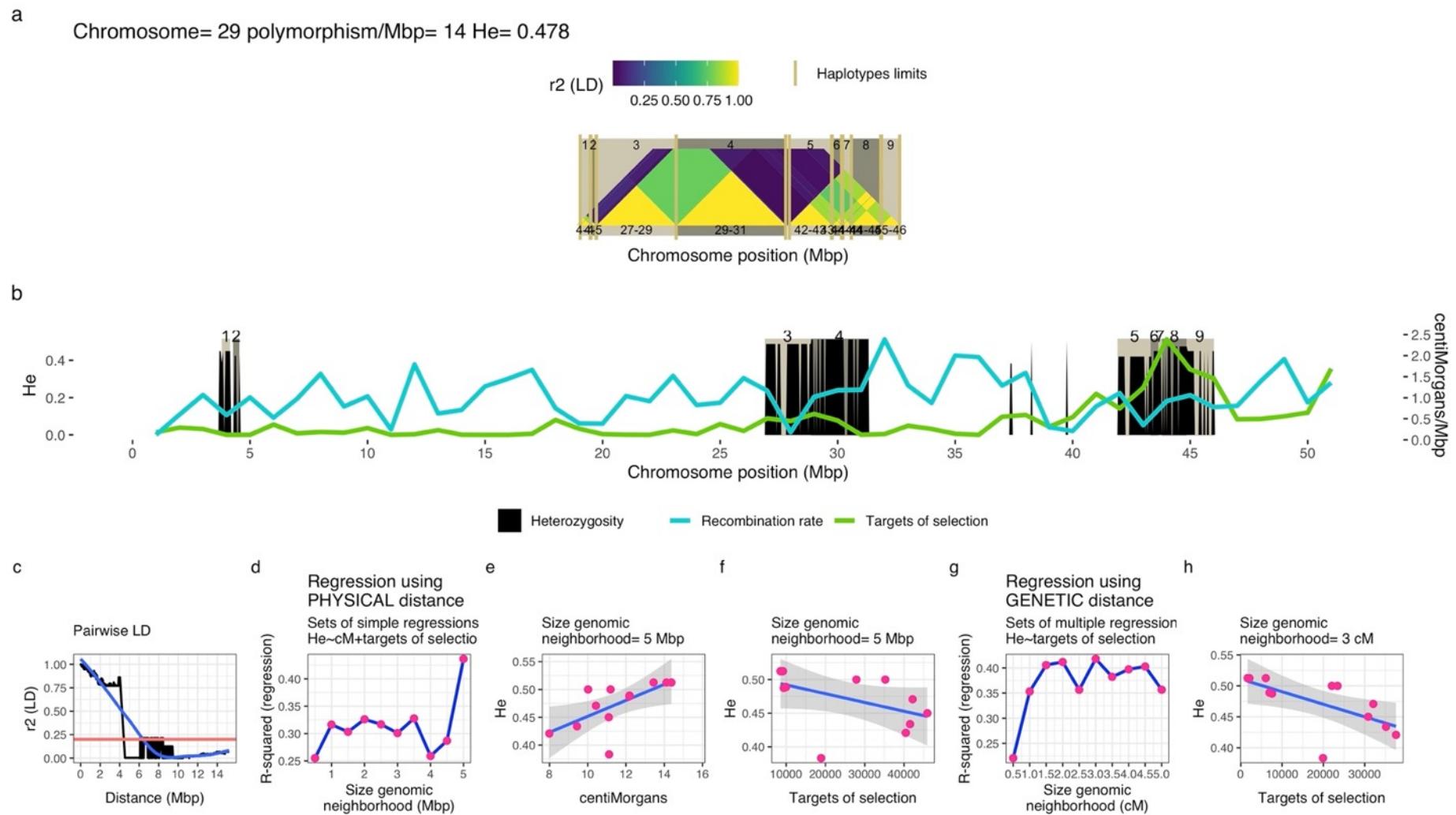












7 References

- Abdellaoui, A., Hottenga, J.-J., De Knijff, P., Nivard, M. G., Xiao, X., Scheet, P., . . . Davies, G. E. (2013). Population structure, migration, and diversifying selection in the Netherlands. *European journal of human genetics*, 21(11), 1277-1285.
- Adrion, J. R., Galloway, J. G., & Kern, A. D. (2020). Predicting the landscape of recombination using deep learning. *Molecular Biology and Evolution*, 37(6), 1790-1808.
- Agrawal, A. F., & Whitlock, M. C. (2012). Mutation load: the fitness of individuals in populations where deleterious alleles are abundant. *Annual Review of Ecology, Evolution, and Systematics*, 43, 115-135.
- Alcala, N., & Rosenberg, N. A. (2017). Mathematical constraints on *FST*: Biallelic markers in arbitrarily many populations. *Genetics*(206), 1581-1600.
- Allendorf, F. W., Luikart, G., & Aitken, S. N. (2013). Conservation and the genetics of populations (Second ed.). Malden, Mass.: Blackwell Publishing.
- Anich, P. S., Anthony, S., Carlson, M., Gunnelson, A., Kohler, A. M., Martin, J. G., & Olson, E. R. (2021). Biofluorescence in the platypus (*Ornithorhynchus anatinus*). *Mammalia*, 85(2), 179-181.
- Arenas, M. (2013). Computer programs and methodologies for the simulation of DNA sequence data with recombination. *Frontiers in genetics*, 4, 9.
- Avery, P. J., & Hill, W. G. (1977). Variability in genetic parameters among small populations. *Genetics Research*, 29(3), 193-213.
- Baguette, M., Blanchet, S., Legrand, D., Stevens, V. M., & Turlure, C. (2013). Individual dispersal, landscape connectivity and ecological networks. *Biological Reviews*, 88(2), 310-326.
- Balkenhol, N., Cushman, S., Storfer, A., & Waits, L. (2015). Landscape genetics: concepts, methods, applications: John Wiley & Sons.
- Balloux, F. (2001). EASYPop (version 1.7): A computer program for population genetics simulations. *Journal of Heredity*, 92(3), 301-302.
- Beaumont, M. A., & Nichols, R. A. (1996). Evaluating loci for use in the genetic analysis of population structure. *Proc. R. Soc. Lond. B*, 263(1377), 1619-1626.

