UNIVERSITY OF SOUTHAMPTON

Exploring models in population biology through the simulation of species invasions, natural selection and market-mediated gene flow

by

Guy Sherwin Jacobs

Supervised by

T. J. Sluckin, T. Kivisild and J. Noble

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
Faculty of Social, Human and Mathematical Sciences
Mathematical Sciences

December 2015

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL, HUMAN AND MATHEMATICAL SCIENCES
MATHEMATICAL SCIENCES

Doctor of Philosophy

Exploring models in population biology through the simulation of species invasions, natural selection and market-mediated gene flow

by Guy Sherwin Jacobs

Supervised by

T. J. Sluckin, T. Kivisild and J. Noble

In this thesis, I apply simulation techniques to investigate three questions in population biology, which focus on movement and natural selection. The first model assesses the theoretical implications of long-range dispersal in species invasions, identifying an important interaction between the representation of a finite population and the rate of population spread. The second investigates the genetic impact of movement distortions among domestic animals due to human economic activity, suggesting that the marketing of animals could fundamentally impact their genetic variation and distribution. My third model considers the problem of detecting evidence of positive natural selection in the genome, refining and testing statistics designed to identify which genes have offered a reproductive advantage in the past using population genetic data. These three simulation studies use very different approaches, and, separately, identify the critical and practical importance of assumptions frequently encountered in population models. Such assumptions - infinite population size, unbiased migration, and constant recombination rate - each lead to interesting properties of model behaviour, and may be relevant to interpretation and prediction in real world problems.

Declaration of Authorship

I, **Guy Sherwin Jacobs**, declare that the thesis entitled **Exploring models in population biology through the simulation of species invasions, natural selection and market-mediated gene flow** and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of this work have been published as [1] and [2]

Signed: ....

Date

Contents

Declaration of Authorship i

Acknowledgements xiii

1 Concepts in modelling and simulation 1

1.1 Modelling in mind and science . . . 2

1.1.1 Modelling in human reasoning . . . 3

1.1.2 Modelling in scientific reasoning . . . 3

1.1.3 How do models provide information about a target system? . . . 4

1.2 Modelling in population biology: some debates . . . 5

1.2.1 A Defence . . . 5

1.2.2 The Strategy . . . 8

1.3 Complexity, tractability and computational simulation . . . 10

1.3.1 Computational simulation . . . 11

1.3.2 Complex simulations of ecological complexity . . . 12

1.3.3 Simple simulations of mathematical simplification . . . 13

1.4 Outline of thesis and main results . . . 15

1.4.1 Chapter 2 - Modelling species invasions . . . 15

1.4.2 Chapter 3 - Modelling animal markets . . . 16

1.4.3 Chapter 4 - Modelling natural selection . . . 17

2 Modelling species invasions: stochastic long-range dispersal and invasion acceleration 19

2.1 Chapter introduction and summary . . . 19

2.2 Long-range dispersal, stochasticity and the broken accelerating wave of advance *(Published manuscript)* . . . 21

2.2.1 Abstract . . . 21

2.2.2 Introduction . . . 22

2.2.3 Simulation modelling . . . 28

2.2.4 Results . . . 35

2.2.4.1 Testing model behaviour . . . 35

2.2.4.2 Model 1 with $N = 1$: Constant velocity waves in a stochastic lattice system, with a notable exception . . . 37

2.2.4.3 Model 2: Acceleration for fat-tailed kernel invasions in a mean-field system . . . 41

2.2.4.4 Model 1, $N > 1$: Reducing dispersal stochasticity leads to slow filling time convergence with long-range kernels 42
2.2.4.5 Truncated power law kernels lead to asymptotically constant wave velocities 45
2.2.4.6 Structural variations of our model support the generality of results 46
2.2.5 Discussion 49
2.2.6 Appendix 1: Simple derivation of a diffusive limit for mean-field Model 2 58
2.2.7 Appendix 2: Periodic boundary conditions 58
2.2.8 Appendix 3: Marginal stability analysis of Model 2 with a nearest-neighbour kernel 59
2.2.9 Appendix 4: Clarifying kernel behaviour 59
2.2.10 Appendix 5: Estimating maximum and minimum wave velocities for the constant velocity waves in Model 1 63
2.2.11 Appendix 6: Finite size scaling of a stochastic system with a bivariate Student's t dispersion kernel 65
2.3 Comments on the manuscript 67
2.4 Representing long-range movements using the dispersal kernel 68
2.4.1 Dispersal and movement 68
2.4.2 Power laws in dispersal data 69
2.4.3 The generative processes of Lévy flights 70
2.4.4 Stretched exponential dispersal kernels and their generative processes 71
2.4.5 From generative process to model 72
2.5 The role of simulation in testing mathematical models 73

3 Modelling animal markets: distributions of species and genes under dynamically biased migration 75
3.1 Chapter introduction and summary 75
3.2 The market for domestic animals 76
3.2.1 The contemporary trade in animal genetic material 76
3.2.2 Literature and linguistics on the ancient livestock trade 77
3.2.3 Domestic animal movements through bones and molecules 79
3.2.4 An overview of movement models 81
3.2.5 Human versus animal migration and the random walk 83
3.2.6 Representing movement biases in models 84
3.3 On the counter-evolutionary effects of market mediated gene flow *(Unpublished manuscript)* 87
3.3.1 Abstract 87
3.3.2 Introduction 88
3.3.2.1 The static and dynamic market filters 89
3.3.3 Methods 90
3.3.3.1 Model design 90
3.3.3.2 The market algorithm 91

3.3.3.3 Simulation modelling . . . 99
3.3.4 Results . . . 100
3.3.4.1 Static Market Filters . . . 100
3.3.4.2 Dynamic Market Filters . . . 103
3.3.4.3 Market mortality and initial conditions . . . 107
3.3.5 Discussion . . . 108
3.3.6 Appendix 1: Auctions and bidding order . . . 116
3.3.7 Appendix 2: The Beverton-Holt model of population regulation . 119
3.3.8 Appendix 3: A static market working with or against selection . 128
3.3.9 Appendix 4: Robustness to initial conditions . . . 131
3.4 Some mathematical comments on the manuscript . . . 134
3.5 Representing the market . . . 138
3.5.1 The structure of our model . . . 138
3.5.2 Some alternative representations of the market . . . 139
3.5.3 The trading network . . . 141
3.6 Conceptual and experimental extensions of the market genetics framework 142
3.6.1 The Market for Lemons and the emergence of breeds . . . 142
3.6.2 Social stratification and the emergence of breeds . . . 142
3.7 Simulations and modelling . . . 143

4 Modelling natural selection: improving statistics to detect selection in population genetic data 145
4.1 Chapter introduction and summary . . . 145
4.2 Signatures of natural selection . . . 147
4.2.1 The concepts of genetic linkage and recombination . . . 147
4.2.2 Measuring linkage disequilibrium . . . 148
4.2.3 Genetic hitchhiking and the interaction between linkage and selection . . . 149
4.2.4 Some signals used to infer selection . . . 150
4.3 Challenges in identifying natural selection . . . 153
4.3.1 Natural selection affecting allele frequencies in humans . . . 154
4.3.2 From differential reproduction to allele frequency changes . . . 155
4.4 Statistical details of genome-wide scans for natural selection . . . 156
4.4.1 The standard model of a genome-wide selection scan . . . 156
4.4.2 Important concepts in GWSSs . . . 157
4.4.3 H_0 and the concept of neutrality . . . 159
4.5 Using simulations to improve test statistics for GWSSs . . . 162
4.5.1 The coalescent simulation method . . . 163
4.6 Refining the use of linkage disequilibrium as a robust signature of selection *(Unpublished manuscript)* . . . 169
4.6.1 Abstract . . . 169
4.6.2 Introduction . . . 170
4.6.3 Selection statistics based on LD . . . 171
4.6.3.1 Challenges faced by Kelly's Z_{nS} and the ω statistc . . . 173
4.6.4 Methods . . . 175

4.6.4.1 Designing selection statistics 175
4.6.4.2 Assessing the power of statistics through coalescent simulation 176
4.6.4.3 Replication of previously suggested selection candidates 177
4.6.5 Results 178
4.6.5.1 Controlling for expected LD increases simulated power of statistics when recombination rate is variable 178
4.6.5.2 Controlling for expected LD increases selection candidate signal replication in HapMap data 181
4.6.5.3 Selection candidate signal replication is not increased for candidate windows containing protein-coding genes . . 185
4.6.6 Discussion 186
4.6.7 Conclusion 189
4.6.8 Extended Methods 190
4.6.9 Appendix 1: Detailed simulation performance results for tested statistics 197
4.7 The role of window size in determining expected average LD 197
4.7.1 Appendix 3: Selection candidate windows 200
4.7.2 Appendix 4: Further selection scans and replication data 200
4.7.3 Appendix 5: Top 1% hits based on selected statistics 200
4.8 Assessing test statistic performance through coalescent simulations . . . 200
4.8.1 Robustness and flexibility of the coalescent 200
4.8.2 Coalescent models and human evolution 203
4.9 Simulation as a tool to improve complex test statistics 204

5 Discussion and conclusions **207**
5.1 Assessing the robustness of my modelling results 208
5.2 The utility of models when robustness is uncertain 209
5.3 Equifinality of models 212
5.4 Closing remarks and conclusion 213

Bibliography **215**

Appendices **253**
A1 Thesis Appendix 1: Presentation of work at conferences 253
A2 Thesis Appendix 2: Simple derivation of a diffusive limit for mean-field Model 2 255
A3 Thesis Appendix 3: Marginal stability analysis of Model 2 with a nearest-neighbour kernel 257
A4 Thesis Appendix 4: Selection candidate windows 259
A5 Thesis Appendix 5: Further selection scans and replication data 262
A6 Thesis Appendix 6: Top 1% hits based on selected statistics 271
A7 Thesis Appendix 7: Glossary of population genetics terms 275

List of Figures

2.1 Schematic of species invasion simulation algorithm 28

2.2 Testing invasion model behaviour 36

2.3 Dependence of filling time on system size in the stochastic and mean-field species invasion models 37

2.4 Dependence of front-velocity on stretched exponential power a the stochastic species invasion model 39

2.5 Placing bounds on velocity in the stochastic species invasion model . . . 40

2.6 The approach of stochastic species invasion model behaviour to mean field model behaviour as N is increased 42

2.7 Visualisation of modelled species invasion dynamics 44

2.8 Convergence of the stochastic species invasion model to the mean-field limit as $N \gg 1$ 44

2.9 Wave velocity behaviour given a truncated power-law dispersal kernel . 45

2.10 The impact of stocasticty on three variations of our species invasion model 47

2.11 The role of the detailed implementation of the logistic effect in the behaviour of the stochastic species invasion model 48

2.12 Filling time of the stochastic species invasion model given periodic boundary conditions 58

2.13 Relationships between diffusivity, D, and asymptotically constant wave velocity, c, under various dispersal kernels 60

2.14 Simulated and analytically estimated diffusion constants for power-law dispersal kernels 61

2.15 Deriving wave acceleration and velocity behaviour given kernel parameterisation using linear and non-linear regression. 63

2.16 Comparison of a stochastic species invasion model behaviour given a bivariate Student's t kernel and a power-law kernel with $\beta = 3$ 66

3.1 Schematic of a Static Filter market model 93

3.2 Schematic of a Dynamic Filter market model 97

3.3 Equilibrium patch occupation given a Static Filter 1 market model and selection against Breed B in both patched 100

3.4 Equilibrium occupation given Static Filters 1–4 as the market model and selection against Breed B in both patches 101

3.5 Temporal evolution of the system given a Static Filter 1 or Static Filter 3 market model 103

3.6 Equilibrium occupation given a Dynamic Filter 1 market model, with selection against Breed B in both patches 104

3.7 Equilibrium occupation given a Dynamic Filter 2 market model, with opposite selection regimes in the two patches 105

3.8 Temporal evolution of the system give a Dynamic Filter 1 or Dynamic Filter 2 market model 106

3.9 Impact of net trade imbalance on mortality in the system, given a market represented by Static Markets 1 or 3 or a Dynamic Market 107

3.10 Equilibrium occupation given a market represented by Dynamic Filters 1 or 2, using different approaches to calculating migration rates 118

3.11 Average stochastic bid-order behaviour as compared to piecewise linear and quadratic approximations 120

3.12 Equilibrium occupation given Static Filter market models and selection against Breed B, using a Beverton-Holt growth function ($R_0 = 2$) 121

3.13 Equilibrium occupation given Static Filter market models and selection against Breed B, using a Beverton-Holt growth function ($R_0 = 1.25$) . . 122

3.14 Equilibrium mortality in the system caused by trade-imbalance and patch-specific population size given Static Filter market models and a Beverton-Holt growth function ($R_0 = 1.5$) 123

3.15 Equilibrium occupation given a Dynamic Filter 1 market model and selection against Breed B, using a Beverton-Holt growth function ($R_0 = 2$) 123

3.16 Equilibrium population size given a Dynamic Filter 1 market model and selection against Breed B, using a Beverton-Holt growth function ($R_0 = 2.0$)124

3.17 Equilibrium occupation given a Dynamic Filter 1 market model and selection against Breed B, using a Beverton-Holt growth function ($R_0 = 1.25$)124

3.18 Equilibrium population size given a Dynamic Filter 1 market model and selection against Breed B, using a Beverton-Holt growth function ($R_0 = 1.25$) 125

3.19 Equilibrium occupation given a Dynamic Filter 2 market model and opposite selection in the two patches, using a Beverton-Holt growth function ($R_0 = 2$) 125

3.20 Equilibrium population size given a Dynamic Filter 2 market model and opposite selection in the two patches, using a Beverton-Holt growth function ($R_0 = 2$) 126

3.21 Equilibrium occupation given a Dynamic Filter 2 market model and opposite selection in the two patches, using a Beverton-Holt growth function ($R_0 = 1.25$) 126

3.22 Equilibrium population size given a Dynamic Filter 2 market model and opposite selection in the two patches, using a Beverton-Holt growth function ($R_0 = 1.25$) 127

3.23 Equilibrium occupation given a migration filter that promotes the movement of breeds to their selectively favoured habitats, with opposing selection regimes in the two patches 129

3.24 Temporal evolution of the system given Static Filter 3 promoting the movement of Breed A animals to their selectively favoured Patch 0 only, with opposing selection regimes in the two patches 130

3.25 Equilibrium occupation given a migration filter that acts against the movement of breeds to their selectively favoured habitats, with opposing selection regimes in the two patches 130

3.26 Equilibrium occupation given a Dynamic Filter 1 market model and selection against Breed B, with initial conditions $[A_0(0), A_1(0)] = [0.99, 0.99]$ 131

3.27 Equilibrium occupation given a Dynamic Filter 1 market model and selection against Breed B, with initial conditions $[A_0(0), A_1(0)] = [0.01, 0.01]$ 132

3.28 Equilibrium occupation given a Dynamic Filter 1 market model and selection against Breed B, with initial conditions $[A_0(0), A_1(0)] = [0.01, 0.99]$ 132

3.29 Equilibrium occupation given a Dynamic Filter 1 market model and selection against Breed B, with initial conditions $[A_0(0), A_1(0)] = [0.99, 0.01]$ 133

3.30 The Edgeworth Box 140

4.1 Schematic diagram of meiotic recombination 148

4.2 Linkage disequilibrium in a northeast Siberian sample around a putatively selected SNP in the *CPT1A* gene 152

4.3 The coalescent under a Wright-Fisher population model 165

4.4 Adding mutations to a coalescent graph, with and without recombination 166

4.5 The average pattern of pairwise linkage disequilibrium and SNP diversity at different stages of a selective sweep 171

4.6 Schamtic illustration of an LD-based selection statistic calculation . . . 176

4.7 Performance of different categories of LD-based selection statistics according to the pAUC metric 179

4.8 The power of variants of the α selection statistic given different approaches to controlling for variable recombination rate 180

4.9 Standardised selection scans on HapMap Phase II data using representative selection statistics 182

4.10 Schematic of the Out of Africa model, showing possible selection phases 195

4.11 Schematic of the step-decay in LD described in the text, and two example statistic window size constructions leading to different expected values of ω. The horizontal bar (here green) indicates the chromosome, with the yellow and purple triangle representing the matrix of expected pairwise LD values between loci along the chromosome. Yellow shows positive LD between SNPs separated by a distance y_c or less, while purple indicates no LD. When the window size is x_0, the value of ω is expected to be 1 in this simple illustrative scenario; when the window size is $x_1 = y_c$, ω is expected to be 2. The 'Along' and 'Over' regions are indicated as α and β respectively. 199

A1 Standardised selection scan using the $\dfrac{\alpha}{E[\alpha]}$ statistic, using the deCODE and HapMap genetic maps 263

A2 Standardised selection scans for statistics based on LD in the Along region that may control for expected LD 265

A3 Standardised selection scans for statistics based on LD in the Over region that may control for expected LD 266

A4 Standardised selection scan using the OmegaPlus program with different window sizes 267

A5 Standardised selection scan using the OmegaPlus program or a similar statistic that controls for expected LD 268

List of Tables

2.1 Approximate wave velocity behaviour given various dispersal regimes and either a stochastic or mean-field model of species invasion 38

2.2 Relationship between wave velocity and diffusion constant for mean-field Model 2 given a stretched exponential dispersal kernel, $K(\frac{l}{\varphi}) = e^{-|\frac{l}{\varphi}|^{\gamma}}$, $l \in \mathbb{Z}$ 60

4.1 The power of representative selection statistics given different recombination rate models 179

4.2 The power of representative selection statistics given an Out of Africa demographic model 180

4.3 The ability of representative selection statistics to replicate previously suggested selection signals in HapMap Phase II data 183

4.4 Previously suggested selection candidate windows replicated by representative test statistics at an empirical p-value of 0.01 184

4.5 Comparison of representative selection statistic results in HapMap Phase II selection candidate windows that either do or do not contain protein coding genes 185

4.6 Example MSMS command lines given representative demographic and selection scenarios 194

4.7 Selection scan studies used to define the candidate selection signal set. . 196

4.8 Complete pAUC performance metric results for all selection statistics tested198

A1 Previously suggested selection candidates used for the CEU population, Chromosome 2 260

A2 Previously suggested selection candidates used for the CEU population, Chromosome 15 261

A3 Previously suggested selection candidates used for the CHB+JPT population, Chromosome 2 261

A4 Replication of previously suggested selection candidates in HapMap Phase II data when controlling for expected LD using different genetic maps . 262

A5 Replication of previously suggested selection candidates, using statistics that focus on the Along region and may control for expected LD 264

A6 Replication of previously suggested selection candidates, using statistics that focuss on the Over region and may control for expected LD 264

A7 Replication of previously suggested selection candidates by the OmegaPlus program using different window sizes, and a variant that controls for expected LD . . . 267

A8 Details of empirical p-values for representative selection statistics in previously identified selection candidate windows using HapMap Phase II data . . . 269

A9 Selection candidates identified by representative selection statistics in this study for HapMap Phase II CEU Chromosome 2 data . . . 272

A10 Selection candidates identified by representative selection statistics in this study for HapMap Phase II CEU Chromosome 15 data . . . 273

A11 Selection candidates identified by representative selection statistics in this study for HapMap Phase II CHB+JPT Chromosome 2 data . . . 274

Acknowledgements

This thesis could not have been produced without the support and guidance of many people.

Firstly, my parents and family, for everything. Words are insufficient.

The many friends that Southampton has brought my way. The iterations of the Padwell Road crew - for making that happen - and Olli especially, always there and so many shared experiences - dark them woods! My ICSS friends, in particular Adam, Rich, Joe and Paul, for humour, debauchery and wicker-based activities. And the many other people in my life who have listened to, inspired and offered friendship. Jiho and Nick, foundations.

To Susie. Such a wonderful, genuine person, for all her support, and for being there. Thank you.

To Mircea and Niraj - an Indian past and an Indian future, together!

The ICSS and Seth, Jason and Hans, for providing me with the opportunity to pursue this PhD, as well as the EPSRC for funding.

Gereon, a friend of ideas (in both senses), for so many stimulating discussions, and for support, kindness and excellent comradeship through our PhDs.

Marta, for her enthusiasm, humanity and inspiration, a light on this path for many years.

Finally, my supervisors - Tim, Toomas and Jason. Jason, for his early role in showing the ICSS was the right place to join, and for providing a warming example of the right way to do (and leave!) academia. Toomas, a true scientist and a person, for his easy generosity and openness, and for sharing his clarity of thought and enthusiasm. Finally, Tim, for his endless patience, his diversity of mind and engaging way of being, who has provided insight, guidance and friendship over these four years. I look back to his ready acceptance of the aimless biological anthropology student with a 2.2 grade in MATH3052, with a smile. To many years of continued friendship!

Academically, several collaborators have contributed to the manuscripts presented in this thesis:

Tim Sluckin suggested the initial idea of modelling species invasions, §2.2. I coded and ran all simulations and conducted all data analysis, with considerable guidance from Tim in the initial stages of the project especially. I wrote the manuscript, with extensive comments from Tim over several iterations on the road to publication. Tim was involved in many stimulating discussions, and provided calculations and insight for sections §2.2.6, §2.2.8 and §2.2.9 especially.

I suggested the initial idea of modelling gene flow through markets, §3.3. I coded and ran all the simulations, with suggestions from Tim Sluckin and Gereon Kaiping, a fellow PhD student supervised by Simon Cox and Tim, and conducted all data analysis. I wrote the manuscript, with comments (especially on the presentation of methods) from Tim and Gereon. Both Tim and Gereon have been involved in many discussions concerning this topic. Early discussions with Marta Lahr were important to the development ideas related to cattle and markets, and while she is not currently included as a co-author on the manuscript - this work was conducted entirely by myself, Gereon and Tim - I hope to revisit and collaborate on this project with her soon.

I suggested the initial idea of exploring signatures of selection in genetic data, §4.6. I designed the command lines for two external coalescent simulators (MSMS [3] and Cosi2 [4]), and coded programs to interpret and perform statistical tests/selection scans on output. I conducted all data analysis. Toomas Kivisild especially provided comments and guidance throughout the project, and Tim Sluckin was also involved in discussion of different selection statistic methods. I wrote the manuscript, with comments from both Toomas and Tim.

Other collaborations that may feature in, but are not core to, this thesis:

I calculated and visualised LD patterns for several population genetic samples for the paper of Clemente *et al* [2], and offered guidance in interpreting the findings (see §4.2.4).

I was involved in discussion and model design, and offered some comments of the manuscript, of Kaiping *et al* [5].

I was involved in discussion and in performing coalescent simulations of Tajima's D signal sharing between populations in Pagani *et al* [6].

During the PhD, I also provided considerable supervision to Susie Martin, an undergraduate at Southampton, on a project searching for selection signals in genomic regions showing phenotypic associations. She received a First for this project, and it may lead to publication

A list of conference contributions is included as Thesis Appendix A1.