- Becher, H., Jackson, B. C., & Charlesworth, B. (2020). Patterns of genetic variability in genomic regions with low rates of recombination. Current Biology, 30(1), 94-100.*
- Begun, D. J., & Lindfors, H. A. (2005). Rapid evolution of genomic Acp complement in the melanogaster subgroup of *Drosophila*. Molecular Biology and Evolution, 22(10), 2010-2021.*
- Bergen, A. C. (2015). Mutation load under additive fitness effects. Genetics Research, 97.*
- Berner, D. (2019). Allele frequency difference AFD—an intuitive alternative to FST for quantifying genetic population differentiation. Genes, 10(4), 308.*
- Bethge, P., Munks, S., Otley, H., & Nicol, S. (2003). Diving behaviour, dive cycles and aerobic dive limit in the platypus *Ornithorhynchus anatinus*. Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology, 136(4), 799-809.*
- Bierne, N., Roze, D., & Welch, J. J. (2013). Pervasive selection or is it...? why are FST outliers sometimes so frequent? Molecular Ecology, 22(8), 2061-2064.*
- Bino, G., Kingsford, R. T., Archer, M., Connolly, J. H., Day, J., Dias, K., ... Griffiths, J. (2019). The platypus: evolutionary history, biology, and an uncertain future. Journal of Mammalogy, 100(2), 308-327.*
- Bino, G., Kingsford, R. T., Grant, T., Taylor, M. D., & Vogelnest, L. (2018). Use of implanted acoustic tags to assess platypus movement behaviour across spatial and temporal scales. Scientific reports, 8(1), 1-12*
- Bino, G., Kingsford, R. T., & Wintle, B. A. (2020). A stitch in time—Synergistic impacts to platypus metapopulation extinction risk. Biological Conservation, 242, 108399.*
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114-2120.*
- Bolker, B. M. (2008). Ecological models and data in R: Princeton University Press.*
- Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. Proceedings of the National Academy of Sciences, 99(suppl 3), 7280-7287.*
- Brandies, P., Peel, E., Hogg, C. J., & Belov, K. (2019). The Value of Reference Genomes in the Conservation of Threatened Species. Genes, 10(11), 846.*

- Brassel, K. E., & Reif, D. (1979). A procedure to generate Thiessen polygons. Geographical Analysis, 11(3), 289-303.*
- Brauer, C. J., Hammer, M. P., & Beheregaray, L. B. (2016). Riverscape genomics of a threatened fish across a hydroclimatically heterogeneous river basin. Molecular Ecology, 25(20), 5093-5113.*
- Broad Institute. (2019). Picard toolkit. Broad Institute, GitHub repository. Retrieved from <http://broadinstitute.github.io/picard/>*
- Broadhurst, B., Ebner, B., Lintermans, M., Thiem, J., & Clear, R. (2013). Jailbreak: a fishway releases the endangered Macquarie perch from confinement below an anthropogenic barrier. Marine and Freshwater Research, 64(9), 900-908.*
- Brooks, A. J., Russell, M., Bevitt, R., & Dasey, M. (2011). Constraints on the recovery of invertebrate assemblages in a regulated snowmelt river during a tributary-sourced environmental flow regime. Marine and Freshwater Research, 62(12), 1407-1420.*
- Brown, J. J., Limburg, K. E., Waldman, J. R., Stephenson, K., Glenn, E. P., Juanes, F., & Jordaan, A. (2013). Fish and hydropower on the US Atlantic coast: failed fisheries policies from half-way technologies. Conservation Letters, 6(4), 280-286.*
- Bryant, L. M., & Krosch, M. N. (2016). Lines in the land: a review of evidence for eastern Australia's major biogeographical barriers to closed forest taxa. Biological Journal of the Linnean Society, 119(2), 238-264.*
- Campos, J. L., Zhao, L., & Charlesworth, B. (2017). Estimating the parameters of background selection and selective sweeps in *Drosophila* in the presence of gene conversion. Proceedings of the National Academy of Sciences, 114(24), E4762-E4771.*
- Casillas, S., & Barbadilla, A. (2017). Molecular population genetics. Genetics, 205(3), 1003-1035.*
- Castellano, D., Coronado-Zamora, M., Campos, J. L., Barbadilla, A., & Eyre-Walker, A. (2016). Adaptive evolution is substantially impeded by Hill-Robertson interference in *Drosophila*. Molecular Biology and Evolution, 33(2), 442-455.*
- Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., . . . De Rosario, M. H. (2018). Package 'pwr'. R package version, 1(2).*

- Chao, A., Ma, K. H., Hsieh, T. C., & Chiu, C. H. (2015). Online program SpadeR (Species-richness prediction and diversity estimation in R). Program and User's Guide published at http://chao.stat.nthu.edu.tw/wordpress/software_download.*
- Charlesworth, B. (1994). The effect of background selection against deleterious mutations on weakly selected, linked variants. Genetical Research, 63(03), 213-227.*
- Charlesworth, B. (2015). Causes of natural variation in fitness: evidence from studies of Drosophila populations. Proceedings of the National Academy of Sciences, 112(6), 1662-1669.*
- Charlesworth, B., & Charlesworth, D. (2010). Elements of Evolutionary Genetics: Roberts and Company Publishers.*
- Charlesworth, B., & Charlesworth, D. (2018). Neutral variation in the context of selection. Molecular Biology and Evolution, 35(6), 1359-1361.*
- Charlesworth, D., & Willis, J. H. (2009). The genetics of inbreeding depression. Nature Reviews Genetics, 10(11), 783-796.*
- Chen, B., Cole, J., & Grond-Ginsbach, C. (2017). Departure from Hardy Weinberg equilibrium and genotyping error. Frontiers in genetics, 8, 167.*
- Chen, J., Gléménin, S., & Lascoux, M. (2017). Genetic diversity and the efficacy of purifying selection across plant and animal species. Molecular Biology and Evolution, 34(6), 1417-1428.*
- Chester, H., & Norris, R. (2006). Dams and flow in the Cotter River, Australia: effects on instream trophic structure and benthic metabolism. Hydrobiologia, 572(1), 275-286.*
- Chu, D., & Wei, L. (2019). Nonsynonymous, synonymous and nonsense mutations in human cancer-related genes undergo stronger purifying selections than expectation. BMC cancer, 19(1), 1-12.*
- Chun, S., & Fay, J. C. (2009). Identification of deleterious mutations within three human genomes. Genome research, 19(9), 1553-1561.*
- Clayton, D. (2015). snpStats: SnpMatrix and XSnpMatrix classes and methods. R package version, 1(0).*

- Clayton, D. (2018). *snpStats: SnpMatrix and XSnpMatrix classes and methods. R package version 1.32.0.**
- Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*: Academic press.*
- Coleman, R., Gauffre, B., Pavlova, A., Beheregaray, L., Kearns, J., Lyon, J., . . . Sunnucks, P. (2018). Artificial barriers prevent genetic recovery of small isolated populations of a low-mobility freshwater fish. *Heredity*, 120(6), 515-532.*
- Comeron, J. M. (2014). Background selection as baseline for nucleotide variation across the *Drosophila* genome. *PLoS genetics*, 10(6), e1004434.*
- Comeron, J. M., Ratnappan, R., & Bailin, S. (2012). The many landscapes of recombination in *Drosophila melanogaster*. *PLoS genetics*, 8(10), e1002905.*
- Corbett-Detig, R. B., Hartl, D. L., & Sackton, T. B. (2015). Natural selection constrains neutral diversity across a wide range of species. *Plos Biology*, 13(4).*
- Cronin, M. A., Cánovas, A., Bannasch, D. L., Oberbauer, A. M., & Medrano, J. F. (2015). Single nucleotide polymorphism (SNP) variation of wolves (*Canis lupus*) in Southeast Alaska and comparison with wolves, dogs, and coyotes in North America. *Journal of Heredity*, 106(1), 26-36.*
- Crow, J. F., & Kimura, M. (1970). *An introduction to population genetics theory. An introduction to population genetics theory*.*
- Cutter, A. D., & Payseur, B. A. (2013). Genomic signatures of selection at linked sites: unifying the disparity among species. *Nature Reviews Genetics*, 14(4), 262-274.*
- Darwin, C. U. (1859). *On the Origin of Species by Means of Natural Selection, Or, The Preservation of Favoured Races in the Struggle for Life*: J. Murray.*
- de Boer, R. A., Vega-Trejo, R., Kotrschal, A., & Fitzpatrick, J. L. (2021). Meta-analytic evidence that animals rarely avoid inbreeding. *Nature Ecology & Evolution*, 1-16.*
- DeAngelis, D. L., & Rose, K. A. (1992). Which individual-based approach is most appropriate for a given problem. *Individual-based models and approaches in ecology: populations, communities and ecosystems*, 67-87.*
- Delourme, R., Falentin, C., Fomeju, B. F., Boillot, M., Lassalle, G., André, I., . . . Marty, A. (2013). High-density SNP-based genetic map development and linkage disequilibrium assessment in *Brassica napus*L. *BMC genomics*, 14(1), 120.*

- Dementieva, N., Mitrofanova, O., Dysin, A., Kudinov, A., Stanishevskaya, O., Larkina, T., . . . Smaragdov, M. (2021). Assessing the effects of rare alleles and linkage disequilibrium on estimates of genetic diversity in the chicken populations. *Animal*, 15(3), 100171.
- Do, C., Waples, R. S., Peel, D., Macbeth, G. M., Tillett, B. J., & Ovenden, J. R. (2014). NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size (N_e) from genetic data. *Molecular Ecology Resources*, 14(1), 209-214.
- Dobzhansky, T. (1955). A review of some fundamental concepts and problems of population genetics.
- dos Santos, G., Schroeder, A. J., Goodman, J. L., Strelets, V. B., Crosby, M. A., Thurmond, J., . . . Consortium, F. (2015). FlyBase: introduction of the *Drosophila melanogaster* Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Research*, 43(D1), D690-D697.
- Dréau, A., Venu, V., Avdievich, E., Gaspar, L., & Jones, F. C. (2019). Genome-wide recombination map construction from single individuals using linked-read sequencing. *Nature communications*, 10(1), 1-11.
- Drechsler, M., & Burgman, M. A. (2004). Combining population viability analysis with decision analysis. *Biodiversity and Conservation*, 13(1), 115-139.
- Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature protocols*, 4(8), 1184.
- Eldridge, M. D. B., Deakin, J. E., MacDonald, A. J., Byrne, M., Fitzgerald, A., Johnson, R. N., . . . Young, A. (2020). The Oz Mammals Genomics (OMG) initiative: developing genomic resources for mammal conservation at a continental scale. *Australian Zoologist*, 40(3), 505-509.
- Ellegren, H., & Galtier, N. (2016). Determinants of genetic diversity. *Nature Reviews Genetics*, 17(7), 422-433.
- Elyashiv, E., Sattath, S., Hu, T. T., Strutsovsky, A., McVicker, G., Andolfatto, P., . . . Sella, G. (2016). A genomic map of the effects of linked selection in *Drosophila*. *PLoS genetics*, 12(8), e1006130.

- England, P. R. (1997). Conservation genetics of population bottlenecks. PhD thesis, Macquarie University, Sydney.,*
- England, P. R., Cornuet, J. M., Berthier, P., Tallmon, D. A., & Luikart, G. (2006). Estimating effective population size from linkage disequilibrium: severe bias in small samples. Conservation Genetics, 7(2), 303-308.*
- Eusebi, P. G., Martinez, A., & Cortes, O. (2020). Genomic tools for effective conservation of livestock breed diversity. Diversity, 12(1), 8.*
- Excoffier, L., Smouse, P. E., & Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes - Application to human mitochondrial-DNA restriction data. Genetics, 131(2), 479-491.*
- Fasiolo, M., Nedellec, R., Goude, Y., & Wood, S. N. (2020). Scalable visualization methods for modern generalized additive models. Journal of computational and Graphical Statistics, 29(1), 78-86.*
- Fisher, R. A. (1930). The genetical theory of natural selection: a complete variorum edition: Oxford University Press.*
- Foll, M., & Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics, 180(2), 977-993.*
- Frankham, R. (1995). Effective population-size adult-population size ratios in wildlife - a review. Genetical Research, 66(2), 95-107.*
- Frankham, R. (2005). Genetics and extinction. Biological Conservation, 126(2), 131-140. doi:10.1016/j.biocon.2005.05.002*
- Frankham, R. (2015). Genetic rescue of small inbred populations: Meta-analysis reveals large and consistent benefits of gene flow. Molecular Ecology, 24(11), 2610-2618.*
- Frankham, R., Ballou, J. D., & Briscoe, D. A. (2009). Introduction to conservation genetics (2nd ed.). Cambridge; New York: Cambridge University Press.*
- Frankham, R., Ballou, J. D., Ralls, K., Eldridge, M., Dudash, M. R., Fenster, C. B., . . . Sunnucks, P. (2017). Genetic management of fragmented animal and plant populations: Oxford University Press.*

- Franklin, I. R., & Frankham, R. (1998). How large must populations be to retain evolutionary potential? *Animal Conservation*, 1(1), 69-70. doi:10.1111/j.1469-1795.1998.tb00228.x
- Frydenberg, O. V. E. (1963). Population studies of a lethal mutant in *Drosophila Melanogaster*. *Hereditas*, 50(1), 89-116.
- Furlan, E. M., Griffiths, J., Gust, N., Handasyde, K. A., Grant, T. R., Gruber, B., & Weeks, A. R. (2013). Dispersal patterns and population structuring among platypuses, *Ornithorhynchus anatinus*, throughout south-eastern Australia. *Conservation Genetics*, 14(4), 837-853.
- Galtier, N. (2016). Adaptive protein evolution in animals and the effective population size hypothesis. *PLoS genetics*, 12(1), e1005774.
- Garant, D., Forde, S. E., & Hendry, A. P. (2007). The multifarious effects of dispersal and gene flow on contemporary adaptation. *Functional Ecology*, 434-443.
- García-Dorado, A., & Caballero, A. (2000). On the average coefficient of dominance of deleterious spontaneous mutations. *Genetics*, 155(4), 1991-2001.
- Georges, A., Gruber, B., Pauly, G. B., White, D., Adams, M., Young, M. J., . . . Unmack, P. J. (2018). Genomewide SNP markers breathe new life into phylogeography and species delimitation for the problematic short-necked turtles (Chelidae: *Emydura*) of eastern Australia. *Molecular Ecology*, 27(24), 5195-5213.
- Gilbert, K., Pouyet, F., Excoffier, L., & Peischl, S. (2020). Transition from background selection to associative overdominance promotes diversity in regions of low recombination. *Current Biology*, 30(1), 101-107.
- Gilbert, K. J., Peischl, S., & Excoffier, L. (2018). Mutation load dynamics during environmentally-driven range shifts. *PLoS genetics*, 14(9), e1007450.
- Gilbert, M. C., Akama, A., Fernandes, C. C., & Albertson, R. C. (2020). Rapid morphological change in multiple cichlid ecotypes following the damming of a major clearwater river in Brazil. *Evolutionary Applications*, 13(10), 2754.
- Gillespie, J. H. (2000). Genetic Drift in an Infinite Population: The Pseudohitchhiking Model. *Genetics*, 155(2), 909-919.
- Gilligan, D. M. (2001). Conservation genetics and long-term survival: testing models using *Drosophila*. Macquarie University,