Chapter 1

Concepts in modelling and simulation

Mathematical models have played a fundamental role in the recent history of population biology. Early examples include descriptions of population growth [7, 8], predator-prey dynamics [9], species dispersal [10, 11], and genetic evolution (e.g. [12, 13, 14]; and prior work). All of these, with the exception of population growth, were formalised in the early 20th century, and all remain active topics of research one hundred years later. My thesis, as an exploration of models and modelling in population biology through simulation, contributes to this tradition. Through three quite different examples - species invasions, genomic signatures of natural selection, and market-mediated gene flow - and three quite different simulation approaches, I will argue that simulations offer a tool with which to flexibly tease apart the implications of modelling assumptions. Through this process, we are able to gain insight into the appropriate use of different models and the generality of their results.

The core research presented here consists of three largely independent studies, which are united in using simulation to explore the role of migration and selection in models of population systems. By taking different approaches to different topical research questions, I aim to draw practical conclusions about the role of simulation models in population biology. Although models have long guided our theoretical understanding of populations, the increasingly widespread availability of computational power allows for the exploration of more complex models that better represent the many interacting, discrete, heterogeneous agents of a population. Many different representations of a system might be chosen. A strong message of my thesis is that these choices can substantially impact model behaviour, and that simple simulation offers a useful tool through which to investigate the importance of different choices.

The challenges in designing [15] models, testing or assessing them [16], and, especially, applying their results to our understanding of the world with appropriate caution

[16, 17, 18] are subjects of continued debate. Given that the role and nature of modelling is not uncontroversial, to draw valuable conclusions form my work about modelling as a process it is is useful to have a clear understanding of what modelling and computational simulation are. This introduction, then, begins with a brief discussion of the theoretical and philosophical background of modelling as an approach in population biology. I then turn to the advantages and disadvantages of different modelling approaches. Readers interested more in the practical role of modelling may prefer to skip to my discussion of modelling in population biology, 1.2.

1.1 Modelling in mind and science

There is a considerable philosophical literature on what models are and how they should be used ([19] and references therein). To summarise even the core historical arguments would rapidly become cumbersome and detract from the essential point of this thesis. I do not propose to rigorously re-analyse the theory and philosophy of model-based science. Details of whether models are fictional [20] or non-fictional [21] entities, or if they are set-theoretic structures [22], are not critical. Rather, I am instead interested in what scientists use models for, how they construct them, and the limitations different types of model may face.

To guide this discussion, it is useful to have a working definition of a model. Drawing on [23, 24, 25, 26], I consider a model to be a modeller's simplified representation of a target entity. This is slightly broader than Maria [25], in that it does not make claims about which features of the system should be represented, and perhaps narrower than Wartofsky [23] who emphasises representation over simplicity. The requirement of simplification is not obviously compatible with ideas of models as isomorphic with the entity they represent (e.g.[27]) - though this may be so if the model is a simplification only from the perspective of the person who designed it or it's user. That a model is of something (be it a biological system, or some data, or a physical object) and provides a representation for someone, who is termed the modeller here without necessitating that they designed the model, is taken at face value.

Our definition makes no statement on the general purpose of modelling, or its specific role in scientific practice. To help clarify this, it is instructive to consider the broader role of models in human reasoning.

1.1.1 Modelling in human reasoning

Under the above definition, it seems likely that models are ubiquitous in human thought. One idea incorporating this is the 'mental model', proposed explicitly by Craik [28] but predicted by Peirce and others [29]. In Craik's formulation, the process of reasoning consists of three steps - mental 'translation' of the external world into some symbolic internal representation, manipulation of this internal representation to arrive at other symbols, and the 'retranslation' of these symbols into their implications for the external world [28, pp50–51]. The internal construction and manipulation of a model yields predictions about the external world. This interpretation of human reasoning has been contrasted (e.g. [30]) to that of formal mental logic (e.g. [31]), while both have been criticised for their difficulties in reconciling partially contradictory beliefs [32]. Whichever conception of human thought is preferred, however, most practical descriptions of our experience and ineraction with the world involve some process of representation.

The use of internal modelling, then, is to conceptualise and reason about the world. In the context of mental models, Craik notes that the approach - translation, manipulation, re-translation - allows us to observe a 'final result similar to that which might have been reached by causing the actual physical process to occur' [28, Chapter 5]. There is some debate about what it is in a model that allows us to retrieve valid conclusions about the target system. This may rely on the entities in the model relating to one another in to a way that reflects reality [33], although Craik did not demand this [29]. As the ability of different representations and simplifications in models to yield valid conclusions is relevant to this thesis, I return to this point with a focus on scientific models shortly.

1.1.2 Modelling in scientific reasoning

If human reasoning involves modelling, then it is inevitable that the action of scientific investigation also involves modelling in a weak sense. The use of modelling in science, however, is more formalised. Boltzmann argued, in the context of the physical sciences, that the task of theory is 'constructing an image of the external world' through 'carrying out globally what on a small scale occurs within us whenever we form an idea' [34, 29]. Parallels between scientific and internal modelling were also noted by other 19th century physicists - Kelvin and Maxwell, for example [29]. This resemblance extends to the process of constructing and using scientific models. Hughes [35] advocates an approach in physics whereby a system is denoted through representation to create a model, with the implications of that model then demonstrated through manipulation before relating these back to the target system through interpretation. The similarity between this approach and that suggested by Craik in the context of mental models, outlined above, is clear.

Given this blueprint for how a model is used in science and reasoning, several immediate questions arise. Assuming that a target system has been chosen, it is not immediately clear how to represent it and how to simplify it. To understand how these decisions are made, it is useful to recognise that the modeller is motivated to learn about a target system. Thus, before turning to the debate on complexity and modelling in biology, I will consider how it is that manipulations of a simplified representation of a system can yield valid conclusions about that system.

1.1.3 How do models provide information about a target system?

That models should be useful at all is not obvious. Representation is often thought of as subjective (e.g. [36, 26]), such that almost any entity can be used to represent (e.g. [33]) or denote [35] any other entity. The implication is that the vast majority of possible simplified representations of systems are very poor, such that any stated correspondence between model and system is either so complex as to obscure understanding or so trivial as to be worthless as an intellectual tool. In response to this problem, further constraints on the types of representation used in modelling are often proposed - ranging from isomorphism to similarity [19]. Strict isomorphism is a rigorous, perhaps impossible, demand, and only useful when it allows a system to be translated into version that is simpler for the modeller to manipulate. Similarity is difficult to define, and the burden of assessing which similarities are important remains with the modeller [26]. An intermediate position is structural representation [33], such that 'the patterns of relations among the constituents of the represented phenomenon is mirrored among the constituents of the representation itself'. This allows the model to be used for 'surrogate reasoning', and appeals to the idea of manipulations of a model reflecting manipulations of the target entity.

And yet 'black boxes' are common in biological modelling, such that models will incorporate different entities to the systems they represent and emphasise some patterns of relationship over others [37]. It is also difficult to reconcile hypothesis testing in statistics with structural representation. The 2 sample t-test might be considered a model in which two samples are normally distributed with the same mean. By applying this model to data and using well understood properties of the model, we can characterise how improbable the data would be under it, and hence learn about the data. The reason that the model is useful is not that it captures a pattern of relations in the data, but because it allows us to constrain the range of possible relations in the data. Another problematic example from statistics is the curve-fitting problem [19], where data are summarised by a curve of a certain form. This may offer useful insight while having an undefined relationship with either the generative process of the data or individual data points.

Given the diversity of models and applications, it is pragmatic to follow Teller [36], who considered the appropriate degree of similarity between model and target system to depend on scientific context. It is likely that novel information about a target system arises through different processes under different circumstances. Partial structural representation may be important at some times, external information - such as knowledge about how to manipulate certain types of model, or the specific context for conducting a statistical procedure - at others.

1.2 Modelling in population biology: some debates

The challenge of modelling in population biology is often characterised as *complexity*. More specifically, we tend to have a large (far greater than one but far less than infinite) number of interacting organisms. Each of these has unique phenotypic characteristics determined by internal biological processes, which are partially heritable but directed by an individual's mental and physical developmental process and environmental context. The types of interaction between organisms are many and varied, and there is extensive structure in the arrangement of these interactions. For example, intra- and interspecific behaviours differ, geographical and temporal proximity have a critical impact on behaviour, and features such as kinship can also play a role. There are, furthermore, abiotic stimuli for each organism, these being heterogeneous in time and space. This intimidating list is not exhaustive.

How might we model such a system to obtain useful insight into it? To begin to answer this question, and to provide some historical context of the difficulties involved, I will introduce two debates on modelling in population biology. The first is J.B.S Haldane's *A defence of beanbag genetics* [38], a response to criticism of mathematical population genetics. This will guide a discussion of the role of mathematics in modelling complex systems. The second is that initiated by Richard Levins' classic paper *The strategy of model building in population biology* [15]. Here, the focus is on model design and the robustness of modelling results. These three topics - the advantages of non-verbal arguments, the challenge of model design and the extent to which we can rely on the output of models - all have relevance to the role of simulation in modelling, a core subject of this thesis.

1.2.1 A Defence

In 1964, Haldane wrote a flamboyant riposte to Mayr's characterisation of mathematical population genetics as 'beanbag genetics' that contributed little to evolutionary theory (Mayr 1963, quoted in Haldane 1964). The reason, Mayr felt, for so limited a contribution was that the classical theory of Wright, Fisher and Haldane (egs. [12, 13, 14]) used too extreme a simplification of the evolutionary process. Selection is modelled as acting on a single allele without epistatic interactions and with no explicit consideration of developmental constraint, an idealisation he considered too unrealistic. Haldane's response was twofold - firstly, to offer examples showing that population genetics has usefully guided our understanding of evolution, and, secondly, to defend the trade-off between the simplifying assumptions used in mathematical arguments and the clarity that this offers when compared to verbal ones.

Haldane's examples from theory are diverse, and some, such as the detailed impact of selective advantage or the relationship between mutation, diversity and selection, are applied with little change today. Recent commentors have argued both that Mayr's characterisation of population genetics misunderstood the extent to which results were robust to different assumptions [39], and that the contribution of population genetics to evolutionary theory is more profound even than Haldane's examples. Ewens proposes that the entire modern synthesis of Darwin's evolutionary theory and Mendel's laws of inheritance is based on mathematical genetics [40]. Updating the defence, he notes how one of the most exciting recent developments in population genetics, the statistical inference of regions of the genome that have been subject to natural selection (explored in Chapter 4), was made possible by mathematical theory. Mathematical modelling has been, and continues to be, a useful tool in characterising evolution; Haldane believed that this is related to its logical exactness.

Mathematical exactness

The essential spirit of Haldane's defence of mathematical reasoning is summarised by a quote he takes from Hume (*A treatise of human nature*, Book 1, Part 3, Section 1, in [41]):

> There remain therefore algebra and arithmetic as the only sciences, in which we can carry on a chain of reasoning to any degree of intricacy, and yet preserve a perfect exactness and certainty.

Developing his position in relation to this, Haldane continues:

> Not only is algebraic reasoning exact; it imposes an exactness on the verbal postulates made before algebra can start which is usually lacking in the first verbal formulations of scientific principles

The premise, then, is that the process of logic in mathematics [1] is exact and that the assumptions made are clearly specified. Thus, 'a mathematical theory may be regarded as a kind of scaffolding within which a reasonably secure theory expressible in words may be built up'.

Leaving aside the problem of representation - the relationship between mathematical models and the world - these two points offer the principle justification of mathematical modelling. The detailed relationship between mathematics and logic depends on what one considers mathematics and logic to be, a subject too distant from applied models in population biology to delve into here. Nonetheless, Haldane is joined by others practitioners (e.g. Kenneth Arrow in the context of the social sciences [43]) in emphasising the especial suitability of mathematics for constructing logically sound arguments, and the utility of this in applied sciences.

Two details should be emphasised here. Firstly, there is nothing that specifically precludes verbal arguments from fully stating their assumptions or making logical statements. Haldane is instead condemning the normal use (c.f. Mayr) of such arguments in evolutionary biology. Verbal reasoning can be quite robust (work on informal logic is relevant here [44]), although it is often difficult to translate mathematical manipulations of a model to verbal ones. Secondly, while mathematical models tend to be more precise in their assumptions, they are not specified entirely. Complete rigour regarding assumptions would be impractical (e.g. this is the objective of Whitehead and Russells *Principia Mathematica* [45]), but a relevant example arises in Kot *et al*'s work introducing integro-difference equations as a continuous-population deterministic representation of animal dispersal [46]. The possible importance of the continuous-population assumption is mentioned only superficially in the final paragraphs of the paper. I build on this model in Chapter 2, and find that correctly representing a population as finite radically alters system behaviour. Later applications of the model do not mention this important disconnect between reality and model at all (e.g. [47]).

Additionally, implicit assumptions may be included in the course of mathematical analysis. An example that I have been involved with investigating, and which is not included in this thesis, is the finding that the specific implementation of natural selection - whether as a modification to the rate of reproduction or of mortality - can impact the behaviour of models in evolutionary game theory [5]. In the context of species invasions, a traditional approach to approximating a discrete-time system with a continuous-time

[1] An over-simplification - Haldane does not make this claim of mathematics in general, while Hume explicitly sets geometry aside from algebra and arithmetic. According to Hume, while it 'far excels [...] the loose judgements of the senses and imagination' (A treatise of human nature, Book 1, Part 3, Section 1), application of geometric knowledge to the real world is limited by our ability to observe the real world. In other words, while we can manipulate theoretical geometric objects, the relationship between these and physical entities is an approximation, such that we cannot, with absolute certainty, apply our geometric theories to the external world [42]. The population genetics of Haldane is only an approximation of the actual details of population biology, such that, given Hume's objection to geometry, the quote is perhaps abused.

one (as performed by Kendall in 1965 on a model of disease spreading, [48], and in Appendix 1 of the paper I present in Chapter 2) has been found, on more detailed analysis, to be inexact when there is a significant time-lag between generations [49].

Finally, it is important to note that mathematical models suffer the same constraints as all other models in striking a balance between accurate representation of a target system and tractability, a subject explored in by Levins.

1.2.2 The Strategy

In 1966, Levins published what has become a highly influential paper on the strategy of designing models in population biology [15]. The challenge of biological complexity has not changed in the intervening fifty years, although our tools, especially computational power, have improved. I will focus on two subjects raised by Levins that are particularly relevant to the work in this thesis - different approaches to modelling complex systems and the robustness of modelling results.

Complexity

Levins suggests that the naïve approach to modelling complexity would be to use a 'brute force' model, consisting of hundreds of partial differential equations through which the processes of nature are accurately represented. This approach fails because it is not analytically solvable, too difficult to compute, and yields results that are opaque to human understanding. In practice, then, he identifies practical strategies, suggesting that models inevitably sacrifice one of three desirable traits - generality, realism or precision [15]. The logical necessity of this trade-off is not proven [50], and ultimately depends on ones definition of these three terms [51].

For our purposes, the details of whether models can be productively characterised on these dimensions, or on any of the many other possible dimensions [52], is not critical. Any choice of model systematics is only important to the extent that one might assess how fully one has modelled a system. More profound is the realisation that, in the practice of modelling, one often has to make sacrifices. Levins' point is best seen in its historical context - as an argument against what he considered the 'brute force' approach of mid-20th-century systems ecology [53], exemplified by the International Biology Program and pioneered by Watt (e.g. [54]) among others. While there has been a modern resurgence of systems ecology modelling, exploiting the vast increase in computational power since the 1960s, there is recognition that some balance between complexity and simplification is needed (e.g. [55]).

Robustness

Levins' suggestion that different types of model, with different focusses and assumptions, can inform on the same target system, raises the question of which modelling results should be believed. Ultimately, he recommends the use of multiple models to yield different insights on a process. This also offers an indication of whether a result is robust, arising in many different partially independent models. There is an important distinction to be made here between the robustness of a modelling result to the choices of the researcher in designing the model and the likelihood of a result applying to the real world target system. The *Strategy* is ambiguous as to which is meant, first implying the former by suggesting that using multiple models indicates 'whether a result depends on the essentials of the model or on the details of the simplifying assumptions', then the latter by implying that it is more appropriate to assert robust theorems as 'biological fact' than fragile ones [15]. Some of the same issues - particularly assessing model independence - arise in both cases, and I consider these before returning to the more complex relationship between model result robustness and expectations concerning real world systems.

Quantifying the independence of results from different models is not simple [50]. The problem has been discussed in the context of climate modelling [56], where an ensemble of models yielding similar results is often taken as an indication of predictive power. It has been suggested that models containing parameters intended to represent different phenomena [57] is an indication of their independence. Another criteria relies on the difference in results between models when given the same input parameterisation [58]. Weisberg suggests a robust result will occur in multiple models, and will have the same 'causal structure' in each [51]. He focusses on the reason that a result is observed rather than the recurrence of a result itself.

Ultimately, the structural difference between models is in their assumptions, both with respect to the aspects of the target system that they hope to capture and the details of how these are represented mathematically or computationally. This suggests that the use of multiple models doesn't necessarily show if a result is robust, but can indicate whether it is robust given certain combinations of assumptions. Although Orzack and Sober [50] reject the idea of exploring this 'assumption space' - the space of assumptions that might be taken in a model - offhand, judicious sampling of it to investigate the role of specific assumptions is a common approach in theoretical population biology. The local assumption space of models is indeed sometimes depicted in studies (e.g. for invasion modelling see Fig. 1.1 in [59], for cell growth Fig. 1 in [60]). That the assumption space of models might be more formally investigated has been suggested in econometrics as 'global sensitivity analysis' [61, 62], although with statistical models such as regression primarily in mind.

One relatively unambiguous use of independent models to confirm robustness is attempted replication. Experimental data may be flawed, mathematical analysis may have errors, and computer programs may have bugs. For example, re-implementation of a computational algorithm by different programmers helps to identify and eliminate bugs. Based on my own programming experience, bugs that lead to extremely unexpected results are easily identified, but, as supported by the change-logs of the many population genetics programs I have used, more subtle errors are common.

I have suggested that using multiple models to test the role of assumptions in modelling results is valuable, and I follow this approach in some of the work presented in this thesis. Even Orzack and Sober consider the use of multiple models a 'useful heuristic', such that my approach is in general agreement with a range of authors [15, 52, 50, 51]. The more profound question is whether a result arising from many partially independent models offers additional evidence for that result holding in the target natural system. Orzack and Sober are suspicious, noting that this would appear to yield information about the world without empirical observation. However, observation may be required in the choice of which models to test and their representation [52], implying that information about the world can be obtained from sets of multiple independent relevant models.

In this thesis, I use simulation to investigate the assumptions of relatively simple models in an effort to identify the implications of different model assumptions, and I now explore how simulations are used in modem scientific research.

1.3 Complexity, tractability and computational simulation

Levins identified the challenges inherent in designing and interpreting brute force models as the primary motivation for using alternative model-building strategies, while Haldane emphasised the value of exactness and clarity in mathematical models. Both authors were in favour of simplification. Given these arguments, a question arises as to the role of increasing computational power in modelling. The ability to do many calculations rapidly and both store and quickly retrieve large amounts of data is highly suited to numerically exploring the behaviour of complicated systems in greater detail. Unless Levins' case against the brute force approach has been tempered, such power may be a mixed blessing. My argument is that a computational simulation of complex systems does not need to be complex to be useful, and that much of the utility of simulation arises in the logical consistency that makes mathematics so effective. To develop these claims it is it is necessary to identify what computational simulation is, and what it is not.

1.3.1 Computational simulation

As with modelling, many definitions of simulation have been suggested. Some examples covering a range of perspectives are

> ‘the operation of a model of a system’ (Maria 1997, [25])

> ‘the process of designing a model of a real or imagined system and conducting experiments with that model’ (Smith 2003, [63])

> ‘the imitation of the operation of a real-world process or system over time’ (Banks 1998, [64])

> ‘a program that is run on a computer and that uses step-by-step methods to explore the approximate behavior of a mathematical model’ (narrowly),
> *or*,
> ‘a comprehensive method for studying systems’ (broadly; both concerning computer simulation, both Winsberg 2015 in [65]).

These definitions vary from very broad [25, 63] to quite narrow [64]. Given this, it is arguably appropriate to merely identify the important common themes between the simulations conducted in this thesis. Thus, I will partially follow Smith and partially Winsberg and view simulations as ‘experiments’ [63] that ‘explore the behaviour of a mathematical model’ [65]. Details such as whether the time evolution of a system is represented [64] - which happens to be the case for each simulation I present - are not important, as there is no reason to believe that this quality is fundamental to the value of simulation in testing models in population biology.

Why might we want to run experiments to explore a mathematical model? An indication is given by Haldane, who highlighted the utility of simplifying assumptions in allowing the pioneering population geneticists to ‘pose problems soluble by the elementary mathematics at [their] disposal’ [38]. The representation of a system mathematically is not only limited by the imagination of the modeller, but also by their ability to manipulate mathematical models so as to solve them. This bias against insoluble models - such that convenience of manipulation is a constraint in model development - is pragmatic, and no doubt responsible for many of the most successful applications of modelling. Nevertheless, the possibility arises of systematic misrepresentation of target systems. Simulations can be used to investigate intractable mathematical models, and, because they are subject to different constraints, lend themselves to different representations of a system. They are useful in this regard because they retain the logical exactness [2] of mathematics and provide evidence of assumptions through the simulation program itself.

[2] Floating point errors aside

The broad definition of simulation proposed above captures a wide range of approaches, from detailed attempts to replicate biological complexity *in silico*, such as the International Biology Program (IBP) or the more recent Human Brain Project [66], to numerical investigation of highly simplified mathematical models. While my focus in this thesis is firmly directed toward the latter, having discussed Levins [15] in some detail it is appropriate to briefly indicate the state of 'brute force' biological models today.

1.3.2 Complex simulations of ecological complexity

Simulations attempting to faithfully capture complex biological systems continue to be widely implemented. A particularly dramatic example is the GUMBO model of ecosystem services, with 930 variables and 1715 parameters used to describe 'dynamic feedbacks among human technology, economic production and welfare, and ecosystem goods and services within the dynamic earth system' [67]. The 43 variables connected by 22 non-linear relationships used in Forrester's World Dynamics model (1971, [68]), with its rather dire predictions for our species' future, pale in comparison. Although GUMBO and its descendants (e.g. MIMES [69]) incorporate a complex biosphere model, they do not focus on biological complexity *per se*. Increasingly, agent-based models (ABMs) have been designed to represent ecological systems [70, 71] as an alternative both to extreme mathematical simplification and to the brute-force use of hundreds of differential equations as envisioned by Levins. ABMs take a 'bottom-up' approach, attempting to capture the microscopic interactions between individual agents (often organisms) and through these understand emergent system-level behaviour.

The use of ABMs in ecology represents a significant methodological development. This is because they represent agents as discrete, autonomous and individually variable, qualities that are usually summarised within the aggregate state variables (e.g. population size or allele frequency) of traditional mathematical models [70]. It has been argued that ABMs can be general, realistic and precise [72], capturing three properties that Levins thought could only be maximised by brute-force models. The critical question is not whether these models disprove Levins' framework, but whether they have substantial advantages as compared to other complex models. Thus, we return to Levins' criticism of brute-force modelling in ecology - that there are many parameters about which we have little prior knowledge, that we cannot solve or sufficiently simulate very complex models, and that the models are so complex that their results have little meaning to us. I suggest that ABMs make most headway, as a modelling strategy, against the issues associated with parametrisation and model interpretation, while computational limitations remain a concern for all complex models.