- Gilligan, D. M., Briscoe, D. A., & Frankham, R. (2005). Comparative losses of quantitative and molecular genetic variation in finite populations of *Drosophila melanogaster*. *Genetical Research*, 85(01), 47-55.*
- Goudet, J. (2005). Hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes*, 5(1), 184-186.*
- Grant, T., & Llewellyn, L. C. (1991). The biology and management of the platypus (*Ornithorhynchus anatinus*) in NSW: NSW National Parks and Wildlife Service.*
- Grant, T. R. (1998). Current and Historical Occurrence of Platypuses, *Ornithorhynchus anatinus*, around Sydney. *Australian Mammalogy*, 20(2), 257-266.*
- Grant, T. R. (2004). Captures, capture mortality, age and sex ratios of platypuses, *Ornithorhynchus anatinus*, during studies over 30 years in the upper Shoalhaven River in New South Wales.*
- Grant, T. R., & Fanning, D. (2007). Platypus: CSIRO PUBLISHING.*
- Grant, T. R., & Temple-Smith, P. D. (2003). Conservation of the platypus, *Ornithorhynchus anatinus*: threats and challenges. *Aquatic Ecosystem Health & Management*, 6(1), 5-18.*
- Griffin, P. C., Hangartner, S. B., Fournier-Level, A., & Hoffmann, A. A. (2017). Genomic trajectories to desiccation resistance: convergence and divergence among replicate selected *Drosophila* lines. *Genetics*, 205(2), 871-890.*
- Grimm, V., Berger, U., Bastiansen, F., Eliassen, S., Ginot, V., Giske, J., . . . Huse, G. (2006). A standard protocol for describing individual-based and agent-based models. *Ecological modelling*, 198(1-2), 115-126.*
- Grimm, V., Railsback, S. F., Vincenot, C. E., Berger, U., Gallagher, C., DeAngelis, D. L., . . . Groeneveld, J. (2020). The ODD protocol for describing agent-based and other simulation models: A second update to improve clarity, replication, and structural realism. *Journal of Artificial Societies and Social Simulation*, 23(2).*
- Groombridge, J. J., Jones, C. G., Bruford, M. W., & Nichols, R. A. (2000). 'Ghost'alleles of the Mauritius kestrel. *Nature*, 403(6770), 616.*
- Gruber, B., Unmack, P. J., Berry, O. F., & Georges, A. (2018). dartr: An r package to facilitate analysis of SNP data generated from reduced representation genome sequencing. *Molecular Ecology Resources*, 18(3), 691-699.*

- Guillot, G., Estoup, A., Mortier, F., & Cosson, J. F. (2005). A spatial statistical model for landscape genetics. *Genetics*, 170(3), 1261-1280. doi:10.1534/genetics.104.033803
- Gunn, M. (2003). *The use of microsatellites as a surrogate for quantitative trait variation.* PhD Thesis, University of New South Wales, Sydney,
- Haag-Liautard, C., Dorris, M., Maside, X., Macaskill, S., Halligan, D. L., Charlesworth, B., & Keightley, P. D. (2007). Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*. *Nature*, 445(7123), 82-85.
- Haddrill, P. R., Loewe, L., & Charlesworth, B. (2010). Estimating the parameters of selection on nonsynonymous mutations in *Drosophila pseudoobscura* and *D. miranda*. *Genetics*, 185(4), 1381-1396.
- Haldane, J. B. S. (1932). *The causes of evolution*: Longmans, Green and co.
- Haller, B. C. (2014). Theoretical and empirical perspectives in ecology and evolution: a survey. *Bioscience*, 64(10), 907-916.
- Haller, B. C., & Messer, P. W. (2019). SLiM 3: forward genetic simulations beyond the Wright–Fisher model. *Molecular Biology and Evolution*, 36(3), 632-637.
- Halliburton, R. (2004). *Introduction to population genetics*: Pearson/Prentice Hall Upper Saddle River, NJ.
- Hardy, G. H. (1908). Mendelian proportions in a mixed population. *Science*, 28(706), 49-50.
- Harris, H. (1966). C. Genetics of Man Enzyme polymorphisms in man. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 164(995), 298-310.
- Hawke, T., Bino, G., & Kingsford, R. T. (2019). A silent demise: Historical insights into population changes of the iconic platypus (*Ornithorhynchus anatinus*). *Global Ecology and Conservation*, 20, e00720.
- Hawke, T., Bino, G., & Kingsford, R. T. (2021). Damming insights: Variable impacts and implications of river regulation on platypus populations. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 31(3), 504-519.
- Hedrick, P. (2011). *Genetics of populations*: Jones & Bartlett Learning.
- Hedrick, P. W. (2005). A standardized genetic differentiation measure. *Evolution*, 59(8), 1633-1638.

- Hedrick, P. W., Peterson, R. O., Vucetich, L. M., Adams, J. R., & Vucetich, J. A. (2014). *Genetic rescue in Isle Royale wolves: genetic analysis and the collapse of the population*. *Conservation Genetics*, 15(5), 1111-1121.
- Hill, W. G., & Robertson, A. (1966). *The effect of linkage on limits to artificial selection*. *Genetics Research*, 8(3), 269-294.
- Hill, W. G., & Robertson, A. (1968). *Linkage disequilibrium in finite populations*. *Theoretical and Applied Genetics*, 38(6), 226-231.
- Hoban, S., Bertorelle, G., & Gaggiotti, O. E. (2012). *Computer simulations: tools for population and evolutionary genetics*. *Nature Reviews Genetics*, 13(2), 110-122. doi:10.1038/nrg3130
- Hoffman, J. R., Willoughby, J. R., Swanson, B. J., Pangle, K. L., & Zanatta, D. T. (2017). *Detection of barriers to dispersal is masked by long lifespans and large population sizes*. *Ecology and Evolution*, 7(22), 9613-9623.
- Hoffmann, A. A., Miller, A. D., & Weeks, A. R. (2020). *Genetic mixing for population management: From genetic rescue to provenancing*. *Evolutionary Applications*.
- Holderegger, R., & Wagner, H. H. (2008). *Landscape genetics*. *Bioscience*, 58(3), 199-207.
- Holleley, C. E. (2009). *Testing hypotheses in molecular ecology: genetic exchange and hybrid performance*. PhD Thesis, University of New South Wales, Sydney,
- Holleley, C. E., Nichols, R. A., Whitehead, M. R., Adamack, A. T., Gunn, M. R., & Sherwin, W. B. (2014). *Testing single-sample estimators of effective population size in genetically structured populations*. *Conservation Genetics*, 15(1), 23-35.
- Holleley, C. E., Nichols, R. A., Whitehead, M. R., Gunn, M. R., Gupta, J., & Sherwin, W. B. (2011). *Induced dispersal in wildlife management: experimental evaluation of the risk of hybrid breakdown and the benefit of hybrid vigor in the F1 generation*. *Conservation Genetics*, 12(1), 31-40.
- Holsinger, K. E. (2020). *Lecture notes in population genetics*. Creative Commons License California, USA.
- Holsinger, K. E., & Weir, B. S. (2009). *Genetics in geographically structured populations: defining, estimating and interpreting FST*. *Nature Reviews Genetics*, 10(9), 639-650.