Correct parametrisation is critical to obtaining meaningful results in both ABMs and other models. The critical difference in ABMs is that the focus of parametrisation has changed, such that many parameters refer to properties of the individual rather than to those of populations. This is an advantage only when it is easier to experimentally measure traits of individuals, and to characterise individual variability, than to quantify more abstract higher-level properties such as predation pressure or carrying capacity. With respect to model interpretation, complex ABMs have a more quantifiable advantage. While it remains possible to summarise the evolution of state variables, and thus retrieve similar information as might be output from more abstract models, it is also possible to follow the actions and experience of individual agents, the 'narrative approach' suggested by Millington *et al* [73]. It may be relatively easy to relate narratives of agents to narratives of the organisms they represent.

Both complicated ABMs and systems of many simultaneous non-linear partial differential equations are unlikely to be mathematically tractable, and hence computational simulations are used to explore their behaviour. For complex models with many parameters, this can require considerable computational power. While simulations of extremely complicated models are now feasible, the number of parameter combinations in a model increases as the power of the number of parameters. Even with modern computers one can only explore a small portion of the parameter space for many-parameter models. Whether this is important depends on the research question.

To summarise, complex models remain widely used in ecology, with a current focus on ABMs. These, in particular, have changed the nature of modelling complex systems, but the problems that Levins identified as limiting the use of brute-force models still apply.

1.3.3 Simple simulations of mathematical simplification

An alternative use of simulations is the numerical investigation of simple models that cannot be comprehensively studied through mathematical analysis[3] . This approach has a long history - for example, in pioneering work modelling the dispersal of mosquitoes published in 1906, Karl Pearson and John Blakeman [10] used mechanical integrators to graphically calculate the distribution of random walkers over short times. They were thus able to observe the approach of the population distribution over time to an easily calculated approximation suitable at long times. The essential objective is similar to that applied in this thesis - to test the robustness of mathematical models to the violation of their assumptions through numerical methods. Confirming analytic results in this way remains a common technique in modelling.

Given that the combined use of simulation and mathematical analysis is already relatively common in theoretical population biology, a question arises as to why I feel it

[3] Often, relatively simple mathematical models can be characterised in certain limits - for example, as certain parameters tend to zero, or are equal - even if they cannot be solved.

appropriate to dedicate a portion of my thesis to the methodological advantages this offers. My answer is two-fold, partly practical and partly abstract. In practical terms, I merely observe that my simulations have actually clarified the results of three models. Two of these are traditional mathematical models that have been studied for a considerable period of time. That my findings were not previously known, and that the techniques I have used, while varied, are not complicated, suggests that the role of simulation in clarifying the impact of the assumptions of models in population biology warrants emphasis.

My second argument in favour of using simple simulation along with mathematical analysis is slightly more abstract. This is that there are various different ways that simulation can be used to investigate model assumptions. I do not intend to develop this point formally, but will return to it repeatedly through the thesis. The simplest approach is to use simulation to explore the parameter space of a model, usually in regions for which mathematical approximations are unavailable or poor. This process is still assessing the impact of different modelling assumptions - each choice of parameters in a model represents a different target system [50]. Some of the assumptions that I test in this thesis are more complicated, in that they are integral to the design of the models. For example, in Chapter 2 I assess the effectiveness of a mathematical model, describing an invasive species spreading over space, in capturing the microscopic properties of the process it purports to represent. In Chapter 3, I modify the structure of a model of animal migration between sub-populations to incorporate a simple representation of animal trading, and, in doing so, find that breaking the assumption of random, unbiased animal movement substantially impacts model results. In Chapter 4, I show that relaxing the common and incorrect assumption of a constant recombination rate along the genome can substantially impact our assessment of test statistics designed to detect genomic signatures of natural selection. In the first case, I am testing an assumption taken to simplify mathematical analysis, in the second, an assumption taken because it is thought to best represent ecological systems, and in the third an assumption which is known to be suspect that is applied in order to constrain parameter space. I suggest that these three examples reflect the breadth of different assumptions that arise in population biology models, and an awareness of the flexibility of simulation in testing different types of assumption for different reasons is useful.

My simulation work has lead to a series of interesting findings, some primarily methodological, others with direct relevance to the real world. I now detail these and describe the structure of my thesis.

1.4 Outline of thesis and main results

The outline of the thesis is as follows. I will begin each chapter by summarising the results and modelling included, and by emphasising the relationship between the work presented and the use of simulation modelling in population biology. I will then give details on the motivation and background of the specific modelling question addressed , as necessary [4], before presenting a manuscript detailing some the findings of modelling work that I have conducted. In each case I am the first author of the manuscript and all co-authors are aware of the inclusion of it in this thesis. Details of the work that myself and different co-authors performed are included in the Acknowledgements section. I also indicate, in Thesis Appendix 1, conferences and forums where our results have been presented. In the case of Chapter 2, on species invasions, the manuscript is the final accepted version of a published manuscript, while those presented in Chapters 3 and 4 are drafts. Each chapter will continue with a more detailed discussion of the methodology and model results, and conclude with some comments on the implications of the work for the practical use of simulations in population biology.

I discuss and integrate my findings in Chapter 5, drawing together the work presented to identify advantages and challenges of biological modelling in general, and the specific role of simulations in improving models.

As the thesis progresses, I will present a number of core findings that are of relevance to the modelling of the systems we explore, as follows:

1.4.1 Chapter 2 - Modelling species invasions

Long-range dispersal, stochasticity and the broken accelerating wave of advance

Published manuscript

Accelerating species invasions may lead to a particularly sudden and dramatic ecological impact, so are of interest to invasion ecologists. We have clarified a critical feature of the dispersal regime that leads to an accelerating species invasion in a finite population model, and confirmed that this is correct for a variety of structurally different models. Specifically, dispersal kernels - the distribution describing the probability of dispersal at any given distance - that decay with a power law tail, $K(x) \propto |x|^{-\beta}$, cause acceleration when $\beta \leq 3$ in a 1-dimensional system incorporating stochasticity in dispersal. Deterministic models, which assume a population is infinitely divisible, yield

[4] In some cases, the literature review of the manuscript presented in the chapter contains much of this information

a weaker constraint, suggesting that all fat-tailed dispersal kernels cause acceleration. This confirms some previous results, but contradicts others.

The implication of this is that incorrectly specifying a model can lead to substantial errors in the output. We have investigated the temporal dynamics of the error associated with incorrectly using a deterministic mean-field model as opposed to a stochastic model. Deterministic models are frequently encountered in the literature.

We further extend our modelling in two ways. Firstly, we note that there are limits to the maximum dispersal distance of any organism, and explore the impact of truncating the dispersal kernel to capture this property. This leads both our stochastic and deterministic models to generate a constant velocity wave of advance, and we again explore the error over time of correctly specifying a truncated kernel but using an incorrect representation of stochasticity. Secondly, we identify important subtleties related to the implementation of logistic density-dependent population growth that impact results from the stochastic, but not deterministic, models of species invasions. Density-dependent population growth is an expected attribute of real populations.

1.4.2 Chapter 3 - Modelling animal markets

On the counter-evolutionary effects of market mediated gene flow

Draft manuscript

The migration of domestic species is usually mediated through human choices, with the trading of domestic animals being a clear example. The impact of gene flow through markets has not, as yet, been characterised. We have adapted a simple model of migration between two patches to include some of the complex migration patterns that might be expected when animals are traded by humans. The patches have unequal wealth distributions, such that individuals from the wealthier patch are better able obtain their preferred types of animal. In this model, wealth inequality can disrupt the action of natural selection, such that an animal or breed with lower fitness may out-compete one with higher fitness, even if it is not favoured in either of the two patches. This result arises due to the interaction of migration and population regulation.

Given that our findings are counter-intuitive, I discuss in detail real world evidence for the population dynamics narratives implied by our model. Current evidence is inconclusive, and the work is best considered an exploratory model showing evidence of the types of phenomena that could occur given market-mediated gene flow rather than a robust statement on expected patterns in the real world. This chapter explores only one of many interesting questions that could be asked about the impact of markets on animal genetic variation.

1.4.3 Chapter 4 - Modelling natural selection

Refining the use of linkage disequilibrium as a robust signature of selection

Draft manuscript

Many statistics have been suggested to detect signatures of natural selection in population genetic data. We use simulations to test the statistical power of a large set of statistics that do so by identifying characteristic distortions in the correlation between the allelic state at different loci (linkage disequilibrium; see Chapter 4 definitions of genetics terminology). An assumption of a constant recombination rate is often used when testing selection statistics. There is considerable evidence that this assumption is unrealistic, and we find that incorporating variation in the recombination rate negatively impacts the performance of several selection statistics. We therefore suggest modifications to currently used statistics, and also propose novel statistic formulations, that improve statistical power to detect recent positive selection.

We further assess these statistics by quantifying their ability to replicate previously identified signals of selection in real genetic data. These results broadly support our proposal that controlling for expected recombination improves the power of selection statistics.

I have also used patterns of linkage disequilibrium generated by positive natural selection to help identify a selected variant in the *CPT1A* gene among a northeast Siberian population. My work also provided additional support in localising the causal SNP, a high-frequency non-synonymous variant known to be associated with child mortality in modern populations but inferred to have been under positive natural selection in the recent past [2].

Chapter 2

Modelling species invasions: stochastic long-range dispersal and invasion acceleration

2.1 Chapter introduction and summary

In this chapter, I explore a model of a species invading previously unoccupied territory through repeated dispersal events that may occur across long distances. I focus on the role of stochasticity in the dispersal process, showing that deterministic and stochastic variants of the same model yield radically different spreading behaviour. The work offers two contributions to the study of species invasions. Firstly, we are able to clarify the conditions under which long-range dispersal may cause the rate of a the species invasion to increase with time. Secondly, we build an awareness of the implications of a widespread modelling assumption - deterministic dispersal due to an infinitely divisible population - in distorting model behaviour. The first result is of practical importance, and is potentially useful when predicting the actual course of species invasions. The second has greater relevance to the core theme of this thesis, highlighting the potential of a convenient mathematical simplification to strongly distort model results.

The approach in this chapter is to focus on the relationship between the model and the world rather than to explore a model thought to be relevant to the world. In finding that a frequently applied approximation, taken for mathematical convenience, is often inappropriate, I am assessing the robustness of certain modelling results and, therefore, the conditions under which the model is valid. In this way, simulation allows us to identify important assumptions and quantify how important they are under different situations. All the simulations I perform are highly idealised, demonstrating the utility of simple simulations in honing our theoretical understanding of invasion events.

The structure of this chapter is as follows. I will begin by reproducing a pre-print version and supplementary material of *Long-range dispersal, stochasticity and the broken accelerating wave of advance* [1], a paper published in Theoretical Population Biology (doi:10.1016/j.tpb.2014.12.003) by myself and Tim Sluckin. This paper presents a range of findings concerning the difference between a deterministic and stochastic model of species invasion given long-range dispersal. We focus on how long an invasion event takes, confirming and expanding on previous observations that the rate of invasion can be qualitatively different under an identical dispersal regime for deterministic and stochastic representations of dispersal. The deterministic representation essentially allows individuals to be infinitely divisible, while the stochastic variant considers a finite population size. We further show that the approach to the deterministic behaviour as population size is increased is very slow under certain long-range dispersal regimes. The implication is that widely used integro-difference models [46], which bear considerable resemblance to our deterministic system, will often yield highly inaccurate results, even if features like population growth and regulation, and dispersal patterns, are accurately characterised. Appendices that were included in the submission for completeness but essentially reproduce known results are included the Thesis Appendices section, §3.

Our approach to representing long-distance dispersal involves describing the relative probability of dispersal over all distances with a simple statistical distribution. We chose to use two families of distribution - power laws and stretched exponentials - that can be used to represent a wide range of possible dispersal regimes. The choice of these specific distributions falls neatly into the narrative of animal movement ecology, in that they have been proposed to represent the movement patterns of various organisms. I explore evidence for these distributions, and their possible generative functions, immediately after the paper.

I close this chapter by considering the broader implications of our findings with respect to the role simulation has in population models. I suggest that the flexibility of simulation allows us to probe the assumptions of mathematical models, and hence assess their relevance to the real world. A similar theme arises in Chapter 4, in which I use simulations as a tool to assess the performance of statistical tests designed to detect natural selection in population genetic data.

2.2 Long-range dispersal, stochasticity and the broken accelerating wave of advance *(Published manuscript)*

G. S. Jacobs and T. J. Sluckin

2.2.1 Abstract

Rare long distance dispersal events are thought to have a disproportionate impact on the spread of invasive species. Modelling using integrodifference equations suggests that, when long distance contacts are represented by a fat-tailed dispersal kernel, an accelerating wave of advance can ensue. Invasions spreading in this manner could have particularly dramatic effects. Recently, various authors have suggested that demographic stochasticity disrupts wave acceleration. Integrodifference models have been widely used in movement ecology, and as such a clearer understanding of stochastic effects is needed. Here, we present a stochastic non-linear one-dimensional lattice model in which demographic stochasticity and the dispersal regime can be systematically varied. Extensive simulations show that stochasticity has a profound effect on model behaviour, and usually breaks acceleration for fat-tailed kernels. Exceptions are seen for some power law kernels, $K(l) \propto |l|^{-\beta}$ with $\beta < 3$, for which acceleration persists despite stochasticity. Such kernels lack a second moment and are important in 'accelerating' phenomena such as Lévy flights. Furthermore, for long-range kernels the approach to the continuum limit behaviour as stochasticity is reduced is generally slow. Given that real-world populations are finite, stochastic models may give better predictive power when long-range dispersal is important. Insights from mean-field models such as integrodifference equations should be applied with caution in such circumstances.

Keywords: long distance dispersal, wave of advance, species invasion, stochastic modelling

2.2.2 Introduction

The manner in which alleles, species and diseases spread over space is of fundamental interest to population biologists. These processes have an important impact on many evolutionary and ecological systems, and are particularly relevant in the modern world, where increasing global trade [74] and highly interconnected transport systems [75] change the dynamics of disease and species dispersal. For example, international air travel has been suggested as a major driver of the spread of disease, including the 2009 H1N1 influenza A swine flu virus pandemic [76]. Anticipating species invasions, and identifying how they might progress in such conditions, is an immediate and relevant problem.

Various models have been constructed in order to theoretically explore the dynamics of spreading populations. These guide our predictions about future genetic, demographic or disease prevalence trends, and our understanding of the history implied by current patterns. A core feature of models is whether they explicitly incorporate stochasticity. Traditional approaches tend to use deterministic approximations of the underlying stochastic process. Here, there is an assumption that over many repeats of an event with a random element the stochasticity will average out, and can be ignored without invalidating results. Such models can often be analysed mathematically, but are sometimes sufficiently complex that a computational solution is necessary.

Stochastic models are usually more computationally intensive and less analytically transparent, but accept that explicitly including the randomness of events is important. It is often unclear which approach is preferable. In the specific case of species dispersal, a finite population of organisms that move and reproduce with a degree of independence implies a finite number of dispersal events. Stochasticity at small scales can have a significant impact on larger scale behaviour, and it is possible that averaging these events has a qualitative impact on model results.

One feature of population spread that is of particular practical interest is the expected rate of invasion. Deterministic equations predict that under many conditions population expansion occurs through a wave of advance travelling at constant velocity [77]. In certain cases, however, where there is a relatively high frequency of long-distance dispersal events, this wave will accelerate indefinitely [46]. The integrodifference model that retrieves this latter result has been widely applied in modelling species dispersal [78, 79, 80, 81, 82, 47, 83]. However, the approach is deterministic, and it is not clear that the underlying stochasticity of dispersal can be ignored without causing inaccuracies. The impact of randomness on the accelerating wave of advance will therefore be the principal subject of this paper. We explore this by considering a range of stochastic models and their mean-field deterministic approximations, in which many dispersal events are described as a single average process.

Fisher-Kolmogorov and its limitations

Classical modelling of population spread has taken the form of reaction-diffusion equations. Here, a diffusion approximation is used to model the underlying stochastic dispersal and reproduction processes, which occur concurrently and independently of one another. This is a macroscopic approximation, obtained from the stochastic description by truncating in space or time to some finite order [84]. The paradigm is the Fisher-Kolmogorov equation [11, 85, 86]:

$$\frac{\partial n}{\partial t} = \alpha n(1 - \frac{n}{K}) + D\nabla^2 n, \tag{2.1}$$

where n is population density at time t, α is the maximum growth rate, K is the carrying capacity (in some suitable units) and D is the diffusion constant. The equation is continuous in space and time and expresses the combination of logistic growth and Fickian diffusion. The diffusion constant, or diffusivity, describes the mean square distance over which a particle diffuses per unit time given a gradient of one unit, and may be expressed in dimensions L^2T^{-1}. A higher diffusion constant implies that the flow of organisms from full to empty space is easier and thus more rapid.

The use of a single parameter D to represent many possible dispersal regimes follows from arguments based on the central limit theorem [87]. It is justified by the relationship between Fickian diffusion and the stochastic process underlying it, Brownian motion. We can describe this process mathematically as a random walk.

A basic random walk is a stochastic system in which the position, x, of a particle is iteratively updated by its jump distance, drawn from a given probability distribution. This probability distribution describes the probability of dispersal over a distance l in a time interval, and is known as the *dispersal kernel*, $K(l)$. If we run many random walks with a given starting position, the distribution of the particles will spread out over time. Supposing a symmetric dispersal process, the mean position remains close to zero, but the diffusivity can be captured by the deviations around this mean. For Brownian motion, and indeed more general random walks,

$$< x(t)^2 >= 2Dt, \tag{2.2}$$

with the constant of proportionality defining the diffusivity. The central limit theorem prescribes that the distribution function of long-time positions is Gaussian so long as the same kernel applies to all particles, there are no long-range correlations in jump-distance, and the kernel has a finite first and second moment. D is related to the variance of the kernel by

$$D = \frac{1}{2}\int_{-\infty}^{+\infty} l^2 K(l)\mathrm{d}l. \tag{2.3}$$

When the variance is unbounded, D is similarly not well defined, a point we return to shortly.

An initially isolated population that behaves according to the Fisher-Kolmogorov equation spreads out over time, creating a ‘wave of advance’, while maintaining a logistically determined level of occupation behind the travelling front. The model has been subject to much mathematical investigation, and a range of velocities can be sustained. However, under suitable initial conditions [85], including those most relevant to biological invasions, the wave speed (after transient acceleration) asymptotically approaches

$$c = 2\sqrt{\alpha D}. \tag{2.4}$$

For c to be asymptotically constant both D and α must exist and be asymptotically constant.

Laying aside model-specific issues such as environmental heterogeneity, advection, and qualities of population growth such as Allee effects, there are two general concerns about the application of the Fisher-Kolmogorov equation. Firstly, long distance dispersal may complicate the diffusion term. Secondly, stochasticity may invalidate results obtained from averaged processes. We deal with these points in turn.

Long-distance dispersal through integrodifference models

Standard theory suggests the diffusivity D can capture a wide range of stochastic dispersal processes through the relationship in Eq. (2.2). In the context of population spread, a naïve assumption of a normally distributed dispersal kernel would seem reasonable. However, many species appear not to follow this dispersal pattern, with dispersal better represented by a ‘fat-tailed’ kernel. These kernels involve an excess probability of dispersal at longer distances; specifically, the tail of the dispersal kernel decays more slowly than an exponential distribution. Such dispersal regimes have been observed in fungal spores [88], plant seeds [89], and in mammals and birds [90]. Under these conditions, it becomes less clear that D will capture the dispersal process faithfully, and there is a strong argument for explicitly incorporating the dispersal kernel itself into a model.

As we have noted, some fat-tailed kernels decay so slowly that the variance or other moments are not well defined. Specifically, when the tail of a kernel decays as a power law, $K(l) \propto l^{-\beta}$ as $l \to \infty$, the $(\beta - 1)^{\mathrm{nth}}$ and greater moments are not finite. This phenomenon is due to the dominant role that rare large values have on the characteristics of the distribution, and is useful for incorporating a relatively high probability of extremely long-range events into the dispersal regime. If $\beta \leq 3$, we can predict dispersal behaviour by considering a particular class of random walks, known as Lévy flights, for which the second moment is undefined [91, 92].

When the variance is unbounded the effective diffusivity increases with time, termed superdiffusion. Given the role of D in the Fisher-Kolmogorov equation, we might expect these kernels to lead to an accelerating wave of advance. Indeed, reaction-diffusion equations of the Fisher-Kolmogorov type that model Lévy flights as fractional diffusion have been investigated, and lead to exponentially accelerating waves [93]. There is a significant body of theoretical and empirical work investigating Lévy flights in the context of foraging behaviour [94].

Several authors [95, 46] have pointed out that systems subject to fat-tailed kernels which nevertheless still possess a finite variance (and hence a well-defined D) may also exhibit anomalous behaviour at long times. This is reflected in the lack of analytic behaviour of $\tilde{K}(k)$ at low k, where $\tilde{K}(k)$ is the Fourier transform of the dispersal kernel $K(l)$. The low k non-analyticity then leads to problems in the application of the central limit theorem at long times. To determine more accurately the implications of such anomalous kernels for species diffusion, we need to describe the dispersal process explicitly, rather than summarising it merely in terms of a diffusivity D.

Kot *et al* [46] achieve this mathematically by incorporating the dispersal kernel directly in an integrodifference equation of the form:

$$n(x, t+1) = \int_{-\infty}^{+\infty} K(x-y) f[n(y,t)] \mathrm{d}y. \tag{2.5}$$

The function $f[n(y,t)]$ applies the population growth process, while the (normalised) dispersion kernel $K(x-y)$ represents the relative probability of dispersal between positions x and y in continuous space, with $l = y - x$. Importantly, there is no assumption that the underlying stochastic dispersal process is Brownian. However, unlike the Fisher-Kolmogorov equation, time-steps are discrete, which can lead to velocity deviations from the continuous time case in similar systems [96, 49]. Furthermore, growth and dispersal are no longer concurrent and independent. Rather, growth and dispersal occur sequentially, such that there is a coupling between the two processes.

We can describe Eq. (2.5) as a mesoscopic representation of the stochastic dispersal and growth processes, in that the random behaviour of individual organisms is averaged as a probability density function [84]. The time-evolution of the probability distribution for occupation over space is then studied, which can be considered population density when a population is large. The approach thus captures elements of organism movement that are summarised by D in the macroscopic reaction-diffusion model of Eq. (2.1). Under certain circumstances, such as a normally distributed dispersal kernel combined with logistically limited growth, Eqs. (2.5) and (2.1) retrieve identical wave velocities [46].

However, this does not imply that the microscopic processes for which they provide deterministic approximations are identical.

In apparent contradiction to Fisher-Kolmogorov predictions, certain fat-tailed dispersal kernels with finite second moments lead to indefinitely accelerating waves of advance [46] in the integrodifference modelling framework. Given that more information is preserved about the dispersal kernel in the integrodifference approach, we might regard it as more broadly applicable. Specific kernels with this effect include stretched exponentials and power laws, where for large dispersal distance l, $K(l) \propto e^{-|l|^{\gamma}} : \ \gamma < 1$ and $K(l) \propto |l|^{-\beta} : \ \beta > 3$ respectively. The stretched exponential kernel leads the spatial extent of the wave to increase as a power law over time, with exponent $\frac{1}{\gamma}$ [84]. Evidence from both reaction-diffusion equations [97, 93] and integrodifference models [46, 98] indicate that power law kernels cause wave velocity to increase exponentially with time, an effect that persists when $\beta > 3$ [97].