- Hoskins, R. A., Carlson, J. W., Kennedy, C., Acevedo, D., Evans-Holm, M., Frise, E., . . . Rossi, F. (2007). Sequence finishing and mapping of *Drosophila melanogaster* heterochromatin. *Science*, 316(5831), 1625-1628.
- Hubby, J. L., & Lewontin, R. C. (1966). A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in *Drosophila pseudoobscura*. *Genetics*, 54(2), 577-594.
- Huber, C. D., Durvasula, A., Hancock, A. M., & Lohmueller, K. E. (2018). Gene expression drives the evolution of dominance. *Nature communications*, 9(1), 2750.
- Huber, C. D., Kim, B. Y., Marsden, C. D., & Lohmueller, K. E. (2017). Determining the factors driving selective effects of new nonsynonymous mutations. *Proceedings of the National Academy of Sciences*, 114(17), 4465-4470.
- Isaac, N. J. B., Turvey, S. T., Collen, B., Waterman, C., & Baillie, J. E. M. (2007). Mammals on the EDGE: conservation priorities based on threat and phylogeny. *PLoS One*, 2(3), e296.
- Johnson, F. M., Kanapi, C. G., Richardson, R. H., Wheeler, M. R., & Stone, W. S. (1966). An analysis of polymorphisms among isozyme loci in dark and light *Drosophila ananassae* strains from American and Western Samoa. *Proceedings of the National Academy of Sciences of the United States of America*, 56(1), 119-125.
- Johri, P., Charlesworth, B., & Jensen, J. D. (2020). Toward an evolutionarily appropriate null model: Jointly inferring demography and purifying selection. *Genetics*, 215(1), 173-192.
- Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., . . . White, S. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, 484(7392), 55-61.
- Jordan, K. W., Carbone, M. A., Yamamoto, A., Morgan, T. J., & Mackay, T. F. C. (2007). Quantitative genomics of locomotor behavior in *Drosophila melanogaster*. *Genome biology*, 8(8), R172.
- Jost, L. (2008). GST and its relatives do not measure differentiation. *Molecular Ecology*, 17(18), 4015-4026.
- Keightley, P. D. (2012). Rates and fitness consequences of new mutations in humans. *Genetics*, 190(2), 295-304.

- Keller, M. C., Visscher, P. M., & Goddard, M. E. (2011). Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data. Genetics, 189(1), 237-249.*
- Kilian, A., Wenzl, P., Huttner, E., Carling, J., Xia, L., Blois, H., . . . Hopper, C. (2012). Diversity arrays technology: a generic genome profiling technology on open platforms. In Data production and analysis in population genomics (pp. 67-89): Springer.*
- Kimura, M. (1964). Diffusion models in population genetics. Journal of Applied Probability, 1(2), 177-232.*
- Kimura, M. (1965). Attainment of quasi linkage equilibrium when gene frequencies are changing by natural selection. Genetics, 52(5), 875.*
- Kimura, M. (1968). Evolutionary rate at the molecular level. Nature, 217(5129), 624-626.*
- Kimura, M. (1983). The neutral theory of molecular evolution: Cambridge University Press.*
- Kimura, M., & Crow, J. F. (1963). The measurement of effective population number. Evolution, 17(3), 279-288.*
- Kimura, M., Maruyama, T., & Crow, J. F. (1963). The mutation load in small populations. Genetics, 48(10), 1303-1312.*
- Kimura, M., & Ohta, T. (1969). The average number of generations until fixation of a mutant gene in a finite population. Genetics, 61(3), 763.*
- Kingman, J. F. C. (1982). The coalescent. Stochastic processes and their applications, 13(3), 235-248.*
- Klamt, M., Thompson, R., & Davis, J. (2011). Early response of the platypus to climate warming. Global Change Biology, 17(10), 3011-3018.*
- Kliman, R. M. (2014). EvolGenius 6.1 program validation. Retrieved from http://www2.cedarcrest.edu/academic/bio/rkliman/EG_Validation.pdf*
- Kokko, H., & Ots, I. (2006). When not to avoid inbreeding. Evolution, 60(3), 467-475.*
- Kolomyjec, S. H. (2010). The history and relationships of northern platypus (*Ornithorhynchus anatinus*) populations: a molecular approach. James Cook University,*

- Kolomyjec, S. H., Chong, J. Y. T., Blair, D., Gongora, J., Grant, T. R., Johnson, C. N., & Moran, C. (2009). Population genetics of the platypus (*Ornithorhynchus anatinus*): a fine-scale look at adjacent river systems. *Australian Journal of Zoology*, 57(4), 225-234.*
- Kolomyjec, S. H., Grant, T. R., Johnson, C. N., & Blair, D. (2014). Regional population structuring and conservation units in the platypus (*Ornithorhynchus anatinus*). *Australian Journal of Zoology*, 61(5), 378-385.*
- Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A., Richardsson, B., . . . Masson, G. (2002). A high-resolution recombination map of the human genome. *Nature genetics*, 31(3), 241.*
- Kousathanas, A., & Keightley, P. D. (2013). A comparison of models to infer the distribution of fitness effects of new mutations. *Genetics*, 193(4), 1197-1208.*
- Krimbas, C. B., & Powell, J. R. (1992). *Drosophila inversion polymorphism*: CRC press.*
- Kryazhimskiy, S., & Plotkin, J. B. (2008). The population genetics of dN/dS. *PLoS Genet*, 4(12), e1000304.*
- Landguth, E. L., Cushman, S. A., Schwartz, M. K., McKelvey, K. S., Murphy, M., & Luikart, G. (2010). Quantifying the lag time to detect barriers in landscape genetics. *Molecular Ecology*, 19(19), 4179-4191. doi:Doi 10.1111/j.1365-294x.2010.04808.x*
- Latter, B. D. H. (1998). Mutant alleles of small effect are primarily responsible for the loss of fitness with slow inbreeding in *Drosophila melanogaster*. *Genetics*, 148(3), 1143-1158.*
- Lawrence, H. A., Taylor, G. A., Millar, C. D., & Lambert, D. M. (2008). High mitochondrial and nuclear genetic diversity in one of the world's most endangered seabirds, the Chatham Island Taiko (*Pterodroma magentae*). *Conservation Genetics*, 9(5), 1293-1301.*
- Lebeuf-Taylor, E., McCloskey, N., Bailey, S. F., Hinz, A., & Kassen, R. (2019). The distribution of fitness effects among synonymous mutations in a gene under directional selection. *eLife*, 8, e45952.*
- Lee, B. T. K., Tan, T. W., & Ranganathan, S. (2004). DEDB: a database of *Drosophila melanogaster* exons in splicing graph form. *BMC bioinformatics*, 5(1), 189.*

- Leffler, E. M., Bullaughey, K., Matute, D. R., Meyer, W. K., Segurel, L., Venkat, A., . . . Przeworski, M. (2012). Revisiting an old riddle: what determines genetic diversity levels within species? PLoS Biol, 10(9), e1001388.*
- Lemopoulos, A., Prokkola, J. M., Uusi-Heikkilä, S., Vasemägi, A., Huusko, A., Hyvärinen, P., . . . Vainikka, A. (2019). Comparing RADseq and microsatellites for estimating genetic diversity and relatedness—Implications for brown trout conservation. Ecology and Evolution, 9(4), 2106-2120.*
- Lesecque, Y., Keightley, P. D., & Eyre-Walker, A. (2012). A resolution of the mutation load paradox in humans. Genetics, 191(4), 1321-1330.*
- Lewontin, R. C., & Hubby, J. L. (1966). A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. Genetics, 54(2), 595-609.*
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics, 25(14), 1754-1760.*
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Durbin, R. (2009). The sequence alignment/map format and SAMtools. Bioinformatics, 25(16), 2078-2079.*
- Li, X., Han, Y., Wei, Y., Acharya, A., Farmer, A. D., Ho, J., . . . Brummer, E. C. (2014). Development of an alfalfa SNP array and its use to evaluate patterns of population structure and linkage disequilibrium. PLoS One, 9(1), e84329.*
- Linck, E., & Battey, C. J. (2019). Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. Molecular Ecology Resources, 19(3), 639-647.*
- Lindsley, D. L., & Grell, E. H. (1968). Genetic variations of *Drosophila melanogaster*. 図書 11-12.*
- Lobkovsky, A. E., Levi, L., Wolf, Y. I., Maiers, M., Gragert, L., Alter, I., . . . Koonin, E. V. (2019). Multiplicative fitness, rapid haplotype discovery, and fitness decay explain evolution of human MHC. Proceedings of the National Academy of Sciences, 116(28), 14098-14104.*

- Ma, L., O'Connell, J. R., VanRaden, P. M., Shen, B., Padhi, A., Sun, C., . . . Liu, G. E. (2015). *Cattle sex-specific recombination and genetic control from a large pedigree analysis*. *PLoS genetics*, 11(11).
- MacEachern, S., Hayes, B., McEwan, J., & Goddard, M. (2009). *An examination of positive selection and changing effective population size in Angus and Holstein cattle populations (Bos taurus) using a high density SNP genotyping platform and the contribution of ancient polymorphism to genomic diversity in Domestic cattle*. *BMC genomics*, 10(1), 1-19.
- Manna, F., Martin, G., & Lenormand, T. (2011). *Fitness landscapes: an alternative theory for the dominance of mutation*. *Genetics*, 189(3), 923-937.
- Marchant, R., & Grant, T. (2015). *The productivity of the macroinvertebrate prey of the platypus in the upper Shoalhaven River, New South Wales*. *Marine and Freshwater Research*, 66(12), 1128-1137.
- Markow, T. A., & Sawka, S. (1992). *Dynamics of mating success in experimental groups of Drosophila melanogaster (Diptera: Drosophilidae)*. *Journal of insect behavior*, 5(3), 375-383.
- Marr, A. B., Keller, L. F., & Arcese, P. (2002). *Heterosis and outbreeding depression in descendants of natural immigrants to an inbred population of song sparrows (Melospiza melodia)*. *Evolution*, 56(1), 131-142.
- Marshall, T. C., & Spalton, J. A. (2000). *Simultaneous inbreeding and outbreeding depression in reintroduced Arabian oryx*. *Animal Conservation*, 3(3), 241-248.
- Martin, H. C., Batty, E. M., Hussin, J., Westall, P., Daish, T., Kolomyjec, S., . . . Grant, T. (2018). *Insights into platypus population structure and history from whole-genome sequencing*. *Molecular Biology and Evolution*, 35(5), 1238-1252.
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., . . . Daly, M. (2010). *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data*. *Genome research*, 20(9), 1297-1303.
- McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., . . . Cunningham, F. (2016). *The ensembl variant effect predictor*. *Genome biology*, 17(1), 122.

- McVean, G. (2009). A genealogical interpretation of principal components analysis. *PLoS Genet*, 5(10), e1000686.
- Meirmans, P. G., & Hedrick, P. W. (2011). Assessing population structure: F-ST and related measures. *Molecular Ecology Resources*, 11(1), 5-18. doi:10.1111/j.1755-0998.2010.02927.x
- Messer, P. W., & Petrov, D. A. (2013). Frequent adaptation and the McDonald–Kreitman test. *Proceedings of the National Academy of Sciences*, 110(21), 8615-8620.
- Mijangos, J. L., Holleley, C. E., Nichols, R. A., Towers, I. N., Jovanoski, Z., Sidhu, H. S., . . . Sherwin, W. B. (2017). Genetic linkage to explain genetic variation. Paper presented at the 22nd International Congress on Modelling and Simulation, Hobart, Australia. mssanz.org.au/modsim2017/A4/mijangos.pdf
- Mijangos, J. L., Pacioni, C., Spencer, P., & Craig, M. D. (2015). Contribution of genetics to ecological restoration. *Molecular Ecology*, 24(1), 22-37.
- Miller, J., Malenfant, R., David, P., Davis, C., Poissant, J., Hogg, J., . . . Coltman, D. (2014). Estimating genome-wide heterozygosity: effects of demographic history and marker type. *Heredity*, 112(3), 240-247.
- Miller, M. P. (2005). Alleles In Space (AIS): Computer software for the joint analysis of interindividual spatial and genetic information. *Journal of Heredity*, 96(6), 722-724. doi:10.1093/jhered/esi119
- Mills, L. S., & Allendorf, F. W. (1996). The one-migrant-per-generation rule in conservation and management. *Conservation Biology*, 10(6), 1509-1518. doi:10.1046/j.1523-1739.1996.10061509.x
- Muller, H. J. (1950). Our load of mutations. *American journal of human genetics*, 2(2), 111-176.
- Nagylaki, T. (1974). Quasilinear equilibrium and the evolution of two-locus systems. *Proceedings of the National Academy of Sciences*, 71(2), 526-530.
- Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences*, 70(12), 3321-3323.
- Nei, M. (1987). *Molecular evolutionary genetics*: Columbia University Press.