There are many documented examples of apparently accelerating species invasions (eg. rice water weevil, *Lissorhoptirus oryzophilus*, in Japan, [99]; cheatgrass, *Bromus tectorum L.*, in North America, [100]; among other diverse plant species, see [101]; potentially Californian sea otters, *Enhydra lutris nereis*, [47]; also see [102]). In some cases, these behaviours may be due to factors other than long-range dispersal. Nevertheless, the link between long distance events and accelerating waves has been explicitly suggested with respect to the spread of several plant and human pathogens using data from empirical studies and observed invasion events [103]. In these cases dispersal is either by wind or via avian vectors.

As Kot *et al* noted, indefinite acceleration is biologically unsustainable, and can break down for several reasons. These include the introduction of an Allee effect, a long-distance cut-off to the dispersal kernel, and effectively introducing a spatially determined cut-off to dispersal by limiting system size [46]. Stochasticity can also have a pronounced effect on system behaviour.

Although the implications of certain of these features remain unclear, integrodifference equations are frequently used in species invasion modelling. The flexibility afforded by explicitly representing the dispersal kernel has allowed authors to explore various phenomena, often incorporating long-distance dispersal [46, 78, 79, 80, 81, 82, 104, 47, 83, 98]. Recently, the approach has been suggested as one of four preferred methods for pest risk analysis [105]. It must be emphasised that models in general, and here the integrodifference equations method in particular, are only an approximation of real-world behaviour, and if the implicit assumptions are incorrect the results will also be unreliable.

Demographic stochasticity: ambiguous results

Both the descriptions of population spread introduced above are deterministic. They are justified by the belief that they will capture the essential behaviour of the underlying stochastic processes of reproduction and dispersal. Under which circumstances they are the correct deterministic limits is unclear.

Demographic stochasticity is known to have an impact on model behaviour. A reduction in wave velocity is usually suggested [106, 107, 108], though contradictory results exist for a two dimensional stochastic cellular automata model [109]. In the simple, linear case, where population growth and dispersal are not density-dependent, work with branching random walks suggests that introducing demographic stochasticity does not generally slow invasions [110]. Separately, the interaction between dispersal kernel and reproductive rate has been highlighted as having an important impact on the structure of the wave [111].

In the context of the non-linear Fisher-Kolmogorov equation, a productive route of enquiry has been approximating stochastic effects by introducing a cut-off to the population growth term, thereby reducing wave velocity [112]. This approach has allowed the characterisation, for example, of stochasticity-induced velocity corrections where the dispersal kernel is exponential, representing the boundary case past which the Fisher-Kolmogorov approach cannot be naïvely applied [113]. In cases where acceleration is predicted by deterministic models, demographic stochasticity has been suggested to break acceleration, even given extremely fat-tailed kernels with unbounded variance [106, 108]. However, noting Lévy flight predictions and Mollison's density-dependent model of epidemic spread in continuous time [95], which preserves acceleration in the context of these kernels, results remain ambiguous.

In summary, models that are able to incorporate long-distance dispersal, demographic stochasticity due to finite population size, and density-dependence (which has a deep theoretical history, [8]) are most biologically plausible but least understood. It is unclear, for example, when exactly wave acceleration should be broken by stochasticity in dispersal or reproduction, and how this might occur. To explore this in more detail, we here present the results of simulation modelling of population spread incorporating two well-known classes of dispersal kernel. Although explicitly incorporating stochasticity into models is intuitively most relevant when predicting the spread of species with lower population size/fewer dispersal events, we find that doing so may be prudent in many situations, and particularly when dispersal is best represented by a fat-tailed kernel.

The layout of the paper is as follows. In §2.2.3 we introduce the modelling framework of our simulations. In §2.2.4 we present results for both stochastic and deterministic models, finding that most waves which accelerate in the mean-field model do not when stochasticity is introduced (§2.2.4.3, 2.2.4.2). However, acceleration persists in the stochastic model for power law kernels when $\beta < 3$. We also examine the impact of reducing stochasticity by increasing the carrying capacity. Here, the behaviour of the finite-population, stochastic model only approaches that of the mean-field system slowly when dispersal is represented by fat-tailed kernels, §2.2.4.4. This has important implications regarding the situations in which averaging approaches such as integrodifference equations may be successfully applied. In our final simulations, we consider the effect of a long-distance kernel cut-off on wave velocity and dispersal dynamics, §2.2.4.5. This represents an interesting case as truncated dispersal kernels, particularly power laws, are often encountered in the literature [114, 115, 116]. We conclude by confirming our results for several variations on our model, §2.2.4.6, and discussing our results in the context of previous work, §2.2.5.

2.2.3 Simulation modelling

We here introduce our two dispersal models, a stochastic model and its mean-field approximation, and describe our methods of data analysis and the simulations conducted. The stochastic model design builds on the *simple epidemic* model explored by Mollison [95] and also bears some resemblance to the dispersal model of Clark *et al* [106] and to Kot *et al's* linear branching random walk model for population spread [110]. Our models incorporate density dependence in the dispersal process, and follow an algorithm that might be described as a 'seeding random walk'. As with the models of Mollison and of Clark *et al*, dispersal and reproduction are united as a single process. As a result, they more closely resemble the emission of seeds by plants than animal dispersal, in which movement is coupled with reproduction. Given this, we also briefly discuss the differences between our model of dispersal and the Fisher-Kolmogorov population diffusion model.

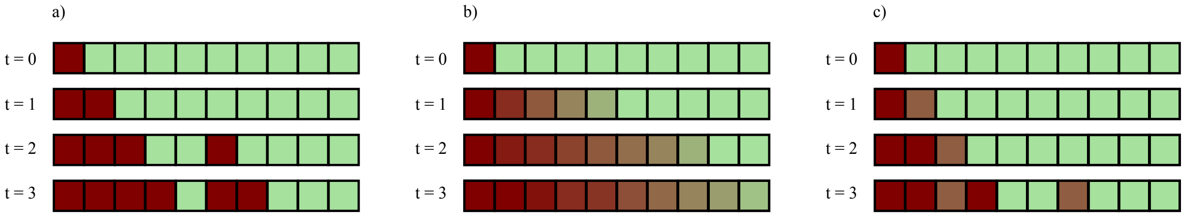


FIGURE 2.1: Diagram of simulation design and example evolution for the models explored. a) The stochastic Model 1, with $N = 1$; b) the mean-field deterministic Model 2 and c) the stochastic Model 1, with $N = 2$ and a resultant reduction in stochasticity.

All invasions occur across a 1-dimensional lattice of size L and are discrete in both space and time. Initial conditions have all sites unoccupied apart from the left-most site, $n(x = 0, t = 0) = 1.0$ and $n(x > 0, t = 0) = 0.0$. A dispersal/growth process is iterated through time, and simulations are terminated when system filling, $n(L)$, reaches 90% ($n(L) = \sum_{x=0}^{L} \frac{n(x,t)}{L} \geq 0.9$). This termination condition implicitly defines our measure of wave velocity, which is estimated using the time until termination for different size systems. For simulations incorporating death we apply other measures of population extent, see §2.2.4.6.

We ignore dispersal events to sites outside of the lattice. Such boundary conditions are biologically reasonable in many cases. However, they lead to two possible concerns. Firstly, there is the possibility of extinction by over-dispersal in some models [86, 117]. We generally ignore death in our simulations in order to focus on stochasticity associated with dispersal, so this is rarely a problem. Secondly, a finite-size system limits the maximum distance that a propagule can travel. As we repeat simulations with increasing system size, an artefactual increase in the distance to which the dispersal kernel stretches occurs. This in turn implies a change in the rate of reproduction. Given these concerns, we explore periodic boundary conditions (see Appendix 2, §2.2.7), with no qualitative change to results.

Stochastic Model 1: We begin by describing our stochastic model algorithm, represented in Fig. 2.1a and 2.1c. Occupation at each lattice site, $n(x,t)$, is a value $\frac{n}{N}$, with integer $n: \quad 0 \leq n \leq N$. After setting up the initial conditions, we proceed through a series of iterated steps:

1. The population of each site *reproduces*, making $n(x,t)N$ reproduction attempts with successful birth probability b for each event.
2. Each newborn individual *disperses* a random distance l drawn independently from the dispersal kernel $K(|l|)$, with a logistic probability of success $1 - n(x + l, t)$.
3. Once each reproduction/dispersal event has occurred, the population of each site is *updated* according to the number of successful propagules arriving at it. Each successful propagule increases site filling by $1/N$.
4. *Death* is now implemented, each individual dying with a probability d, such that there are up to n deaths in a site with filling $n(x,t) = n/N$.
5. For any site with filling $n(x,t) > 1.0$, the population is reduced to $n(x,t) = 1.0$.
6. Steps 1 to 5 are repeated until the termination condition is met, or population extinction occurs.

Dispersal can occur in either direction. When $N = 1$, dispersal is highly stochastic, each occupied site making a single dispersal attempt per generation. As N is increased, a fully occupied site releases increasingly many propagules of decreasing size. Stochasticity decreases, and we approach the mean-field model presented below; we conjecture that in the limit $N \to \infty$ we indeed recover this mean-field model. In these models, stochasticity arises both in the dispersal/birth process and through death. However, here we focus on the randomness of dispersal, and usually set $b = 1$ and $d = 0$. Our main results are found to be robust to positive d and $b < 1.0$, as explored in §2.2.4.6.

Mean field Model 2: This is an average deterministic representation of Model 1, represented in Fig. 2.1b. We now follow a mean-field approach, in that we assume that the dispersal of very many interacting propagules can be described in terms of a single average process. The occupation of each site can now be defined so that $0.0 \leq n(x) \leq 1.0$. After setting up initial conditions, we update the system iteratively using our mean-field equation,

$$n(x, t+1) = (1-d)n(x,t) + (1-d)b[1-n(x,t)] \sum_{l=-\infty}^{+\infty} K(|l|)n(x+l, t), \tag{2.6}$$

where d and b are considered the average effect of the birth and death probabilities in Model 1, with $0 \leq d$, $b \leq 1.0$. In this equation, site occupation at $t+1$ is determined by a logistically limited dispersal from all other sites, with birth rate b, followed by a death stage. The model resembles the integrodifference version of the simple epidemic explored by [104], though differences exist in our application of death and in the lattice structure. Given that birth and dispersal are not always commutative in reaction-diffusion systems [118], these differences may be important. As such, we intend this equation as a tool to investigate our stochastic algorithm rather than a general representation of most dispersal processes.

Note that this is not a spatially discretised version of the integrodifference equation Eq. 6 in Kot *et al* [46]. In that model, a potentially non-linear population growth function is followed by dispersal that is not logistically limited. Integrodifference models of this form are popular in the literature and more closely reflect the life history certain organisms than our models. We therefore confirm our main results in a lattice variant of this integrodifference system in §2.2.4.6.

Dispersal kernels: To represent complex dispersal regimes, we follow Kot *et al* [46] in describing a dispersal kernel. Our models are discrete in space, and the probability of dispersal between sites x and y, over a distance $l = y - x$, is:

$$K(|l|) \propto f(|l|), \qquad l \in \mathbb{Z}, l \neq 0 \tag{2.7}$$

We define $K(0) = 0$ in all cases, in order to avoid an infinite probability of zero-length jump distance for power law kernels.

To represent kernels incorporating long-distance dispersal, we use two classes of function that can lead to fat-tailed distributions:

$$f_a(l) = e^{-|l|^\gamma}, \tag{2.8}$$

$$f_b(l) = |l|^{-\beta}, \tag{2.9}$$

with $l \neq 0$ and $l \in \mathbb{Z}$ in both cases. When functions $\{f_a\}$ or $\{f_b\}$ apply for $|l| \to \infty$, they respectively describe the stretched exponential (if $\gamma < 1.0$) and inverse power law functions. In the real world, short-range behaviour may deviate substantially from these idealised distributions. However, such deviations are unlikely to impact long-time invasion behaviour. For this reason, we also expect that distortions to our dispersal kernels in the model, due to the discrete lattice or constraining $K(0) = 0$, to have minimal impact on qualitative system behaviour.

Our chosen kernel forms offer flexibility in investigating wave of advance behaviour. For functions $\{f_a\}$, $\gamma = 1$ is an exponential distribution and $\gamma = 2$ is a Gaussian. Functions $\{f_b\}$ with $\beta \leq 1 + z$ lack finite moments greater or equal to the z^{th} moment. In one dimension, kernels with $\beta \leq 3$ do not have finite variance, leading to interesting behaviour. These are the kernels leading to Lévy flights.

Having determined the relative probability of a dispersal at each distance, we normalise to obtain the kernel:

$$\sum_{-l_{\max}}^{+l_{\max}} K(|l|) = 1. \tag{2.10}$$

Note that taking the absolute value of l in the above formula corresponds to symmetrical dispersal behaviour in both directions. Although the kernel should theoretically extend to $l_{\max} = \infty$, for practical reasons we define a cut-off distance over which dispersal cannot occur. This is set to $l_{\max} = 2 \times 10^8$, far larger than most of the invasion lattices explored, and can be regarded as a sum out to infinity.

Some caution must be applied here. One needs to have $l_{\max}$ large, but small enough for computer memory considerations. In the case of power-law kernels (2.9) we can check the validity of this approach through an alternative method of constructing the kernel, using the Hurwitz zeta function [119]. This check sometimes indicated an artefactual increase in filling time for very large systems ($L \geq 10^7$), particularly using long range power law kernels, $\beta < 3$. In some of our simulations we investigate the impact of a cut-off to dispersal distance, by reducing $l_{\max}$ to values less than the system size and normalising as above, §2.2.4.5.

Comparison of Models: The above models share various essential features. Specifically, they are discrete in space and time and incorporate dispersal and growth through a logistically limited "budding" process. Dispersal distance is determined according to a dispersal kernel. Initial conditions, the treatment of boundaries, and termination conditions are the same. Finally, in the main body of the work all models ignore population death, other than that implicitly included when population growth is seen as a net process of death and birth. We assess the effects of including death in §2.2.4.6. In many important points the two models above are comparable.

The algorithm for dispersal and birth in our models well-represents the spread of many plants, and combining the two processes follows the approach of other workers (eg. [106] in a population dispersal context, and [95] for modelling epidemic spread). However, the differences between this and the Fisher-Kolmogorov model could lead to difficulties in interpreting our results and situating them historically. In addition to the non-Markovian characteristics of our model, whereby dispersal and reproduction no longer occur independently and concurrently, a particular difference between the Fisher-Kolmogorov model appears in the application of the logistic effect. To investigate this latter point, we give a basic derivation of a diffusive limit of our mean-field equation (2.6) in Appendix 1 (§2.2.6).

This derivation highlights similarities with a diffusion approximation of Mollison's simple epidemic [59], and we find that several analytic results from his model hold when we reduce the spatial and temporal scale of our system, §2.2.4.1. Our application of the logistic effect based on filling at the target site translates to a non-linearity in the diffusion term, see Eq. (A2.8), having greatest impact when the system nears filling. We nevertheless expect our mean-field results to be qualitatively quite general. The Linear Conjecture states that the speed of a wave of advance is governed by the linear properties of the governing partial differential equation far ahead of the front itself, and is thought to apply when the average reproductive rate of an individual is maximised in virgin territory and individuals have negligible impact on the environment far from their current location [120]. These conditions hold for our model. Note that the latter point refers not to long-distance dispersal, but to effects such as local population growth causing global environmental degradation [121].

In support of this argument, we find agreement between our simulations and standard wave number selection methods [122] for determining wave speed, see Fig. 2.2 and Appendix 3 (§2.2.8). We also find that explicit mean-field simulations in which the logistic effect is contributed by the home site yield very similar results to Model 2, although there are interesting deviations in the stochastic case under some long-range dispersal regimes, see §2.2.4.6. As a more general point, the partitioning of the logistic effect between the home range, the target of dispersal, and sites along the route of dispersal technically depends on the life-history one seeks to describe. As such details can impact deterministic and stochastic systems in different ways, there is an argument for considering them when designing models for ecological applications, especially when long-range dispersal is important.

Methodological comments: To investigate the impact of stochasticity on the dispersal process, we perform simulations of the two models for various parameterisations of the two types of dispersal kernel. This allows us to explore a highly stochastic scenario with $N = 1$, a deterministic scenario, and the effects of reducing randomness by increasing N.

To quantify the velocity behaviour of the different dispersal regimes, we use finite size scaling [123]. Here, the size of the lattice is varied over several orders of magnitude, $10^2 \leq L \leq 10^7$ where possible, and the filling time recorded. This method effectively switches the dependent and independent variables, such that we no longer have to follow the wavefront explicitly. The approach allows us to obtain reliable results given noisy systems, in which the wavefront can consist of a large, sometimes widening, region of sparse filling that is difficult to define. The behaviour of filling time as system size is increased is used to identify the relationship between the dispersion kernel and wave velocity and acceleration. Appropriate functional forms are used to interpret results in the different cases.

Finite size scaling: For systems in which invasion takes place through a constant velocity wave of advance, the filling time can be approximated as

$$T \approx \frac{L}{c}, c \approx \frac{L}{T}, \tag{2.11}$$

where L is system size, T is filling time (or average filling time in stochastic systems) and c is the velocity of the wave of advance.

Some accelerating waves can be approximately described by the relation

$$L \approx aT^B, \tag{2.12}$$

with parameter a dominating early time velocity and parameter $B > 1$ describing acceleration. Log-log plots of $\ln T$ against $\ln L$ enable us to approximate the value of these parameters using a linear fitting

$$\ln T \approx \frac{1}{B} \ln L - \frac{1}{B} \ln a. \tag{2.13}$$

In certain systems, rapid exponential acceleration has been observed [46, 97, 93]. In such cases, filling time and system size scale as

$$T \approx g \ln L + h, \tag{2.14}$$

where the parameter h describes early time behaviour and parameter g the acceleration effect. A linear fitting to a semi-log plot of T against $\ln L$ is appropriate for estimating the parameters. This model can be seen as a correction to the non-local dispersal case, where propagules disperse to random sites on the lattice and total system filling, $n(L)$, follows

$$n(L) \approx e^{\alpha t} \quad \Rightarrow \quad t \approx \frac{1}{\alpha} \ln n(L), \tag{2.15}$$

for early times.

These relations represent tools with which we can characterise our idealised systems in the parameter space explored, and do not necessarily correspond to analytically retrievable behaviour.

The relationships of the various variables above to the kernel parameters were estimated using the non-linear regression analysis package GraphPad Prism version 6.0 (GraphPad Software, La Jolla California USA, www.graphpad.com).

Supplementary investigations: We also conducted several additional investigations in order to clarify the behaviour of our simulations and facilitate comparison with the work of other authors:

(a) We have verified the behaviour of our simulation model in several ways. Allowing the model to approach the limits of continuous space and time, we retrieve analytic results for the simple epidemic (§2.2.4.1). Our dispersal kernels yield expected diffusion coefficients, Fig. 2.14, and wave velocity given a fat-tailed kernel in our stochastic system remains within expected bounds, Fig. 2.5, as defined by an approach suggested in Clark *et al* [106].

(b) To assess the implications of the various dispersion kernels, we directly estimated D using a random walk simulation and Eq. (2.2). The relationship between D and wave velocity is explored in Appendix 4, §2.2.9.

(c) We also investigated the impact of stochasticity in birth and death, a stochastic model following the integrodifference equation of Kot *et al*, and a version of Model 1 with the logistic effect acting on propagules but arising at the home site, §2.2.4.6. These models confirm our main results.

(d) To clarify the behaviour of potentially more realistic dispersal scenarios, simulations were conducted using the bivariate Student's t dispersion kernel, as recommended and investigated by Clark *et al* [106], Appendix 6 (§2.2.11). Our main results also hold qualitatively when using different methods of measuring wave velocity, eg. Fig. 2.10, and when applying periodic boundary conditions, Appendix 2 (§2.2.7).

2.2.4 Results

One of our key observations will be that, when long-range dispersal is important, demographic stochasticity gives rise to qualitative differences in wave velocity behaviour, as compared to predictions from integrodifference modelling [46] and our own mean-field model (See Fig. 2.3 and Table 2.1). We find that stochasticity breaks wave acceleration caused by fat-tailed kernels, except in the case of power law kernels with $\beta < 3.0$. This critical point supports some previous results [95] but deviates from expectations based on the behaviour of similar systems in which stochasticity also disrupts acceleration in the Lévy flight case [106, 108].

The difference in wave velocity behaviour between stochastic and deterministic models, which can persist even when N is large (see Figs. 2.6, 2.8), suggests that a realistic representation of demographic stochasticity is important in models of species invasions incorporating long-range dispersal.

2.2.4.1 Testing model behaviour

We applied several tests to confirm that our programs display expected model behaviour. Firstly, we obtained the asymptotic velocity of our mean-field Model 2 for several simple kernels (nearest-neighbour: $K(l = \pm 1) = 0.5$, $K(l \neq \pm 1) = 0.0$; normal distribution; exponential distribution). When $b = 1$ and our standard spatial scale is applied, these deviate somewhat from both Fisher-Kolmogorov expectations and those for the infinite-population limit of the simple epidemic in continuous space and time (see [59], Table 1).

We therefore explored the impact of increasing the resolution of the spatial lattice, by defining our kernels as $K(|l_\varphi|)$, with $l_\varphi = \frac{l}{\varphi}$, $\varphi l \in \mathbb{Z}$, and of the time steps, by reducing b. Here, b serves as a temporal scale rather than a birth rate, so we normalise the resulting velocity according to $\frac{1}{b}$. Results obtained for the simple epidemic model with a nearest neighbour dispersal kernel are approached as $b \to 0.0$, see Fig. 2.2a. For the normal and exponential distributions, analytic results are approximated when $b \ll 1.0$ and $\varphi \gg 1$. These findings are unsurprising, as low-b, large φ systems more closely resemble models that are continuous in time and space, and the simple epidemic model structurally resembles our system.

We also applied a well-known argument based on the marginal stability of the linearised form [77, 122, 124] of the population function far ahead of the wave front, see Appendix

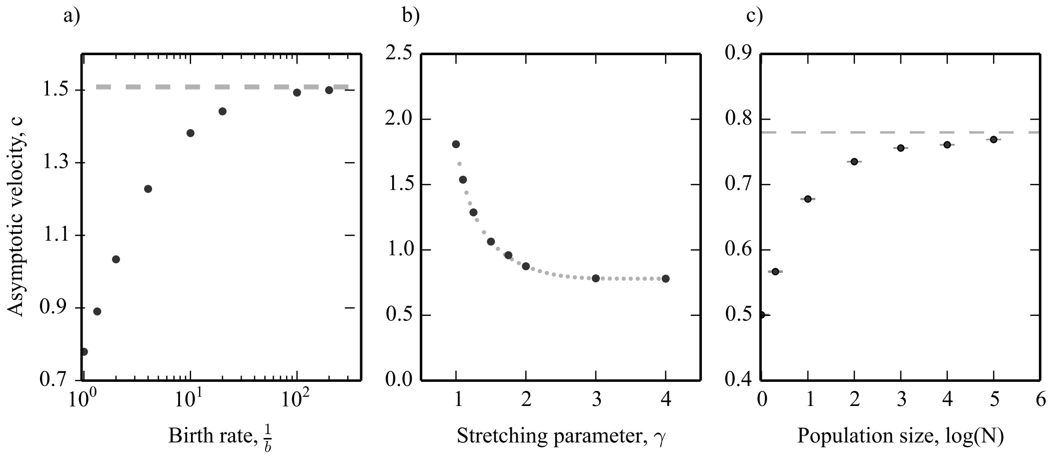


FIGURE 2.2: Testing model behaviour: a) Approach to expected asymptotic velocity for the nearest-neighbour kernel ($D = 0.5$) as birth rate, b, is reduced, with velocity rescaled as described in the main text; b) Agreement between asymptotic velocity obtained by simulation and expected velocity obtained using the marginal stability of the linearised wave front for exponential family kernels, $\gamma \geq 1.0$; c) Approach to mean-field (Model 2) asymptotic velocity as N is increased in Model 1, thereby reducing demographic stochasticity, using the nearest neighbour kernel and 5-20 replicates. Relative standard errors are plotted, but are minimal.

3 (§2.2.8). This allowed us to correctly estimate the wave velocity for the nearest-neighbour ($c \approx 0.78$) and exponential family of kernels ($\gamma \geq 1.0$) in the mean-field Model 2 case, both for $b = 1.0, \varphi = 1$ (see Fig. 2.2b) and when $b < 1.0, \varphi > 1$.

Finally, we confirmed the convergence of wave velocity to the mean-field Model 2 result as stochasticity is reduced by increasing N. We use the nearest-neighbour kernel, Fig. 2.2c, on an lattice of size $L = 10^4$.

The difference between asymptotic wave velocity and the Fisher-Kolmogorov prediction suggests that the simple relationship between diffusion constant and velocity, $c = 2\sqrt{\alpha D}$, may not hold. As results are not critical to this study, we offer a basic investigation of this in Appendix 4, §2.2.9. Briefly, a relationship of the form $c = \mu D^{\rho}$ is apparent, with ρ moderately close to 0.5, for all cases where the kernel leads to a constant velocity wave. However, some deviations exist. These appear to be structural, such that reducing b and increasing φ does not negate them, in agreement with results in Mollison [59] on the simple epidemic. For example, simulations with $b = 0.01$ and $\varphi = 1$ yield $c \approx 2.26D^{0.59}$ for the mean-field model with the exponential family of kernels, $\gamma > 1.0$. We can also quite accurately approximate the impact of discrete time and space using the linearisation method. This allows us to estimate the relationship between c and D for the $b = 0.00001, \varphi = 10^5$ system, which would require lengthy simulations, as $c \approx 2.599D^{0.578} \approx \dfrac{3\sqrt{3}}{2}D^{\frac{1}{\sqrt{3}}}$.

2.2.4.2 Model 1 with $N = 1$: Constant velocity waves in a stochastic lattice system, with a notable exception

In the highly stochastic Model 1, an asymptotically constant wave velocity is generally observed, apparent in the linear scaling of filling time with system size (Figure 2.3a, c). This is usually true even when long-distance dispersal is incorporated through a fat-tailed dispersal kernel, contrasting with predictions from integrodifference equations [46] and our own deterministic model (see below). An accelerating wave of advance is seen, however, for kernels that lack a finite second moment. Approximate velocity behaviours obtained by non-linear regression on filling time behaviour or velocity are given in Table 2.1.

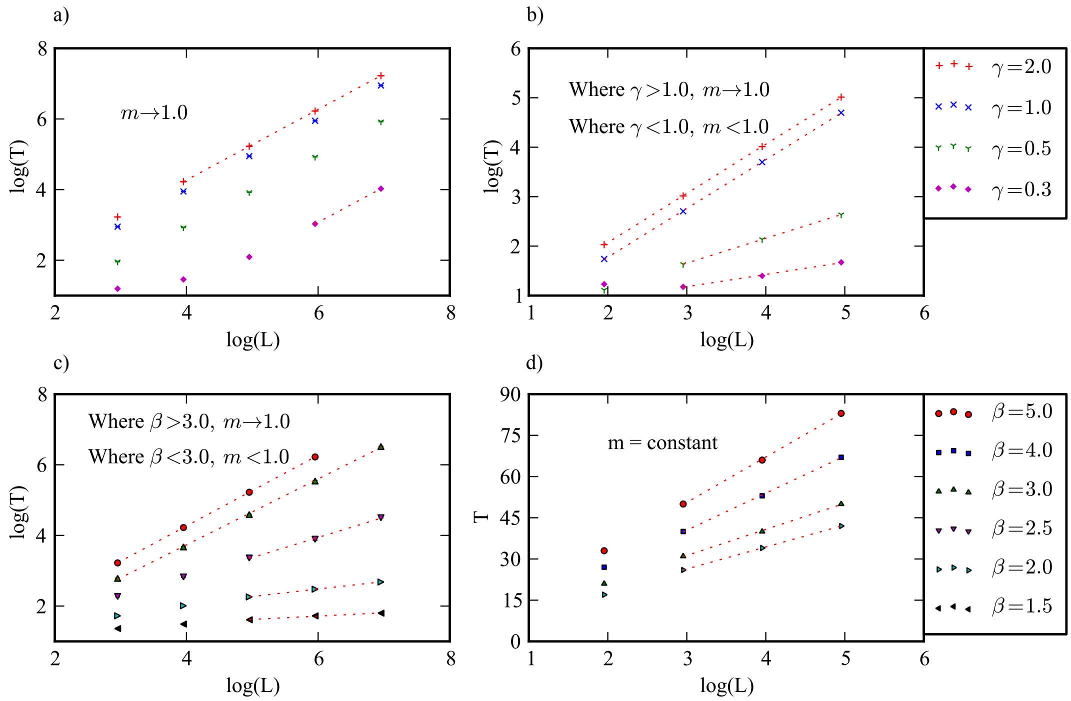


FIGURE 2.3: Log-log/semi-log plots of filling time, T, against system size, L, for the two kernel families following stochastic and mean-field systems. m indicates the gradient of a linear fitting to the data, with $m < 1.0$ indicative of acceleration in the log-log plots. Top-left (a): Stretched exponential kernel stochastic behaviour, $N = 1$; Top-right (b) Stretched exponential kernel mean-field behaviour; Bottom-left (c) Power law kernel stochastic behaviour, $N = 1$; Bottom-right (d) Power law kernel mean-field behaviour. The form of acceleration for fat-tailed kernels in the mean-field systems are in agreement with previous results [46, 97, 93], velocity increasing as a power law and exponentially with time for the stretched exponential and power law kernels respectively. For stochastic models, minimum replicates were: $L = 10^2$, 100; $L = 10^3$, 100; $L = 10^4$, 100; $L = 10^5$, 50; $L = 10^6$, 10; $L = 10^7$, 1.

For the exponential family of kernels, $K(l) \propto e^{-|l|^\gamma}$, a constant velocity wave is observed even when the kernel is fat-tailed ($\gamma < 1.0$). So long as $\gamma > 0$, the diffusivity D remains well-defined. Velocity as determined through our simulations (Fig. 2.4) was exceptionally large when γ is small. Our results can be quite closely approximated using

Kernel form	Model 1 (stochastic, $N = 1$)		Model 2 (mean-field)	
	Equation	Behaviour	Equation	Behaviour
Exponential, $\gamma > 1.0$	$0.5 + (0.52 \pm 0.01)\gamma^{-3.49\pm0.12}$	Asymptotically constant	$0.78 + (1.03 \pm 0.03)\gamma^{-3.23\pm0.19}$	Asymptotically constant
Exponential, $\gamma < 1.0$	Inaccurate fitting	Asymptotically constant	$\frac{(0.94\pm0.03)a_1}{\gamma^{1.10\pm0.05}} t^{\left(\frac{0.94 \pm 0.03}{\gamma^{1.10\pm0.05}}-1\right)} \approx \frac{a_1}{\gamma} t^{(\gamma^{-1}-1)}$	Accelerating
Power law, $\beta > 3.0$	$0.5 + (0.14 \pm 0.01)\beta_1^{-1.09\pm0.06}$	Asymptotically constant	$g^{-1}\exp(\frac{t}{g})$, where	Rapidly accelerating
Power law, $\beta < 3.0$	$a_2(1.0 + B)t^B$, where $B \approx (3.21 \pm 0.26)\beta_2^{1.59\pm0.23}$	Accelerating	$g \approx (1.45 \pm 0.02)\beta - (0.04 \pm 0.16)$	Rapidly accelerating

TABLE 2.1: Table showing wave velocity behaviour given exponential, $K(l) \propto e^{-|l|^\gamma}$, and power law, $K(l) \propto |l|^{-\beta}$, dispersal kernels, as estimated using non-linear regression on finite-size scaling simulation results.These correspond to observed behaviour in our simulations and do not necessarily describe analytically retrievable relationships. $\beta_1 = \beta - 3$ and $\beta_2 = 3 - \beta$, while a_1 and a_2 are prefactors for the velocity behaviour - these tend to be important at early times. We find that $a_1 \approx \frac{0.22 \pm 0.07}{(1.0-\gamma)^{0.96\pm0.16}}$. We could not obtain an accurate fitting for a_2, which was $< 10^{-3}$ when $\beta \leq 2.25$ and very small indeed when $\beta < 2.0$, suggesting an important influence even when t is relatively large. We discuss this briefly in the main text. We fit this regime in the region $2.95 \leq \beta \leq 2.15$. For more details on the parameter ranges for which each fitting was retrieved, see Appendix 4 (§2.2.9) and Fig. 2.15.

the theoretical diffusion constant [125] of the stretched exponential kernel (Appendix 4, §2.2.9), $D \approx \frac{\Gamma\left(\frac{3}{\gamma}\right)}{\Gamma\left(\frac{1}{\gamma}\right)}$ where Γ is a gamma function.

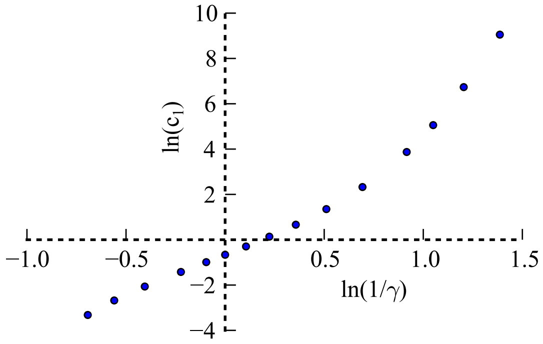


FIGURE 2.4: Logarithmic plot of the dependence of the front velocity on stretched exponential power γ. Note the apparent regime change close to $\gamma \approx 0.5$ ($\ln \frac{1}{\gamma} \approx 0.69$), rather than at $\gamma = 1.0$ as might be expected. The right hand side of the plot corresponds to long-range kernels with $\gamma < 1.0$. Plotting $c_1 = c - c_{\text{min [Model 1]}} = c - 0.5$, rather than the actual velocity c, avoids a spurious saturation in the bottom left quadrant as $c_1 \to 0, \gamma \to \infty$.

Simulations using a power law kernel, $K(l) \propto |l|^{-\beta}$, show a constant velocity when $\beta > 3.0$, Fig. 2.3c. However, when $\beta < 3.0$, accelerated invasion fronts persist, as predicted by [95]. Acceleration is not exponential as in our mean-field Model 2 (see §2.2.4.3). Rather, the linear fit in Fig. 2.3c suggests acceleration occurs as a power law, as in the case of the stretched exponential kernels in Model 2.

Estimates of Lévy flight behaviour [108, 126] yield

$$\left(< x(t)^2 >\right)^{1/2} \approx t^{1/(2-\beta_2)}, \tag{2.16}$$

where $\beta_2 = 3.0 - \beta$. The diffusion constant for such distributions is undefined. However, in a single generation we take a finite number of samples from the kernel, and the variance is finite. We might expect the diffusion constant to increase with system filling, then, implying the acceleration effect that we observe. We were able to estimate an expression for velocity behaviour when β is not too small, $2.95 \geq \beta \geq 2.15$, which supports a strong dependence of velocity on time, see Table 2.1.

When $\beta \leq 2.0$, our attempts to apply non-linear fitting to simulation results failed on two accounts. Firstly, the inferred prefactor to the velocity behaviour, a_2 in Table 2.1, becomes extremely small. Secondly, Akaike's information criterion now suggests that a stretched exponential fit is preferred, $L \approx e^{pT^q}$. It is therefore possible that power-law acceleration no longer adequately describes model behaviour when $\beta \leq 2.0$. However, as these waves travel extremely quickly it was not possible to confirm behaviour over long times.

The power law kernel with $\beta = 3.0$ represents the marginal case between the constant velocity and accelerated front regimes. Given this transition, we expect such invasions to display strong fluctuations, and indeed we observe a sharp peak in the relative standard error for filling time here. An example of filling behaviour in this region can be seen in

Fig. 2.7.

Bounds on the constant velocity of waves under stochastic Model 1, $N = 1$

We can independently confirm the constant velocity results by obtaining estimates of the maximum and minimum wave velocities for each kernel. To do this, we perform a lattice version of the analysis conducted by Clark *et al* [106] for their dispersal model. We describe this in detail in Appendix 5, §2.2.10. Briefly, the approach is as follows. In each generation, we consider the propagule dispersing furthest ahead of the wavefront as the "extreme disperser". This propagule defines the wavefront in the next generation, and the distance it travels indicates the wave velocity that generation. Given this, the maximum wave velocity can be estimated from the distribution of extreme dispersal distances when a large area of contiguous occupation stretches out behind the wavefront. Conversely, this distribution offers the minimum velocity when the wavefront consists of a single isolated occupied site. The asymptotic wave velocity results from our simulations lie between these maximum and minimum bounds, Fig. 2.5.

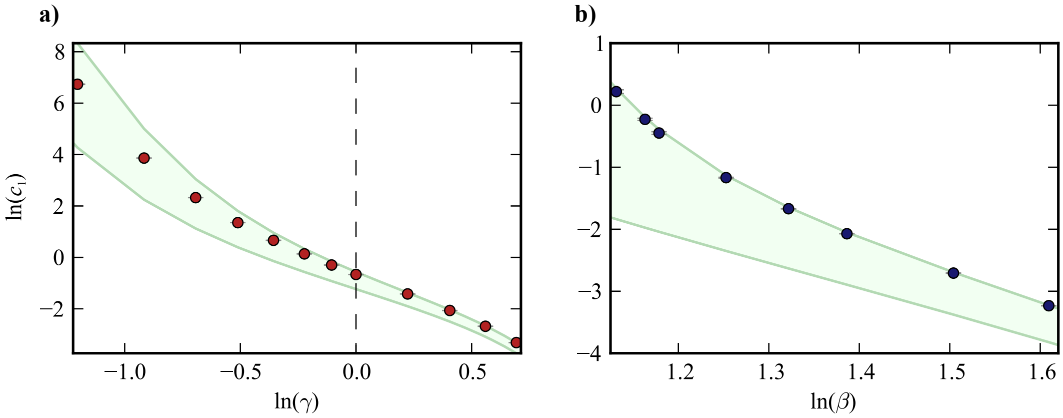


FIGURE 2.5: Placing bounds on velocity for the stochastic Model 1 simulations using a lattice variant of Clark *et al's* method, see Appendix 5 (§2.2.10). Shaded area indicates the region between velocity minimum estimate with $N_s = 1$ and velocity maximum estimate with $N_s = 10^5$, where N_s is the length of contiguous occupation behind the furthest forward site. Kernel length $l_{\max} = 10^5$, which is shorter than in our explicit simulations and may slightly bias upper velocity estimates downwards for the most fat-tailed kernels. a) Exponential family of dispersion kernels, $0.3 \geq \gamma \leq 2.0$, fat-tailed kernels are to the left; b) power law family of dispersion kernels, $\beta > 3$. Circles represent explicit Model 1, $N = 1$, simulation velocity results, with $L = 10^6$, minimum 10 replicates, based on time taken to reach 90% system filling. Other measures of velocity (eg. time until first dispersal to the right-most 10% of the system) give similar results, as do simulations with $L = 10^7$ (performed when $\gamma \leq 0.5$ or $\beta \leq 3.5$). Relative errors are shown, though were small.

Our work indicates that Clark *et al* were correct in their prediction that dispersal kernels without a fat-tail lead to waves that are supported by a large region of occupation to their rear, thus attaining their maximum velocity. While fat-tailed kernels do lead to a more sparsely occupied wavefront, the minimum velocity was not achieved in our simulations. The implication is that this minimum value does not appear to be a good estimate of wave velocity given long-distance dispersal regimes. The importance of dispersal from behind the main wavefront is also supported by more complex models of plant dispersal [127].

2.2.4.3 Model 2: Acceleration for fat-tailed kernel invasions in a mean-field system

In our lattice mean-field simulations, dispersal according to a fat-tailed kernel leads to accelerating waves of advance. Importantly, this agrees with analytic predictions [46] for a spatially continuous version of the system. The signal of acceleration is apparent in the sub-linear scaling of filling time with system size (Figure 2.3b-d). Again, velocity behaviour obtained by non-linear regression is given in Table 2.1.

In the case of short-range kernels, $K(l) \propto e^{-|l|^{\gamma}}$ with $\gamma \geq 1.0$, the wave of advance travels at an asymptotically constant velocity. In the limit $\gamma \rightarrow \infty$, this case corresponds to the nearest-neighbour dispersal model, with $c = c_{\text{min [Model 2]}} \approx 0.78$. For these kernels, we can independently verify simulation results. We do this by numerically obtaining the wave number and corresponding velocity according to eqs. (A3.7) and (A3.6) obtained in our marginal stability analysis of Model 2, Appendix 3 (§2.2.8). This approach yields extremely similar results to our full simulation studies, as shown in Fig. 2.2.

When $\gamma < 1.0$, these kernels become fat-tailed, and acceleration is observed. The form of acceleration follows theoretical expectations quite closely (eg. [84, pp 176]), with filling behaving as $L \approx T^{\frac{1}{\gamma}}$ such that the acceleration $B \approx \gamma^{-1}$. When $\gamma = 0.5$, $B_1 = B - 1 \approx 1.0$, in agreement with the appropriately parameterised Eq.(21) of [46].

Power law kernels, $K(l) \propto |l|^{-\beta}$, lead to waves of advance with extreme accelerating behaviour in the mean-field model. The form of acceleration is exponential, in agreement with previous studies [46, 97, 93], and can be modelled as a correction to random filling Eq. (2.15), using Eq. (2.14) $T \approx g \ln L + h$. We obtain a linear relation, $g \approx \frac{3}{2}\beta$ (see Table 2.1), reflecting the strong accelerating effect even when β is large. Each order of magnitude increase in system size corresponds to a constant increase in filling time. If the continuous model is to be believed, dispersal kernels with this structure would have catastrophic implications in the case of a species invasion. No unusual behaviour is observed at $\beta < 3.0$, as might be expected given the divergence of the diffusion constant in this region.

Theoretical predictions suggest $g \approx \beta$ (eg. [84, pp 171-173]), such that while we recover the exponential form of acceleration, our simulations give a larger value of g.

2.2.4.4 Model 1, $N > 1$: Reducing dispersal stochasticity leads to slow filling time convergence with long-range kernels

As N is increased, stochasticity due to dispersal decreases and we expect to approach the mean-field approximation of Model 2.

In Fig. 2.6 we present finite size scaling results comparing Model 1 with $N = 1, 10, 10^3, 10^5$ to the mean-field Model 2, for three fat-tailed kernels: the stretched exponential kernel with $\gamma = 0.5$ and the power law kernels with $\beta = 2.5,\ 3.5$. The two power laws kernels gave rise to different behaviour in the $N = 1$ stochastic system, but the same exponential acceleration in the mean-field case. The filling time for each N relative to that of the mean-field model is also plotted. This shows the divergence between stochastic and mean-field models over time.

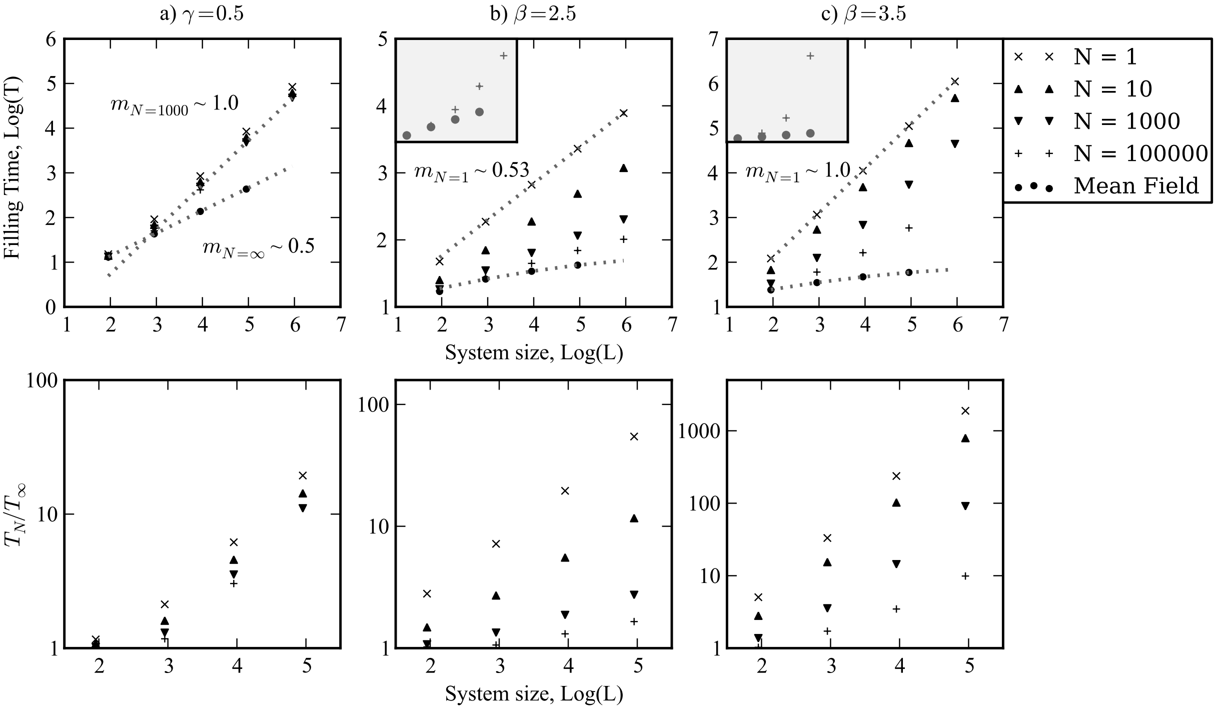


FIGURE 2.6: Comparing the behaviour of Model 1 with $N > 1$ to the mean-field prediction given 3 representative dispersion kernels. Upper plots show finite size scaling behaviour, lower plots the divergence of filling time predicted by each model. Plot a) follows a stretched exponential kernel, plots b) and c) power law kernels. Semi-log plots of T against $\log L$ are inserted in the case of the power law kernels. Minimum replicates were: 10^2: 100; 10^3: 100; 10^4: 50; 10^5: 5; 10^6: 1. Only a single simulation was conducted for the $\beta = 3.5, N = 10^5, L = 10^5$ system; otherwise, more simulations than the minimum number were possible. Relative error bars are shown in the upper plot, but are generally small.