- Nislow, K. H., Hudy, M., Letcher, B. H., & Smith, E. P. (2011). Variation in local abundance and species richness of stream fishes in relation to dispersal barriers: implications for management and conservation. Freshwater Biology, 56(10), 2135-2144.*
- O'Grady, J. J., Brook, B. W., Reed, D. H., Ballou, J. D., Tonkyn, D. W., & Frankham, R. (2006). Realistic levels of inbreeding depression strongly affect extinction risk in wild populations. Biological Conservation, 133(1), 42-51. doi:10.1016/j.biocon.2006.05.016*
- O'Leary, S. J., Puritz, J. B., Willis, S. C., Hollenbeck, C. M., & Portnoy, D. S. (2018). These aren't the loci you're looking for: Principles of effective SNP filtering for molecular ecologists. Molecular Ecology, 27(16), 3193-3206.*
- Pacifci, M., Santini, L., Di Marco, M., Baisero, D., Francucci, L., Marasini, G. G., . . . Rondinini, C. (2013). Generation length for mammals. Nature Conservation, 5, 89.*
- Paetkau, D., Vazquez-Dominguez, E., Tucker, N. I. J., & Moritz, C. (2009). Monitoring movement into and through a newly planted rainforest corridor using genetic analysis of natal origin. Ecological Management & Restoration, 10(3), 210-216. doi:10.1111/j.1442-8903.2009.00490.x*
- Pavoine, S., Ollier, S., & Dufour, A. B. (2005). Is the originality of a species measurable? Ecology Letters, 8(6), 579-586.*
- Peck, S. L. (2004). Simulation as experiment: a philosophical reassessment for biological modeling. Trends in Ecology & Evolution, 19(10), 530-534.*
- Peñalba, J. V., & Wolf, J. B. W. (2020). From molecules to populations: appreciating and estimating recombination rate variation. Nature Reviews Genetics, 1-17.*
- Pew, J., Muir, P. H., Wang, J., & Frasier, T. R. (2015). related: an R package for analysing pairwise relatedness from codominant molecular markers. Molecular Ecology Resources, 15(3), 557-561.*
- Portanier, E., Garel, M., Devillard, S., Maillard, D., Poissant, J., Galan, M., . . . Itty, C. (2019). Both candidate gene and neutral genetic diversity correlate with parasite resistance in female Mediterranean mouflon. BMC Ecology, 19(1), 12.*

- Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., . . . Santpere, G. (2013). Great ape genetic diversity and population history. Nature, 499(7459), 471-475.*
- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics, 155(2), 945-959.*
- Privé, F., Aschard, H., Ziyatdinov, A., & Blum, M. G. B. (2018). Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. Bioinformatics, 34(16), 2781-2787.*
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., . . . Daly, M. J. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics, 81(3), 559-575.*
- R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria., URL <https://www.R-project.org/>.*
- Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W., & Cavalli-Sforza, L. L. (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proceedings of the National Academy of Sciences, 102(44), 15942-15947.*
- Rumball, W., Franklin, I. R., Frankham, R., & Sheldon, B. L. (1994). Decline in heterozygosity under full-sib and double first-cousin inbreeding in *Drosophila melanogaster*. Genetics, 136(3), 1039-1049.*
- SAC. (2020). Scientific Advisory Committee. Final Recommendation on a Nomination for Listing *Ornithorhynchus anatinus* Shaw 1799 - Platypus. File No.: FF/54/3795. Retrieved from East Melbourne VIC 8002: https://www.environment.vic.gov.au/_data/assets/pdf_file/0032/496346/03-FINAL-Platypus-FFR.pdf*
- Saccheri, I. J., Whiteford, S., Yung, C. J., & van't Hof, A. E. (2020). Recessive Z-linked lethals and the retention of haplotype diversity in a captive butterfly population. Heredity, 1-12.*

- Sainsbury, A. W., & Vaughan-Higgins, R. J. (2012). Analyzing disease risks associated with translocations. *Conservation Biology*, 26(3), 442-452.
- Sakharkar, M. K., Chow, V. T. K., & Kangueane, P. (2004). Distributions of exons and introns in the human genome. *In silico biology*, 4(4), 387-393.
- Salvador-Martínez, I., Coronado-Zamora, M., Castellano, D., Barbadilla, A., & Salazar-Ciudad, I. (2017). Mapping Selection within *Drosophila melanogaster* Embryo's Anatomy. *Molecular Biology and Evolution*, 35(1), 66-79. doi:10.1093/molbev/msx266
- Santos, J., Pascual, M., Simões, P., Fragata, I., Lima, M., Kellen, B., . . . Matos, M. (2012). From nature to the laboratory: the impact of founder effects on adaptation. *Journal of Evolutionary Biology*, 25(12), 2607-2622.
- Schaffner, S. F. (2004). The X chromosome in population genetics. *Nature Reviews Genetics*, 5(1), 43-51.
- Schmidt, T. L., Jasper, M., Weeks, A. R., & Hoffmann, A. A. (2020). Unbiased population heterozygosity estimates from genome-wide sequence data. *Biorxiv*.
- Schou, M. F., Loeschke, V., Bechsgaard, J., Schlötterer, C., & Kristensen, T. N. (2017). Unexpected high genetic diversity in small populations suggests maintenance by associative overdominance. *Molecular Ecology*, 26(23), 6510-6523.
- Seddon, P. J., Armstrong, D. P., & Maloney, R. F. (2007). Developing the science of reintroduction biology. *Conservation Biology*, 21(2), 303-312. doi:10.1111/j.1523-1739.2006.00627.x
- Sella, G., Petrov, D. A., Przeworski, M., & Andolfatto, P. (2009). Pervasive natural selection in the *Drosophila* genome? *PLoS Genet*, 5(6), e1000495.
- Semenov, G. A., Safran, R. J., Smith, C. C., Turbek, S. P., Mullen, S. P., & Flaxman, S. M. (2019). Unifying theoretical and empirical perspectives on genomic differentiation. *Trends in Ecology & Evolution*, 34(11), 987-995.
- Serena, M., & Williams, G. (1998). Rubber and plastic rubbish: a summary of the hazard posed to platypus *Ornithorhynchus anatinus* in suburban habitats. *Victorian Naturalist*, 115, 47-49.

- Serena, M., & Williams, G. (2013). Movements and cumulative range size of the platypus (*Ornithorhynchus anatinus*) inferred from mark–recapture studies. *Australian Journal of Zoology*, 60(5), 352-359.
- Servedio, M. R., Brandvain, Y., Dhole, S., Fitzpatrick, C. L., Goldberg, E. E., Stern, C. A., . . . Yeh, D. J. (2014). Not just a theory—the utility of mathematical models in evolutionary biology. *PLoS Biol*, 12(12), e1002017.
- Sherwin, W. B., Chao, A., Jost, L., & Smouse, P. E. (2017). Information theory broadens the spectrum of molecular ecology and evolution. *Trends in Ecology & Evolution*, 32(12), 948-963.
- Siewert, K. M., & Voight, B. F. (2017). Detecting Long-Term Balancing Selection Using Allele Frequency Correlation. *Molecular Biology and Evolution*, 34(11), 2996-3005. doi:10.1093/molbev/msx209
- Silva, A. T., Lucas, M. C., Castro-Santos, T., Katopodis, C., Baumgartner, L. J., Thiem, J. D., . . . Braun, D. C. (2018). The future of fish passage science, engineering, and practice. *Fish and Fisheries*, 19(2), 340-362.
- Silva, G. M. F., Andrade, M. C., Silva, B. R. M., Palheta, I. S., Gonçalves, L. B., Rocha, R. M., & Ferreira, M. A. P. (2020). Has a river dam affected the life-history traits of a freshwater prawn? *Ecology and Evolution*, 10(13), 6536-6548.
- Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, 139(1), 457-462.
- Smit, A. F. A., Hubley, R., & Green, P. (2015). RepeatMasker Open-4.0. Retrieved from <http://www.repeatmasker.org>
- Smith, J. M., & Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical Research*, 23(01), 23-35.
- Søndergaard, M., & Jeppesen, E. (2007). Anthropogenic impacts on lake and stream ecosystems, and approaches to restoration. In: Wiley Online Library.
- Steane, D. A., Nicolle, D., Sansaloni, C. P., Petroli, C. D., Carling, J., Kilian, A., . . . Vaillancourt, R. E. (2011). Population genetic analysis and phylogeny reconstruction in *Eucalyptus* (Myrtaceae) using high-throughput, genome-wide genotyping. *Molecular phylogenetics and evolution*, 59(1), 206-224.

- Stumpf, M. P. H., & McVean, G. A. T. (2003). Estimating recombination rates from population-genetic data. Nature Reviews Genetics, 4(12), 959-968.*
- Sun, H., Rowan, B. A., Flood, P. J., Brandt, R., Fuss, J., Hancock, A. M., . . . Schneeberger, K. (2019). Linked-read sequencing of gametes allows efficient genome-wide analysis of meiotic recombination. Nature communications, 10(1), 1-9.*
- Sunde, J., Yıldırım, Y., Tibblin, P., & Forsman, A. (2020). Comparing the performance of microsatellites and RADseq in population genetic studies: Analysis of data for pike (*Esox lucius*) and a synthesis of previous studies. Frontiers in genetics, 11, 218.*
- Szulkin, M., Bierne, N., & David, P. (2010). Heterozygosity-fitness correlations: a time for reappraisal. Evolution: International Journal of Organic Evolution, 64(5), 1202-1217.*
- Szulkin, M., Stopher, K. V., Pemberton, J. M., & Reid, J. M. (2013). Inbreeding avoidance, tolerance, or preference in animals? Trends in Ecology & Evolution, 28(4), 205-211.*
- Takahata, N. (1983). Gene identity and genetic differentiation of populations in the finite island model. Genetics, 104(3), 497-512.*
- Thurmond, J., Goodman, J. L., Strelets, V. B., Attrill, H., Gramates, L. S., Marygold, S. J., . . . Trovisco, V. (2018). FlyBase 2.0: the next generation. Nucleic Acids Research, 47(D1), D759-D765.*
- Thurmond, J., Goodman, J. L., Strelets, V. B., Attrill, H., Gramates, L. S., Marygold, S. J., . . . Trovisco, V. (2019). FlyBase 2.0: the next generation. Nucleic Acids Research, 47(D1), D759-D765.*
- Tigano, A., & Friesen, V. L. (2016). Genomics of local adaptation with gene flow. Molecular Ecology, 25(10), 2144-2164.*
- Tyers, M. (2017). riverdist: River network distance computation and applications. R package Version 0.15. 0.*
- Veyrunes, F., Waters, P. D., Miethke, P., Rens, W., McMillan, D., Alsop, A. E., . . . Schatzkamer, K. (2008). Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. Genome research, 18(6), 965-973.*

- Vicoso, B., & Charlesworth, B. (2006). Evolution on the X chromosome: unusual patterns and processes. *Nature Reviews Genetics*, 7(8), 645-653.
- Wallace, B. (1975). Hard and soft selection revisited. *Evolution*, 29(3), 465-473.
- Wang, J., Hill, W. G., Charlesworth, D., & Charlesworth, B. (1999). Dynamics of inbreeding depression due to deleterious mutations in small populations: mutation parameters and inbreeding rate. *Genetics Research*, 74(2), 165-178.
- Wang, J., & Whitlock, M. C. (2003). Estimating effective population size and migration rates from genetic samples over space and time. *Genetics*, 163(1), 429-446.
- Waples, R. S., & Do, C. (2008). LDNE: a program for estimating effective population size from data on linkage disequilibrium. *Molecular Ecology Resources*, 8(4), 753-756.
- Waples, R. S., Elz, A., Arnsberg, B. D., Faulkner, J. R., Hard, J. J., Timmins-Schiffman, E., & Park, L. K. (2017). Human-mediated evolution in a threatened species? Juvenile life-history changes in Snake River salmon. *Evolutionary Applications*, 10(7), 667-681.
- Warren, W. C., Hillier, L. W., Graves, J. A. M., Birney, E., Ponting, C. P., Grützner, F., . . . Chinwalla, A. T. (2008). Genome analysis of the platypus reveals unique signatures of evolution. *Nature*, 453(7192), 175-183.
- Watson, D. (2013). Contouring: a guide to the analysis and display of spatial data: Elsevier.
- Watson, D. F., & Philip, G. M. (1985). A refinement of inverse distance weighted interpolation. *Geo-Processing*, 2(4), 315-327.
- Weinberg, W. (1908). Über den nachweis der vererbung beim menschen. *Jh. Ver. vaterl. Naturk. Wurttemb.*, 64, 369-382.
- Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, 38(6), 1358-1370.
- Whitlock, M. C. (1992). Temporal fluctuations in demographic parameters and the genetic variance among populations. *Evolution*, 46(3), 608-615.
- Whitlock, M. C., & Lotterhos, K. E. (2015). Reliable detection of loci responsible for local adaptation: Inference of a null model through trimming the distribution of F ST. *The American Naturalist*, 186(S1), S24-S36.

- Willi, Y., Van Buskirk, J., & Hoffmann, A. A. (2006). *Limits to the adaptive potential of small populations*. In *Annual Review of Ecology Evolution and Systematics* (Vol. 37, pp. 433-458). Palo Alto: Annual Reviews.
- Williams, J. L., Hall, S. J. G., Del Corvo, M., Ballingall, K. T., Colli, L., Ajmone Marsan, P., & Biscarini, F. (2016). *Inbreeding and purging at the genomic Level: the Chillingham cattle reveal extensive, non-random SNP heterozygosity*. *Animal genetics*, 47(1), 19-27.
- Woinarski, J., & Burbidge, A. (2016). *Ornithorhynchus anatinus*. In *The IUCN Red List of Threatened Species 2016: e. T40488A21964009*: IUCN.
- Wood, S. N., Goude, Y., & Shaw, S. (2015). *Generalized additive models for large data sets*. *Journal of the Royal Statistical Society: Series C: Applied Statistics*, 139-155.
- Wright, S. (1931). *Evolution in Mendelian populations*. *Genetics*, 16, 97-159.
- Wright, S. (1943). *Isolation by distance*. *Genetics*, 28, 114-137.
- Xie, X., Yang, Y., Ren, Q., Ding, X., Bao, P., Yan, B., . . . Qiu, Q. (2018). *Accumulation of deleterious mutations in the domestic yak genome*. *Animal genetics*, 49(5), 384-392.
- Yandell, M., & Ence, D. (2012). *A beginner's guide to eukaryotic genome annotation*. *Nature Reviews Genetics*, 13(5), 329-342.
- Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). *GCTA: a tool for genome-wide complex trait analysis*. *The American Journal of Human Genetics*, 88(1), 76-82.
- Yates, A. D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., . . . Bennett, R. (2020). *Ensembl 2020*. *Nucleic Acids Research*, 48(D1), D682-D688.
- Yıldırım, Y., Tinnert, J., & Forsman, A. (2018). *Contrasting patterns of neutral and functional genetic diversity in stable and disturbed environments*. *Ecology and Evolution*, 8(23), 12073-12089.
- Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., . . . Girón, C. G. (2017). *Ensembl 2018*. *Nucleic Acids Research*, 46(D1), D754-D761.
- Zhao, L., & Charlesworth, B. (2016). *Resolving the conflict between associative overdominance and background selection*. *Genetics*, 203(3), 1315-1334.