The stretched exponential kernel leads to an accelerating wave in the mean-field case. However, we find an asymptotically constant velocity wave in the highly stochastic $N = 1$

case, and even large N systems rapidly converge to a constant velocity. For the power-law kernels, increasing N leads to more extreme acceleration. When $\beta = 3.5$, transient acceleration persists for a longer period, but waves tend toward the constant velocity behaviour observed in the highly stochastic $N = 1$ system. Qualitatively, acceleration does not reach the extreme mean-field form of $T \approx g \ln L + h$ for either power law explored in our simulations, evident in the semi-log plot inserts in which a linear relationship is not achieved. The implication is that the degree of acceleration predicted for power law kernels using reaction-diffusion equations with fractional diffusion [93], for example, may not persist in real world populations due to stochastic effects.

The difference between stochastic and mean-field results for $T(L)$ increases with system size (see Fig. 2.6, lower plots). This reflects the greater wave acceleration apparent in mean-field systems, such that velocity diverges between the two models over time. We can use this as an indication of how the mean-field approximation deteriorates given the dispersal kernel and N. For small systems, or at early times, $L \approx 10^2$, the filling time results of stochastic and mean-field simulations are similar. For larger systems, eg. $L \geq 10^4$, results can diverge substantially. This is particularly severe for the shorter-range power law kernel ($\beta = 3.5$, $D \approx 1.15$) and the stretched exponential kernel ($\gamma = 0.5$, $D \approx 71.6$). Here, filling is ≈ 91 and ≈ 11 times slower respectively than the mean-field system when $N = 10^3$ and $L = 10^5$.

Figs. 2.7 and 2.8 show the effect of decreasing stochasticity on the structure of the wave of advance and on filling time for a $\beta = 3.0$ power law kernel and a stretched exponential with $\gamma = 0.5$. As N increases, acceleration becomes more apparent at this scale for the power-law system but not for the stretched exponential - the former shows an increasingly concave interface between occupied and unoccupied space, and a substantial reduction in filling time. Acceleration occurs through jumps, with wave structure becoming patchier. This kernel is at the transition between well-defined and infinite variance, so rough and highly variable behaviour is expected on general grounds.

Regions of filling between 30% and 70% for the power law kernel with $N = 10^8$ have been shaded on top of the mean-field filling plot (top rightmost, Fig. 2.7a). Systems with extremely large N approximate the mean-field dynamics closely at early times, but the patchiness created by stochasticity soon reappears.

The impact of increasing N on wave of advance velocity has also been assessed for a short-range kernel (nearest-neighbour, see Fig. 2.2). Even with reasonably small N, wave velocity is close to mean-field predictions of $c \approx 0.78$ ($N = 100$, $c \approx 0.74$), suggesting that the mean-field approximation accurately reflects the underlying stochastic behaviour for very short-range kernels.

Our results indicate that the mean-field model gives a reasonable estimate of stochastic behaviour given short-range kernels, or for fat-tailed kernels at very short times,

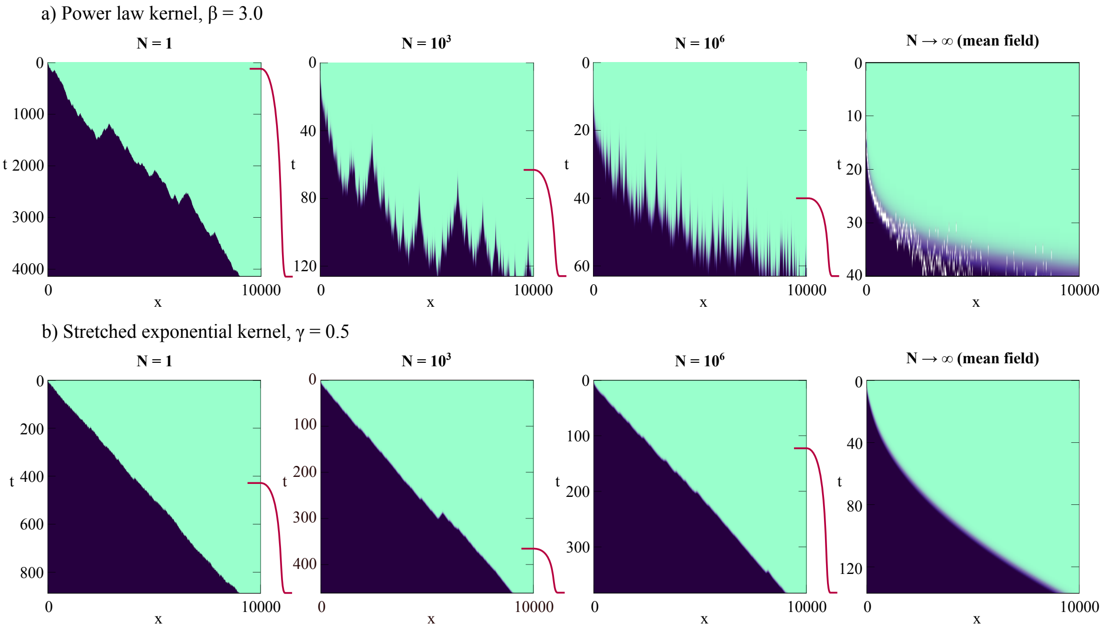


FIGURE 2.7: Plots of the filling processes of individual simulations using Model 1 with a range of N values and two dispersion kernels. Shading represents population density at each position, x, on the 1-dimensional lattice (x-axis). Dark regions corresponds to complete square filling and light regions to empty space. Time progresses from top to bottom on the y-axis, such that systems begin with a single full site at the left-most position $x = 1$. Kernels are labelled accordingly; from left to right, $N = 1, 10^3, 10^6, \infty$ (mean-field). The time scale of each simulation has been rescaled to facilitate comparison of filling dynamics, with relative scale indicated to the right of each plot. The white shading on the upper mean-field plot with a power law kernel at $\beta = 3.0$ represents the filling process for $N = 10^8$, see text. System size, $L = 10^4$.

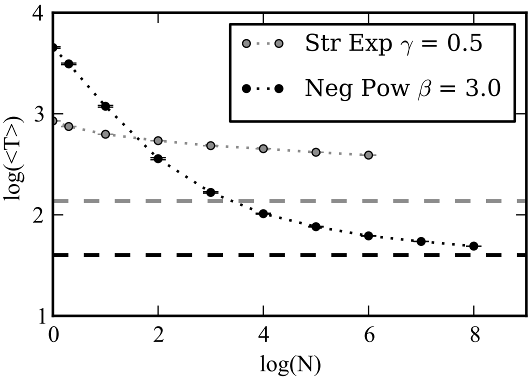


FIGURE 2.8: Convergence of filling time toward the mean-field limit (dashed line) as N is increased for two fat-tailed kernels, $L = 10^4$. Replicates for $N = 1, 2, 10 : 100$; $N = 10^2$ to $10^5 : 50$; $N = 10^6 : 10$; $N = 10^7$ and $10^8 : 1$.

especially if N is large. However, this and similar deterministic models, such as integrodifference equations, tend to over-estimate wave of advance velocity, an error that increases with time when long-distance dispersal is important. Even when filling time is well-predicted by the mean-field approximation, as in some very large N systems, features such as the patchiness of the invasion are poorly described.

2.2.4.5 Truncated power law kernels lead to asymptotically constant wave velocities

Incorporating a long-distance cut-off to power law kernels has been found to describe well the movement patterns of various species [114, 115, 116], and has been investigated in the context of the Fisher-Kolmogorov equation with fractional diffusion [128]. Often, a limit to the distance at which dispersal can occur is also biologically reasonable, due to factors such as energetic constraints or a finite lifespan. It is also possible for such kernels to be retrieved from field data erroneously due to insufficient sampling of rare long-distance events, and an awareness of the errors that this might generate is useful. We here assess the impact of kernel truncation on wave velocity.

A basic prediction is that incorporating a cut-off by reducing $l_{\max}$, see Eq. 2.10, will lead to a slower wave of advance. A cut-off also causes power law kernels with $\beta < 3.0$ to have a defined variance. Asymptotically accelerating waves in either the mean-field or stochastic system are not expected.

We applied a long-distance truncation to power law kernels with various β, and to the stretched exponential case with $\gamma = 0.5$. Behaviour for the power law kernels is shown in Figure 2.9. In the mean-field system, the wave of advance accelerates according to the standard Model 2 until a time t_0. Velocity then approaches a finite value through a series of oscillations. The oscillations may be an example of the Gibbs phenomenon [129] due to the abrupt nature of our cut-off, and if so would not be expected in real systems.

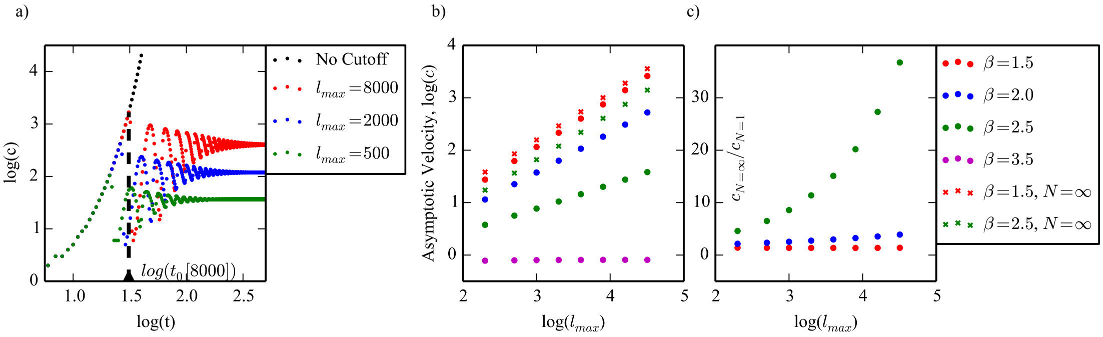


FIGURE 2.9: The impact of a long-distance cut-off on wave velocity given a truncated power law kernel. a) Log-log plot of velocity of wave front over time when $\beta = 2.5$, showing damped oscillations after a time period t_0. b) Log-log plot showing relationship between cutoff, $l_{\max}$ and asymptotic wave velocity, c, for both the stochastic ($N = 1$) and mean-field models. c) Relationship between asymptotic wave velocity reached in the mean-field system and the stochastic $N = 1$ system, for various power laws and sizes of cut-off. Stochastic systems were taken to have reached a stable velocity when the estimated velocity of the wave increased by no more than 1% upon an order of magnitude increase in system size. The stabilisation of mean-field systems was determined visually, with the $l_{\max} = 500$ case providing an example in a), velocity taken as the average filling increase over 10 generations after stabilisation. Replicates for stochastic systems were $L = 10^5 : 100$; $L = 10^6$: 50; $L = 10^7$: 20. Relative standard errors were minimal and are not shown

In the stochastic case, which might be considered a more realistic dispersal model, a finite velocity is again achieved for each dispersal kernel. Unsurprisingly, a cut-off will have particularly dramatic effects for kernels that would otherwise lead to accelerating waves of advance - power laws and stretched exponentials in the deterministic case, and power laws with $\beta < 3.0$ in the stochastic case. For power law kernels in the region $1.25 \leq \beta \leq 2.75$ with $N = 1$ we found that the heuristic fitting $c = a_3 l_{\max}^{(-0.26\pm0.09)(3-\beta)^2+(0.97\pm0.18)(3-\beta)+(0.03\pm0.06)} \approx a_3 l_{\max}^{\frac{-(3-\beta)^2}{4}+(3-\beta)}$ captured asymptotic velocity behaviour. a_3 depends on β and is $0.15 \leq a_3 \leq 0.7$ in this region, with the larger values observed when β is closer to 3.

Asymptotic wave velocity in the mean-field Model 2 can also be fitted to $c = al_{\max}^{B}$, but here B remains close to 0.9 for power laws $1.25 \leq \beta \leq 3.5$ (see Fig. 2.9b) in the parameter space explored. However, as $l_{\max}$ becomes large there are indications of a gradual trend toward $B = 1$. This linear relationship is easily seen for the uniform kernel, which is recovered when $\beta = 0.0$. To further investigate this, we repeated the non-linear regression analysis, this time for $8000 \leq l_{\max} \leq 32000$ and $0.0 \leq \beta \leq 3.5$. The behaviour $c = (0.49 \pm 0.03)e^{-(0.99\pm0.08)\beta} l_{\max} + a_4 \approx 0.5e^{-\beta} l_{\max}$ is supported.

Long-range cut-offs may apply to many biological kernels, and given this the asymptotically stable and substantially reduced difference between wave velocity in the stochastic and mean-field models is of some interest. The mean-field model particularly well approximates the stochastic case given low-β power law kernels with moderately short-range truncation (Fig. 2.9c), and is also quite effective given a stretched exponential kernel with $\gamma = 0.5$ (where $\frac{c_{N=\infty}}{c_{N=1}} \approx 5.2$ when $l_{\max} = 5000$).

2.2.4.6 Structural variations of our model support the generality of results

As mentioned in the introduction, several studies have found that stochasticity breaks wave acceleration induced by Lévy flight dispersal kernels [106, 108]. Given that some of our simulations are seemingly at variance with these results, we have investigated three variations on the stochastic Model 1 presented above:

1. *Case 1* - Our stochastic Model 1 with $d > 0.0, b < 1.0$.

2. *Case 2* - A model following our Model 1 algorithm, but with the logistic effect applied to newborn individuals at the home site rather than the target site. The mean-field approximation is now Eq. (2.17). Note that this system does not result in an advancing wave when $N = 1$. For $N > 1$ and even, we set initial conditions to $n(0, 0), n(1, 0) = 0.5$.

3. *Case 3* - A stochastic version of Kot *et al's* integrodifference model [46]. In each generation, a population growth stage occurs first within each site, such that every

individual reproduces with probability $b[1 - \frac{n(x,t)}{N}]$ and dies with probability d. A dispersal stage then occurs, in which both newborn and older individuals disperse with probability Δ. Usually, $\Delta = 1.0$.

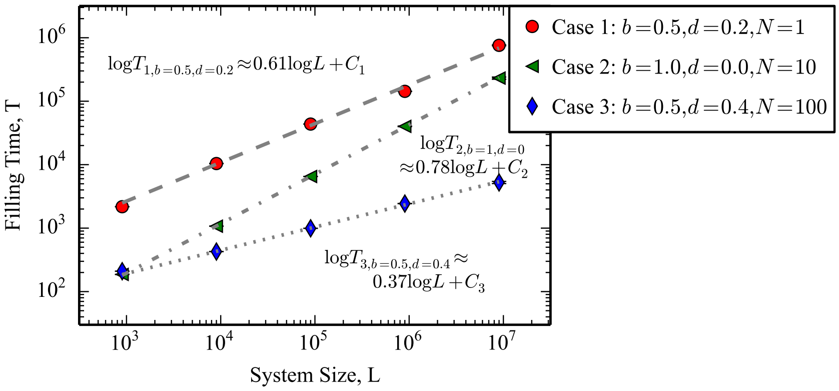


FIGURE 2.10: The impact of stochasticity on three variations of our Model 1, see main text. When dispersal follows a power law with $\beta = 2.5$, long-term acceleration of the wave of advance occurs in each case. Minimum replicates for different systems sizes were $L = 10^3, 10^4$:100; $L = 10^5$: 50; $L = 10^6$: 20; $L = 10^7$: 5. Relative standard errors are plotted, but are small.

In these models, we sometimes incorporate death, $d > 0$. An equilibrium system filling of $n(L) < 1.0$ is possible. We therefore chose to terminate the system when a site in the last 10% of lattice space has filling $n(x > 0.9L, t) \geq 0.1$. This approach is more closely allied with the traditional mathematical approach of tracking the furthest forward position with filling greater than some constant value, $n(x_{\text{Max}}, t) \geq n_{\text{Min}}$. We also tried terminating a Case 1 ($b = 0.5, d = 0.1$) system when system filling exceeds 50%, and found results to be qualitatively similar (not shown).

If the acceleration due to Lévy flights observed in Model 1 is quite general, we would expect it to hold under different implementations of birth and death, and of dispersal and crowding effects. Indeed, we find that acceleration is observed for the $\beta = 2.5$ power law kernel in each model variant (Fig. 2.10), and that velocity always increases as a power law with time. The different acceleration exponents, as indicated by the gradients of linear fittings in Fig. 2.10, are largely due to differing N. However, the structure of the model can also be important - with $N = 100$, the Case 2 system still accelerates comparatively slowly, as long as d is small. We discuss this below. The three model variations also support results from Model 1 for the fat-tailed stretched exponential kernel, $K(l) = e^{-|l|^{\gamma}}$, $\gamma = 0.5$, which creates waves with asymptotically constant velocities (results not shown).

The specific implementation of the crowding effect has interesting implications. In Models 1 and 2, dispersal is logistically limited by occupation at the target site. In Case 2, the logistic effect is applied to new propagules from their home site. The equivalent mean-field equation for Case 2 is

$$n(x,t+1) = (1-d)n(x,t) + (1-d)b \sum_{l=-\infty}^{+\infty} K(|l|)n(x+l,t)[1-n(x+l,t)]. \tag{2.17}$$

Simulations using this equation and the Model 2 mean-field equation, Eq. (2.6), lead to very similar results, even when long-range dispersal is important. The difference between filling times for a system of size $L = 10^5$ was less than 1% for stretched exponential kernels with $\gamma = 2.0, 1.0, 0.5$, and $\approx 2\%$ and $\approx 3\%$ for the power law kernels with $\beta = 2.5, 3.5$ respectively.

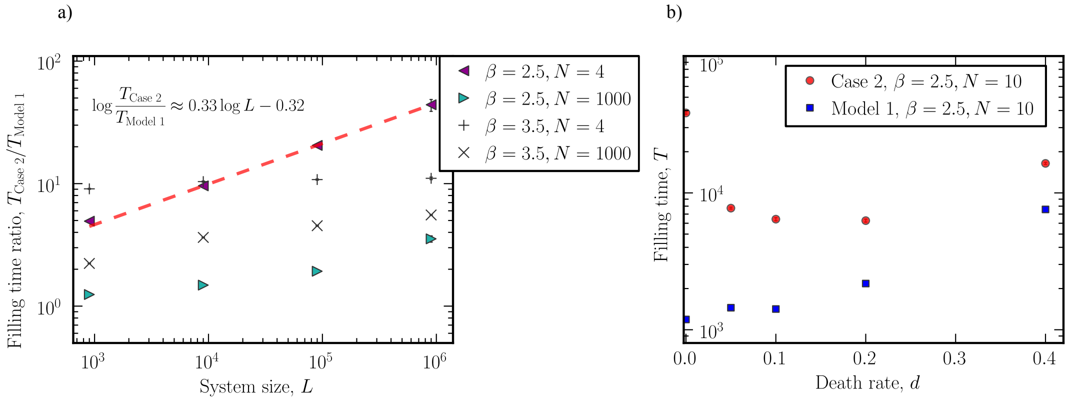


FIGURE 2.11: When only newborn individuals disperse in a stochastic system, wave velocity and acceleration depend on whether the logistic effect arises at the home site (Case 2, see text) or at the site targeted for dispersal (Model 1). This is particularly the case for low-d, small-N systems incorporating long range dispersal. a) The ratio of Case 2 filling time to Model 1 filling time as system size L is varied, with no death. Larger values indicate greater wave velocity deviations between the two systems. The ratio converges toward a constant value when the kernel leads to a finite-velocity wave, as for $\beta = 3.5$. b) Faster system filling occurs when $d > 0$ for the Case 2 system, $L = 10^6$. Minimum replicates are $L = 10^3 : 100$; $L = 10^4 : 50$; $L = 10^5 : 20$; $L = 10^6 : 5$.

However, significant differences appear when stochasticity is introduced. Here, full sites frequently occur, but cannot contribute to population growth. Understandably, this effect is particularly pronounced when $d = 0$ and N is small. For example, when $N = 4$ and dispersal follows a stretched exponential kernel ($\gamma = 0.5$), the Case 2 model leads to an asymptotic wave velocity of approximately $\frac{2}{3}$ that seen for Model 1. The effect is especially striking in the case of power law kernels, and the deviation between filling time results for Model 1 and Case 2 as L increases is shown in Fig. 2.11a.

The impact of increasing death, $d > 0.0$, when the logistic effect arises at the home site is shown in Fig. 2.11b. Dispersal from sites that are far behind the main front and close to equilibrium filling is again possible, such that incorporating some death can increase wave velocity.

2.2.5 Discussion

Conclusions of our study

When a population of an invasive species colonises a new region, it can grow and spread. This creates a wave of advance, which travels across the landscape with potential impacts on agriculture and natural ecosystems. Understanding the dynamics of invasions can help workers predict disruption and orchestrate a response. Simulation and mathematical models provide one route toward this understanding. In this study, we have identified important consequences of different modelling approaches, finding that stochasticity implied by a finite population fundamentally alters the behaviour of the wave of advance when long-distance dispersal is important. Results obtained by deterministic methods such as integrodifference equations or models of fractional diffusion risk significant inaccuracies in such circumstances, Figs. 2.6, 2.10.

Our simulations involve a simple model of population dispersal with tuneable demographic stochasticity. The models incorporate a flexible dispersal regime, with a logistic effect limiting the dispersal/growth process. In this way, the structure is similar to that of Mollison's simple epidemic [95], but occurs in discrete time, and indeed we confirm several of his fundamental results (for example, the different conditions on accelerating waves for stochastic and deterministic models). Some of these results appear to contradict, at first glance, recent work on stochastic models that incorporate long-range dispersal [106, 108]. We continue by discussing this discrepancy, after reviewing the core assumptions and detailed features of our simulation framework.

Dimensionality

Our models are one dimensional, a feature that is biologically plausible for the spread of species along rivers, coastal strips and island chains, but does not represent the general case. This approach is not unusual with most work on population spread historically conducted in one dimension for the sake of simplicity. Although for linear models the relationship between high-dimensional behaviour and the one-dimensional case is often straightforward, this is not true for non-linear stochastic models [121] such as ours. In the context of our work, there exists an obvious situation where dimensionality is important, specifically when dispersal occurs according to a power-law. In this case,

$$D = \int l^2 d^{d_I} l K(l) \sim \int_{r_{\min}}^{\lambda} l^2 l^{d_I - 1 - \beta} \mathrm{d}l = \int_{r_{\min}}^{\lambda} \frac{l^{d_I + 1}}{l^{\beta}} \mathrm{d}l \tag{2.18}$$

with d_I the dimensionality. We take a lower limit $r_{\min}$, defined so as to avoid divergence at small scales, and approximate the kernel using its power-law decay, which dominates as $\lambda \to \infty$. This integral describes the leading order behaviour of D as a function of λ, and if it diverges D does not exist. Given our results, stochastic Model 1 is expected to produce accelerating waves when $\beta \leq d_I + 2$.

To check whether the behaviour implicit in (2.18) is observed, we conducted limited simulations on a triangular lattice in two-dimensions. Our current implementation of the explicit simulation was impractical for lattices with sides of length greater than 10^4, i.e. 10^8 sites, in Model 1, and considerably smaller systems in the mean-field Model 2. Nevertheless, accelerating waves were seen in the mean-field system with fat-tailed dispersal kernels. This acceleration was found to break down when stochasticity was added, with the exception of systems with certain power-law kernels. We were unable to confidently determine the critical value of β under which acceleration is preserved despite stochastic dispersal. For example, acceleration was not observed in the highly stochastic Model 1 system when $\beta = 6.0$, but was for the limited range of lattice sizes explored when $\beta = 5.0$. Equation (2.18) suggests that this is likely to be a transient feature.

An example of a similar system explored in two dimensions is the model of Kawasaki *et al* [109]. Here, the authors compare a stochastic cellular automaton model in which dispersal occurs randomly at a given rate with a deterministic variant. In both cases, site occupation is limited to 0 or 1. The deterministic case represents dispersal as a constant pressure exerted by occupied sites on empty sites within their dispersal range. Colonisation occurs when the cumulative dispersal pressure experienced by an empty site exceeds 1. The stochastic version of the model, which resembles our Model 1 with $N = 1$, yields a higher wave velocity than the deterministic case, even with the nearest-neighbour dispersal kernel. This is a result of the rougher wavefront for stochastic systems, such that there are more isolated sites, and the greater probability of propagules from isolated sites being successful.

Although this effect is in principle possible in our system, we did not observe greater wave speeds for our Model 1 than Model 2 in two dimensions. This is probably due to the potential for sites to have filling $0.0 \leq n(x,t) \leq 1.0$ in our deterministic model, such that newly arriving propagules immediately begin to contribute to the filling of nearby sites. There is a time-delay in the deterministic system of Kawasaki *et al*, as a site has to wait for sufficient dispersal pressure to transition from state 0 to 1. Despite this difference, the wavefronts in our stochastic Model 1 are rougher than those in the deterministic system. It is possible that both features are relevant, and a detailed characterisation of the behaviour of our model in two dimensions represents an important future step.

Modelling assumptions

A particular structural feature of our models is discreteness in time and space. Biologically, seasonal reproduction is seen in many species, and discrete-time models are often desirable. While one could consider discrete space to represent minimum territory size or regular fragmentation of a habitat, a more general interpretation views this as an artificially imposed lattice that averages effects within regions. This is a common approach used to simplify models, but can create an artefactual population crowding effect [130]. The lattice will also distort the dispersal kernel. Our replication of several forms of wave velocity behaviours caused by long-range dispersal [95, 46, 97, 93] suggests that our lattice dispersal kernels approximate sufficiently their continuous forms. Indeed, at long distances distortion to the dispersal kernel is minimal as the rate of tail decay is low.

A core subject of this paper is the relationship between deterministic and stochastic models of dispersal. We investigate the relationship between the stochastic Model 1 and a mean-field Model 2 by increasing the carrying capacity of sites, N. When $b = 1$ and $d = 0$, this only reduces the amount of stochasticity associated with the dispersal process. Model 2 is a mean-field model that offers a heuristically appealing approximation of Model 1, but we have not formally shown it to be correct. Indeed, in one case - that of stretched exponential kernels - the approach to mean-field wave velocity behaviour as N is increased is particularly slow, Fig. 2.8. While this could indicate that another formulation is more appropriate, radically different behaviour in stochastic and mean-field systems for these dispersal kernels is theoretically expected [95].

Increasing N in Model 1 also leads to a second transition, in that the filling impact of each propagule decreases. Biologically, this corresponds to the varied demographic dynamics of different species. Low N systems resemble the dispersal of larger organisms, such as trees with few viable seeds per generation or various mammals. When N is larger, organisms such as insects or micro-organisms are better described. Long-distance dispersal is possible for both, with the latter case often relying on aerial or aquatic currents, or living vectors (e.g. [131, 132]).

Principle characteristics of model behaviour

In this work, superficially similar models are found to yield qualitatively different conclusions. A fundamental difference in wave velocity behaviour is apparent between the stochastic and deterministic versions of our system. This sort of phenomenon has been noted by other workers in the context of long-range dispersal [95, 108]. The implication is that care is needed when designing models of species dispersal, and the consequences of mathematical or computational simplifications should be explored where possible.

Our simulations support the observation that fat-tailed kernels can give rise to accelerating waves of advance in deterministic models of population spread [95, 46, 97, 93]. The stochastic Model 1 also supports aspects of previous work [95]. Here, indefinite acceleration is seen to break down for some fat-tailed kernels. However, this is not always the case. Reflecting some previous results [95], but not others [106, 108], we find that kernels described by a power law with $\beta < 3$, which lack a finite second moment, can lead to long-term acceleration. Random walks according to similar distributions lead to Lévy flight superdiffusion. Our Model 1 is more complex, with a non-traditional branching rule and density-dependent dispersal. These differences do not appear to break the fundamental Lévy-flight-like dynamics. The apparent disagreement between this result and behaviour seen in similar systems [106, 108] is interesting, and we discuss it in detail below.

In Fig. 2.6 we examine the time evolution of the difference between our mean-field and stochastic systems. The mean-field Model 2 gives a reasonable estimate of the short-time wave velocity. However, at longer times the wave accelerates rapidly away from that predicted by the stochastic model. This effect is disrupted when a cut-off is applied to the kernel, §2.2.4.5, and in some cases (eg. low-β power law kernels, see Fig. 2.9c) the velocities achieved for the mean-field and stochastic systems are rendered similar.

Acceleration observed in mean-field models seems to be caused by the logistic growth in low-occupation sites well ahead of the travelling wave. Demographic stochasticity usually disrupts this. However, when the dispersion kernel lacks a second moment, we regain acceleration. We speculate that this is a result of a superdiffusive effect from behind the main front. The advance of the wave is no longer strongly coupled with occupation close to the main front. This leads both to a patchy front and to a velocity that increases with total system occupation.

Should such conditions arise in the real world? In our introduction §2.2.2 we noted both that long-distance dispersal and accelerating waves of advance have been separately observed. Laying these points aside, we recall the extensive theoretical work on the Lévy flight foraging hypothesis [94]. When dispersal to new sites of reproduction is an extension of the foraging process, this hypothesis would suggest that the power law dispersion kernels are expected. Even when the spread of organisms to new territories is clearly separated from foraging, the potential to 'get lost' while foraging may facilitate rare long-distance dispersal events that would be expected to eventually dominate system behaviour. Anomalous kernels that resemble those explored above seem possible, although an eventual cut-off is inevitable in most real-world situations.

Relationship to other work

The detailed form of reproduction

Our observation of an accelerating wave of advance in a stochastic system due to power-law dispersal with $\beta < 3.0$ appears to contradict the results of Brockmann and Hufnagel [108]. In this study, the authors found that stochasticity disrupts Lévy flight superdiffusion in a two-particle reaction-dispersal system. Briefly, the model consists of a space containing a large number of particles of type A and B. Both particles disperse by Lévy flights, and can also react according to $A_x + B_x \xrightarrow{k_1} 2A_x$ and $A_x + B_x \xrightarrow{k_2} 2B_x$. Our system is similar in having two particle states, 'empty' and 'full', and a transition of empty to full space due to population growth. However, in contrast, fluctuations do not tame superdiffusion caused by Lévy flights in our study.

We can suggest several reasons as to why this may be the case. The constant filling rate behaviour of the Brockmann-Hufnagel system is interpreted as being due to the probability of absorption outweighing local exponential growth when particles are rare. We see parallels between this and an emergent Allee effect. Given this, any critical difference between the models is likely to impact population growth in the low-population regime. The simple possibility that contrasting results relate to explicitly including stochasticity in birth and death is not supported, Fig. 2.10.

One fundamental difference between all our models and that of Brockmann and Hufnagel is that their system is Markovian, with reproduction and dispersal occurring independently and concurrently. The detailed manner in which dispersal and growth are combined in our system might explain the difference in behaviour. For example, an $A_x + B_x \rightarrow 2A_x$ reaction event in [108] leads to a local reduction in per-particle reaction rate, as in the Fisher-Kolmogorov equation, impacting both the parent and new-born organism. In our model, dispersal and growth are combined, such that a new propagule far ahead of the main front is likely to move to an unoccupied site. Neither parent nor new-born experiences a direct reduction in growth rate, although this is implicitly, and stochastically, implemented by local crowding. If such differences are critical, the implications for modelling long-range dispersal are quite profound, in that apparent subtleties of reproductive behaviour might lead to highly variable patterns of population spread. We certainly do not show this to be the case, however, and believe that further modelling or analysis of the two forms of system is necessary to clarify the importance of this effect.

Other explanations are plausible. The potential for both particles to disperse with Lévy flights in the Brockmann-Hufnagel system facilitates long-range counter-invasions. Alternatively, the acceleration in our systems may merely reflect an extremely long transient effect. This final point seems unlikely given the large systems explored and superdiffusive behaviour of Lévy flight random walks. Although acceleration sometimes appears to slow down slightly when $L \geq 10^7$, this is likely a consequence of our constructing the kernel to a finite distance $l_{\max} = 2{*}10^8$. Explicitly calculating the dispersal distance using the Hurwitz Zeta function appears to remove the effect.

The extreme disperser approximation

In the population dispersal literature, the model of Clark *et al* [106] resembles our simulations. The authors suggest that even dispersal kernels without a finite variance do not lead to indefinite acceleration, apparently at variance with our results (Figs. 2.3, 2.10).

The argument of ref. [106] focusses on 'extreme dispersers'. An occupied territory consists of one or more of organisms, each producing on average R_0 dispersing propagules per generation. The extreme disperser is the propagule that travels furthest ahead of the wavefront in a generation, and defines the wavefront in the next generation. Based on the increasing patchiness of occupied areas when dispersal is fat-tailed, the authors raise the possibility of a transition between two regimes. Initially, the wave of advance moves at the edge of a region of high population density. After a long period of time, the furthest forward individual in generation $t+1$ is merely the extreme disperser of the furthest forward individual at time t. The former case leads to the maximum wave velocity, while the latter approximates the minimum wave velocity. The authors suggest that this transition, combined with the fact that a sample of finite size from a kernel with infinite variance has finite variance, may lead such kernels to create constant velocity waves of advance.

We find that, for fat-tailed kernels with a finite variance, the minimum velocity approximation is not particularly effective in our system, §2.2.4.2. Furthermore, for Model 1 with $N = 1$, our initial condition ($n(0,0) = 1.0$) of a single isolated occupied site represents the minimum velocity case, and transient acceleration is nevertheless observed.

Clark *et al* [106] carry out explicit simulations of invasion dominated by extreme dispersal in the case of a bivariate Student's t kernel [133]. This kernel is fat-tailed and lacks a well-defined second moment,

$$K(l) = \frac{1}{2\sqrt{2u}\left(1+\dfrac{l^2}{2u}\right)^{\frac{3}{2}}}, \tag{2.19}$$

where u is a scale parameter for the distribution. This kernel closely resembles a power law at $\beta = 3$ over long distances, but unlike all our fat-tailed kernels it is convex at its source. Importantly, this power law represents the boundary case of undefined variance in one dimension, and we expect acceleration caused by such kernels to be marginal and difficult to detect. Indeed, we have performed simulations using the bivariate Student's t kernel to assess the importance of convexity at source, and results closely reflect the power law at $\beta = 3$. Specifically, the wave accelerates for some time, but appears to approach a constant velocity eventually (for details, see Appendix 6, §2.2.11). The precise from of the short-range dispersal regime does impact the length of transient acceleration, but does not change the long-term dynamics. For example, the time taken for acceleration to cease is greater when parameter u is large, as is the case for the species of spruce (*Picea*) modelled by Clark *et al*, where $u = 5531$.

We therefore suspect that the constant velocity behaviour suggested by Clark *et al* for the bivariate Student's t kernel reflects the fact that it is the boundary case for accelerating waves (i.e. $\lim_{l \to \infty} K(l) \propto |l|^{-3}$), rather than the finite number of dispersal events per generation or factors such as its short-range behaviour. Acceleration may nevertheless be visible when using different methods of data analysis. We expect that further work will resolve behaviour caused by such marginal dispersal kernels.

Our work suggests that the method used by Clark *et al* would give unreliable results for kernels with an undefined second moment that are non-marginal, as these are expected to accelerate due to the dispersal impact of sites far behind the main front. However, we note that the framework provides very accurate estimates of velocity behaviour for waves generated by power law kernels with $\beta > 3$, Fig. 2.5, using the maximum rather than minimum velocity approximation. The method is also reasonably effective for stretched exponential kernels when γ is not too far below 1.0, and it therefore remains very interesting in these two cases. It is also certainly possible that further refinement to the approach will allow the recovery of the velocity given by low-γ stretched exponential kernels, or the acceleration behaviour caused by power law kernels with $\beta < 3.0$.

Extending our model

As discussed in §2.2.3, our method of combining birth and dispersal leads to significant differences from the traditional Fisher-Kolmogorov equation. Whether one interprets our approach as the release of dispersing propagules, or as the dispersal of adults after their generation-long maturation and local reproduction, our algorithm does not accurately represent the behaviour of certain organisms. Given this, we performed several extensions to our model, §2.2.4.6.

These modifications consisted of incorporating stochasticity in birth and death ($b > 0$, $d < 1$), applying the logistic effect at the home site as in the Fisher-Kolmogorov equation, and implementing a stochastic version of the integrodifference equation studied by Kot *et al* [46], Eq. (2.5). In each case, acceleration of the wave of advance was observed in the stochastic system given Lévy flight dispersal. These variations and re-interpretations of our model suggest our results are, qualitatively, quite general.

The detailed implementation of the logistic effect has interesting modelling implications. When only propagules disperse and these are subject to intra-specific competition, applying the logistic effect at the home site (Case 2 in §2.2.4.6), as is traditional in population dispersal modelling [86, 46], can cause the wave of advance to stop before complete filling. This occurs when all occupied sites are completely full, and is inevitable when $N = 1$. If $N > 1$, the wave of advance should not stop permanently unless $d = 0$. When $d > 0$, transient pauses are still possible, and are more likely when N is small, dispersal is short-range and $d \ll 1.0$.

With respect to long-range power-law dispersal ($\beta < 3.0$), the rate of acceleration tends to be substantially lower for the Case 2 model than in Model 1. The significant influence on the wave of advance from sites behind the main front is damped, and we might expect velocity to scale with the total number of partially-full sites. For this reason, death has an unusual effect in such systems, and the maximum wave speed is sometimes achieved when $d > 0$, Fig. 2.11. This behaviour is interesting from a modelling perspective, and although the Case 2 system is somewhat contrived (particularly when $d = 0$) the effect may in principle play a role in real species invasions. Relevant situations include cases in which crowding has a significant impact on parental reproductive potential, or where newborn offspring experience intense intra-specific competition before dispersal from the home site. Slight changes to the stochastic algorithm, such as allowing reproductive individuals to disperse, are likely to have a significant impact on the behaviour of this model.

Final Remarks

Taken together, our results suggest that the modelling strategy employed should depend strongly on the dispersal regime under consideration, the time-scale of interest, and the life-history details of the organism in question. In general, methods that poorly represent the long-distance tail of dispersal regimes, such as integrodifference equations, should be used with caution when these regions determine wave behaviour. Such methods are frequently encountered in the literature, and are often applied specifically for their ability to include long-range dispersal kernels. There is a danger here of using sophisticated mathematics to derive qualitatively incorrect conclusions. Our work helps to clarify the conditions under which random effects due to demographic stochasticity are important, and the severity of errors expected when they are ignored.

Comparing our results to other work highlights the potential for subtleties in model design to create apparently contradictory system behaviours. This does not necessarily reduce the validity of the different approaches. However, which method is appropriate will depend critically on the real-world scenario one is seeking to explore. Conclusions for one field or problem may not translate simply to other applications; and, perhaps worryingly, it can take structural investigation of a model rather than basic parameter sweeps to identify this.

Note added in proof: After this paper had been submitted, we became aware of recent work by Hallatschek and Fisher [134] and by Chatterjee and Dey [135] which addresses mathematically the impact of power law dispersal on stochastic system behaviour. These studies are of especial relevance to our simulation results in §2.2.4.2.

Acknowledgements

We would like to thank Lee Hazelwood (University of Cambridge), Anne Kandler (City University) and James Steele (UCL) for important input at the beginning of this project. We also appreciate helpful comments from Dirk Brockmann (Humbolt University), and from Tim Reluga (Penn State) and two anonymous reviewers.

Supplementary Information

Long-range kernels, stochasticity and the broken accelerating wave of advance

G. S. Jacobs and T. J. Sluckin

2.2.6 Appendix 1: Simple derivation of a diffusive limit for mean-field Model 2

This appendix is included as Thesis Appendix A2 at the end of the thesis.

2.2.7 Appendix 2: Periodic boundary conditions

How one treats boundary conditions is a methodological question encountered in many modelling scenarios. Frequently, periodic boundary conditions are applied, which can make a system 'neater' from a physical perspective. However, in explicit simulation particularly, we have flexibility in this regard, and conditions that greater resemble the scenario being modelled are appropriate. We have chosen to ignore dispersal outside the system in the majority of our simulations, which is reasonable if we are modelling invasion of a stretch of viable habitat along a coastline or river. Periodic boundary conditions better resemble the coast of an island for which the entirety of its coast is habitable for a species, but dispersal cannot occur over the main landmass.

To consider this case, and for modelling completeness, we here briefly present results for the highly stochastic Model 1 system with periodic boundary conditions. Our essential results for the Model 1 stochastic system are preserved in the case of periodic boundary conditions, with wave acceleration only observed for power law kernels with $\beta < 3.0$.

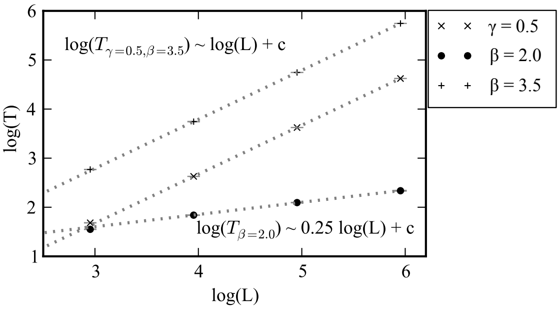


FIGURE 2.12: Filling time behaviour for Model 1 with periodic boundary conditions. Constant velocity is suggested for waves arising from both the stretched exponential kernel, $\gamma = 0.5$, and the power law kernel with $\beta = 3.5$. Acceleration is apparent for the power law kernel with $\beta = 2.0$, as evident in the log-log gradient of filling time against system size < 1.0. The acceleration parameter of 0.25 is estimated as similar to that of our main simulations (~ 0.22).

Under periodic boundary conditions, long-range dispersal at distances far greater than the system size roughly equates to random positioning of the propagule after a large number of 'laps' of the lattice, creating the possibility of unrealistic dispersal events for smaller system sizes.

2.2.8 Appendix 3: Marginal stability analysis of Model 2 with a nearest-neighbour kernel

This appendix is included as Thesis Appendix A3 at the end of the thesis.

2.2.9 Appendix 4: Clarifying kernel behaviour

The diffusion constant, D, is the traditional variable used to describe dispersal in the Fisher-Kolmogorov equation Eq. (2.1). This value was determined for various kernel parameterisations using the root mean square displacement of five thousand 50,000-step random walks and Eq. (2.2). The accepted D was the average of 50 repetitions of this process. Velocity for constant speed systems was retrieved from the filling time of systems with $L = 10^6$, using Eq. (2.11).

Having obtained values for D and c for different kernel parameters, we can estimate the relationship between these quantities using non-linear regression, Fig. 2.13. This identifies slight departures from the $c \approx 2\sqrt{\alpha D}$ relationship, Eq. (2.4), derived from the Fisher-Kolmogorov equation. Given that a diffusion approximation of our model shows important differences from the Fisher-Kolmogorov equation, see Appendix 1 (§2.2.6), this is not surprising. A better point of comparison is the simple epidemic explored by Mollison [95, 59], which is very similar to our model but is continuous in space and time, and again does not adhere to Eq. (2.4). Indeed, by increasing the spatial and temporal resolution of our model, we are able to retrieve velocity results derived for the simple epidemic. The implication is that the relationships we find between D and c reflect both the structure of our model and discretisation effects.

We approximate c for a continuous-time, continuous-space model by applying the linearisation method detailed in Appendix 3, §2.2.8. This time, b is reduced to 10^{-5} so as to substantially increase the temporal resolution of the system. The spatial detail is increased by using the modified kernel $K(\frac{l}{\phi}) = e^{-|\frac{l}{\phi}|^{\gamma}}$, $l \in \mathbb{Z}$, with $\phi = 10^5$. The value of D in continuous space is approximated using this kernel and the method described above. For the deterministic approximation of the simple epidemic, an exponential kernel is known to lead to a wave of advance with $c \approx \frac{3\sqrt{3}}{2}D^{\frac{1}{2}}$, while normal kernels give $c \approx \sqrt{2De}$ [59]. These results are retrieved almost exactly, Table 2.2. Finally, we can explore the behaviour of c given D for our exponential family of kernels, $\gamma \geq 1$, by applying non-linear regression on our results. We identify the relationship $c \approx 2.5992D^{0.5777} \approx \frac{3\sqrt{3}}{2}D^{\frac{1}{\sqrt{3}}}$ for this approximation of a continuous-space, continuous-time system.

We have also checked the accuracy of D, as estimated through the random walk simulations. This involved obtaining the values of the diffusion constant for power laws with

Time scale (b)		1	10^{-5}	1	10^{-5}	D		Wave velocity estimates	
Lattice scale ($\frac{1}{\varphi}$)		1	1	10^{-5}	10^{-5}	Discrete	Continuous	FKPP	Simple Epidemic
$\gamma =$	2,	0.875	1.645	0.654	**1.166**	0.572	0.25	1	**1.165**
	1.75,	0.946	1.757	0.721	1.276	0.638	0.290	1.078	
	1.5,	1.064	1.948	0.833	1.458	0.763	0.370	1.217	
	1.25,	1.288	2.309	1.045	1.802	1.019	0.530	1.457	
	1.1,	1.538	2.710	1.283	2.185	1.335	0.739	1.719	
	1.0	1.809	3.142	1.544	**2.598**	1.715	1	2	**2.598**
$c = \mu D^{\rho}$,	μ	1.2684	2.2859	1.5452	2.5992				
	ρ	0.6565	0.5891	0.6184	0.5777				

Table 2.2: Relationship between wave velocity and diffusion constant for mean-field Model 2 given a stretched exponential dispersal kernel, $K(\frac{l}{\varphi}) = e^{-|\frac{l}{\varphi}|^{\gamma}}$, $l \in \mathbb{Z}$. The spatial scale is given by $\frac{1}{\varphi}$ and temporal scale incorporated as the birth rate, b. Low values correspond to a fine scale in both cases. Death is ignored in each case, such that b is essentially the traditional Malthusian parameter, $\alpha = b - d$. D and c were estimated as described in the main text. Velocity predictions derived from the Fisher-Kolmogorov equation follow the formula $c = 2\sqrt{\alpha D}$, while those according to the simple epidemic use formula given in [59]. The dependence of c on D was estimated for each system by non-linear regression on log-transformed data. When $b, \varphi = 1$ this relationship corresponds well with that obtained through explicit simulations, Fig. 2.13.

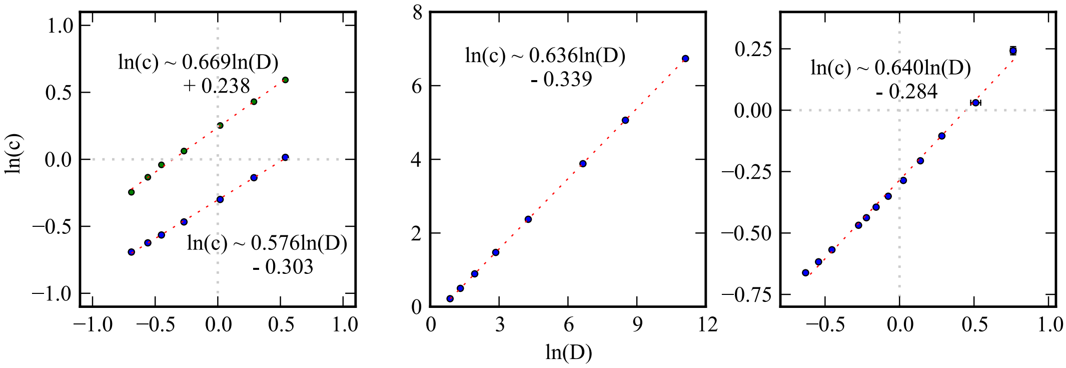


Figure 2.13: Relationships between diffusivity, D, and asymptotically constant wave velocity, c, under various dispersal kernels. For stochastic systems, $N = 1$. Left - exponential family kernels, $\gamma \geq 1.0$, stochastic Model 1 (blue) and mean-field Model 2 (green); centre - stretched exponential kernels, $\gamma < 1.0$, stochastic Model 1; right - power law kernels, $\beta \geq 3.0$, stochastic Model 1.

$\beta > 3.0$ by applying Eq. (2.3), yielding the formal relation:

$$D_{\beta} = \frac{I_{\beta-2}}{2I_{\beta}}, \tag{2.20}$$

where

$$I_{\beta} = \sum_{n=1}^{\infty} \frac{1}{x^{\beta}} = \zeta(\beta) \tag{2.21}$$

is the Riemann zeta function. The close correspondence between predicted and observed values of D shown in Fig. 2.14 serves as a simple check on our random walk results.

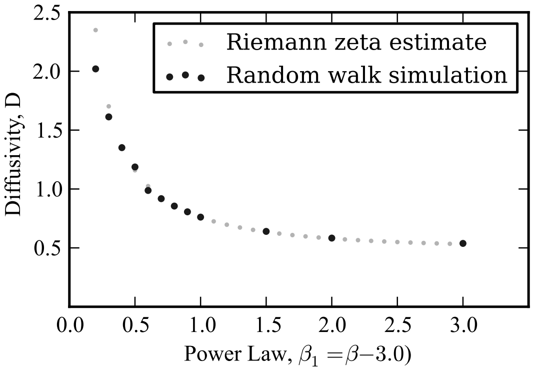


FIGURE 2.14: Diffusion constant for power law kernels, estimated both numerically and by simulation of 5,000 random walks to time 50,000. Results show close correspondence, but begin to diverge as β approaches 3.0.

Note the divergence of the sum at $\beta - 2 = 1$, corresponding to the loss of a finite second moment at this point.

Estimating D for power law kernels when $\beta \gtrsim 3.0$

We can estimate the value of D, where it exists, for power law kernels in the region $\beta \gtrsim 3.0$ as follows. In general,

$$K_\beta(l) = \frac{|l|^{-\beta}}{\sum\limits_{l=1}^{\infty} l^{-\beta}} = \frac{1}{l^\beta \zeta(\beta)}, \tag{2.22}$$

where $\zeta(\beta)$, as above the Riemann zeta function, is the correct normalisation factor. In the region $\beta \gtrsim 3.0$, the principal contributions to the diffusion constant come from very large l. The lower limits (close to zero) are unimportant, and the sum can be replaced by an integral, as in Eq. (2.3). We truncate the lower limits in order to avoid the unphysical low l divergence. Then the diffusion constant is given by Eq. (2.3), and is here

$$D_\beta \approx \frac{1}{2\zeta(\beta)} \int_1^\infty \frac{l^2}{l^\beta} \mathrm{d}l. \tag{2.23}$$

The key part of this integral is in the integral of $\frac{1}{l^{(\beta-2)}}$, which diverges when $\beta - 2 = 1$, or $\beta = 3$. Close to $\beta = 3$ the divergence dominates the behaviour:

$$D_\beta \approx \frac{1}{2\zeta(3)} \frac{1}{(\beta - 3)}. \tag{2.24}$$

Once $\beta \leq 3.0$ the integral diverges, and D is no longer defined.

Behaviour of D in stretched exponential kernels

We can also consider the behaviour of D for the stretched exponential kernels, $K(l) \propto x^{-|l|^{\gamma}}$. We obtain the variance of a spatially continuous stretched exponential kernel by applying the expression

$$\theta^2 = \frac{\Gamma\left(\frac{3}{\gamma}\right)}{\Gamma\left(\frac{1}{\gamma}\right)}, \tag{2.25}$$

given in [125]. We can retrieve the diffusion constant using $D = 0.5\theta^2$, and obtain values that closely correspond to our random walk procedure on a fine spatial lattice ($\phi = 10^-5$), presented in Table 2.2. For the spatially discrete system, there are, unsurprisingly, deviations. Using (2.25) to estimate D is generally better for moderately small γ, with the difference less than 20% when $0.3 \leq \gamma \leq 0.5$; for $\gamma = 0.9$ the difference is about 60%. This is related to the growing role of long-range dispersal as $\gamma \to 0.0$, such that the long tail, where spatial discretisation has least impact, increasingly dominates dispersal behaviour.

We can use these values to roughly estimate c, either by assuming a Fisher-Kolmogorov relationship, $c = 2\sqrt{\alpha D}$, or by applying our heuristically derived fitting for Model 1, $c \approx 0.71D^{0.64}$. Clearly the latter is more accurate for our model.

Behaviour of D as γ becomes large

The estimates of D using Eq. (2.25) are less accurate for larger γ. It is therefore interesting to consider the behaviour of the discrete system in the limit $\gamma \to \infty$. This system is a perturbation around the nearest neighbour kernel.

The kernel is

$$K_x = Ce^{-|x|^{\gamma}}, \tag{2.26}$$

The decay is very fast, such that the kernel can be reasonably represented only by the first two terms: $C \sim 1 + e^{-2^{\gamma}}$. The diffusion constant can then be calculated:

$$D \approx \frac{1}{2C}\left(1 + 4e^{-2^{\gamma}}\right) \approx \frac{\left(1 + 4e^{-2^{\gamma}}\right)}{2\left(1 + e^{-2^{\gamma}}\right)} \approx \frac{1}{2}\left(1 + 3e^{-(2^{\gamma}-1)}\right). \tag{2.27}$$

This is a reasonably accurate approximation when $\gamma \geq 1.5$. A similar correction can be made in principle to the nearest neighbour wave of advance velocity, but we do not pursue this here.

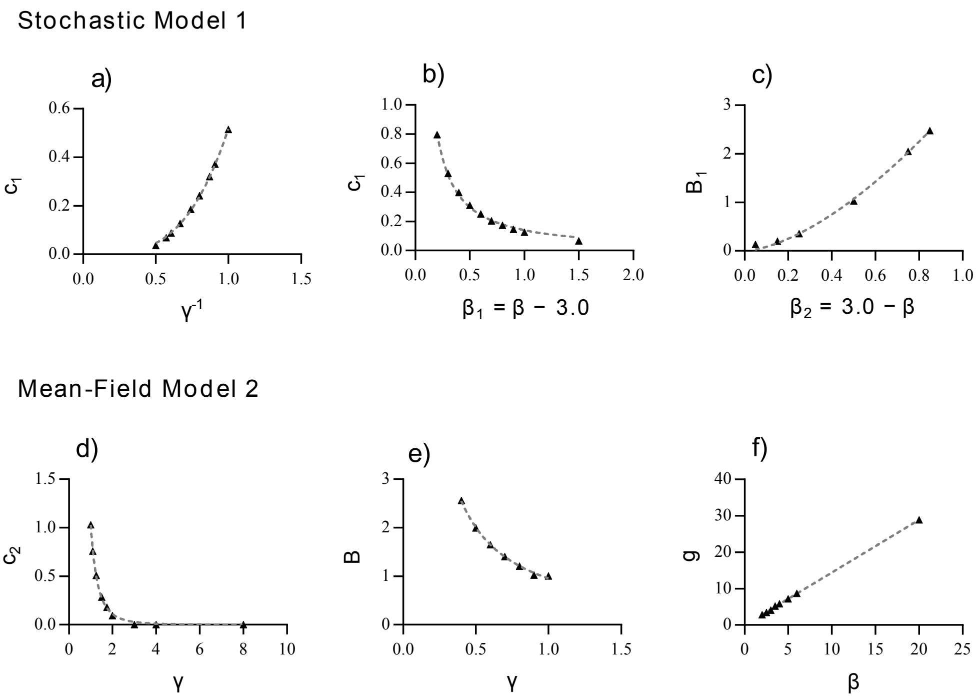


FIGURE 2.15: Deriving wave acceleration and velocity behaviour given kernel parameterisation using linear and non-linear regression. For equations, see main text. Fittings are to modified velocities $c_1 = c - 0.5$, $c_2 = c - 0.78$ and acceleration $B_1 = B = 1.0$. These fittings clarify qualitative behaviour only. a) Model 1: Exponential family kernels: $1.0 \leq \gamma \leq 2.0$; b) Model 1: Power law kernels: $3.2 \leq \beta \leq 4.5$; c) Model 1: Power law kernels: $2.15 \leq \beta \leq 2.95$; e) Model 2: Exponential family kernels: $1.0 \leq \gamma \leq 8.0$; f) Model 2: Exponential family kernels: $0.3 \leq \gamma \leq 1.0$; g) Power law kernels: $2 \leq \beta \leq 20$

2.2.10 Appendix 5: Estimating maximum and minimum wave velocities for the constant velocity waves in Model 1

Model 1 corresponds well with a lattice version of the model suggested by Clark *et al* [106] with their parameter for number of offspring, $R_0 = 1$. We can therefore follow a lattice equivalent of their method for estimating minimum and maximum wave velocities based on the idea of "extreme dispersers". These are the dispersal events that travel furthest ahead of the wave front in each generation, and in doing so define both the wave front for the next generation and the velocity of the wave. The probability distribution of distances travelled by extreme dispersal events depends on occupation in the region of the lattice that can contribute the extreme disperser in a generation.

Two examples are of particular interest. Our one dimensional simulations start with a single occupied site at the far left of the lattice. In this case, occupation at the wavefront is so sparse that only one site is able to contribute the extreme disperser. The distance travelled by the wave is described simply by the dispersal kernel. Alternatively, the wavefront may be densely packed, such that occupation stretches some long distance rearward from the furthest forward occupied site. In this case, the extreme disperser could be contributed by many different sites, and the average advance of the wave in a generation will be greater. If we accept that the extreme disperser defines wave velocity and that it can only originate from a given region of the wavefront, these cases can be used to retrieve estimates of the minimum and maximum wave velocities respectively.

We begin with the case of estimating minimum velocity based on dispersal from an isolated occupied site. In the continuous case studied by Clark *et al*, the probability density function (PDF) of the extreme dispersal event with distance is

$$p(x;1) = R_0 K(x) \left[\int_{-\infty}^{x} K(y) \mathrm{d}y \right]^{R_0 - 1} \qquad -\infty < x < \infty, \tag{2.28}$$

where $p(x; N_s)$ represents the probability of a single propagule travelling distance x from N_s evenly spaced occupied sites being the extreme disperser. R_0 is the number of offspring from a single occupied site per generation and $K(x)$ is the dispersal kernel. This equation corresponds to Eq. (1) of [106]. In our lattice system, we instead consider

$$p(x;1) = R_0 K_f(x) \left[\sum_{0}^{x} K_f(y) \right]^{R_0 - 1}, \tag{2.29}$$

where $K_f(x)$ is the 'forward dispersal kernel' such that $K_f(0) = 0.5$, with normalisation condition $\sum_1^\infty K_f(x) = 0.5$, so as to represent the possibility of both backward and forward dispersal. The cumulative distribution function (CDF) is

$$P(x;1) = \left[\sum_{0}^{x} K_f(y) \right]^{R_0}. \tag{2.30}$$

When the front is a consecutive series of occupied sites on our lattice, the CDF becomes

$$P(x;N_s) = \prod_{d=0}^{N_s} \left[\sum_{0}^{x} K_f |y + d| \right]^{R_0}, \tag{2.31}$$

by which the dispersals of every site in the wavefront stretching back a distance N_s travel to distance x from the furthest forward occupied site or nearer. As the CDF is a discrete series, we simply obtain the PDF using

$$p(x;N_s) = P(x;N_s) - P(x-1;N_s). \tag{2.32}$$

The average velocity of the travelling wave is the expected dispersal distance of the extreme disperser. This is the weighted sum of the PDF,

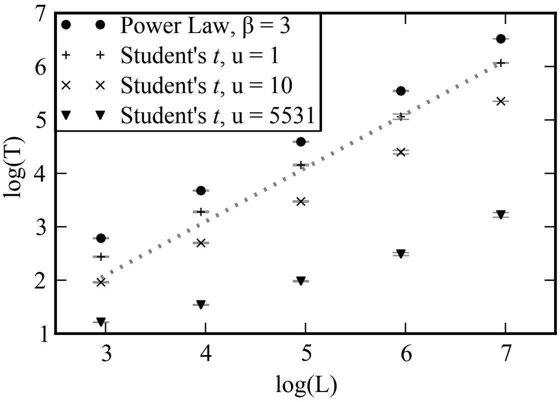
$$E(c; N_s) = \sum_{x=0}^{\infty} p(x; N_s)x. \tag{2.33}$$

We can use these equations to obtain minimum ($N_s = 1$) and maximum ($N_s \to \infty$) velocity estimates. Of course, practically N_s cannot go to infinity in our numerical simulations, so we instead use $N_s = 10^5$, and the length of the kernel is similarly limited to $b = 10^5$. Velocity estimates were obtained through this method using Python, and were found to successfully bound the velocities retrieved from explicit simulation for those kernels that lead to constant velocity waves in Model 1. For short range kernels, the maximum velocity estimated by this method very closely matches the observed velocity. As Clark *et al* suggest, the asymptotic wave velocity of the wave of advance created by fat-tailed dispersal kernels drops away from the maximum value, though does not reach the velocity predicted by the $N_s = 1$ limit. Note that in all our models, $R_0 = 1$. An increase in R_0 does not correspond simply with an increase in bN in Model 1 as R_0 does not take into account the reduced impact of each of the dispersal events. Large R_0 system would correspond to large bN systems if carrying capacity of our sites was constrained to 1 and each full site made N dispersal attempts. See Fig. 2.5 in the main paper for results.

2.2.11 Appendix 6: Finite size scaling of a stochastic system with a bivariate Student's t dispersion kernel

A kernel given by the bivariate Student's t distribution $\left(\text{see Eq. (2.19)}, K(l) = \dfrac{1}{2\sqrt{2u}\left(1 + \dfrac{l^2}{2u}\right)^{\frac{3}{2}}}\right)$ has been discussed in ref [106]. This kernel is both convex at its source and lacks a finite second moment. It is extremely similar at long ranges to a power law kernel with $\beta = 3.0$. Unsurprisingly, we find that wave acceleration behaviour is akin to that observed for this kernel, which is explored in more detail in the main body of this study. Specifically, acceleration is marginal in the highly stochastic Model 1 when parameter $u = 1$, with a gradient on the log-log system behaviour plot, Fig. 2.16, of 1. Note that acceleration persists for long times when u is large.

FIGURE 2.16: Model 1 comparison of finite-size scaling results with a bivariate Student's t kernel and a power law kernel with $\beta = 3.0$. Note the essential similarity of the system behaviour, both tending to finite velocity over time. A linear fitting to the later stages of the simulations using a bivariate Student's t kernel with parameter $u = 1$ gives a gradient of ~ 1, dotted line. Minimum replicates were as follows: $L = 10^3$, 50; $L = 10^4$, 50; $L = 10^5$, 20; $L = 10^6$, 5; $L = 10^7$, 1. Relative errors shown.



2.3 Comments on the manuscript

Shortly before our manuscript was accepted, but long after our first submission, work on a very similar question was published by Fisher and Hallatschek [134] in PNAS. They made far more progress analytically, and, as such, their work may be more methodologically valuable in the long term. However, by using the flexibility of simulations to characterise a much wider variety of systems in detail, our study offers a breadth that is complementary. I also feel that we discuss the literature in more detail, which may be useful for other workers. Nevertheless, I certainly recommend their work for interested readers.

Our findings appear to contradict those of Brockmann and Hufnagel [108], as discussed above. As I was unable to obtain the program used to conduct their stochastic simulations, I re-implemented their model. I used Gillespie's algorithm [136] to simulate a continuous-time, discrete-space stochastic system, with the rates of different reactions (population growth and dispersal) as described by their equations Eq. (1) and Eq. (2). I was unable to replicate their results using my re-implementation. However, without comparing my program to theirs I could not determine the cause of this anomaly - an error in my model implementation, or theirs, or in my interpretation of their system. We therefore didn't include this work in our published paper. Importantly, whether the fault is mine or theirs, bugs in programs and misspecification of models can happen. Unlike complex experimental, or indeed simulation, work, re-implementing simple computational models like ours or that of Brockmann and Hufnagel is relatively easy. Although this particular case has not been resolved, this is a major advantage of simple simulations.

Certain other findings are not presented in the above paper. For example, I implemented a two dimensional model on a hexagonal lattice. I intended to use this for further theoretical investigation and for an adapted version of our work describing human movement over space. Unfortunately, I was unable to replicate the finite-size scaling investigation applied in the paper due to computational limitations associated with dimensionality. While I could simulate systems with over 10^7 sites in either model, this is only about 3150 by 3150 in two dimensions . I was nevertheless able to retrieve some results, detailed in the paper. I note in passing that I was able to replicate plausible rates of human spreading across a 2D lattice representing the Americas by applying a dispersal kernel fitted to modern human hunter gatherers. However, our model is an extreme idealisation of the dispersal process and, as such, I do not consider these speculative results of sufficient interest to detail here.

2.4 Representing long-range movements using the dispersal kernel

In the paper presented above, we represent long-range dispersal using power-law and stretched exponential dispersal kernels. We also consider, in less detail, the impact of truncating the tails of these kernels. These distributions were chosen for two main reasons. Firstly, they incorporate, together, certain mathematical properties that have been shown to impact invasion models. The stretched-exponential and power-law with $\beta > 3$ give two contrasting examples of fat-tailed distributions with a finite second moment. When $\beta \leq 3$ the power-law distribution is fat-tailed and does not have a finite second moment, which can be restored by truncating the tail of the distribution. Secondly, several of these distributions have been widely fitted to movement or dispersal data in ecological studies.

Our work is primarily intended to explore the impact of methods used to model plausible dispersal regimes rather than as a statement on what might be expected in nature. Representing dispersal as a series of independent draws from a single distribution, as we do, will often be a gross simplification, suitable only at certain spatial and temporal scales. Nevertheless, it is useful to have an idea of how closely our model and kernels might correspond to real dispersal behaviour.

2.4.1 Dispersal and movement

The discussion of motility in the ecological literature is complicated by the immense variety of observed movement patterns, both within and across species. Nevertheless, an important distinction is often made between movement and dispersal, the latter having an intrinsic temporal scale of generation time, while the former often involves multiple processes acting at different scales (eg. among elks [137] and aphids [116]). In the context of the kernels we employ, the distinction between movement and dispersal has significant implications. For example, there is a huge literature on movement according to power-laws, focussing on the Lévy flight foraging hypothesis. This combines the observation that animal movement patterns can often be represented using Lévy flights with results on the high efficiency of Lévy flight movement in finding sparsely distributed finite resources to suggest the possibility of adaptation for Lévy flight foraging [94, 138]. The argument also applies to the search, by dispersal, for unoccupied sites suitable for colonisation [139].

I do not discuss in further detail the Lévy flight foraging hypothesis directly, the accuracy of which remains a matter of debate [140, 141]. However, we can draw two things from this literature. Firstly, power-laws with $\beta < 3$, sometimes truncated, yield better fits to sampled movement data among many organisms than other plausible distributions such as normal or exponential movement kernels (see references in [141]). The relevance of this data to dispersal would have to be determined on a species-by-species basis based on knowledge of an organism's biology and its environment, and lies outside the scope of my predominantly modelling work. Nevertheless, movement approximating one of our dispersal kernels over some spatial and temporal scales is known. Secondly, research on Lévy flights foraging has prompted a wide-ranging discussion of how such movement patterns arise. These mechanisms may be of equal interest in explaining dispersal as movement.

2.4.2 Power laws in dispersal data

There is relatively unambiguous evidence for dispersal kernels approximated by power-law kernels among a range of species. Many of these undergo aerial or aquatic dispersal, while others hitch-hike on other motile species. The origin of evidence for Lévy flight dispersal can be experimental, analysing the scatter pattern of seeds or spread of a disease in a controlled environment, or observational, based on tracking animals or less complete first-sighting records of invasive species. Examples from experimental settings include the aerial dispersal of plant seeds (*Calluna vulgaris* and *Erica cinerea*, see mixed model in [89]) and fungal plant pathogens (eg. wheat stripe rust [103]). Observational examples include freshwater plants and marine algae ([142] find truncated power laws with $\beta < 3$ and exponential kernels), individual British starling movements [143], certain fungal spores [144] and various plant and animal diseases [103].

Many organisms are also thought to undergo long distance dispersal by 'hitchhiking' on the movements of other species. An obvious example is the internal microbiome of animals, but others can be found in the movement of bacteria on zooplankton [131], flower mites on various flower visitors [145] and shrimp eggs on water birds [146]. The role of human action is considerable, facilitating many species invasions [74], with specific examples including the movement of seeds on cars [147], sessile marine species on ocean-drifting plastic [148], and both invertebrate species and zooplankton in ballast water [149]. This suggests that the relatively strong case for Lévy-flight-like movement, which has also been suggested for humans [150, 151], is also of relevance to dispersal.

An alternative method of long-distance dispersal involves the active or passive use of aerial currents. These movements likely correspond to dispersal as they are not simply reversed. The passive dispersal of pollen, spores and seeds is complemented by the active jumping mechanism of *Equisetum* plant spores as they settle then dry from a wet, structurally tangled state [152]. Among arthropods, the rearing behaviour of ice plant scale (*Pulvinariella mesembryanthemi*) which catches air currents [153] is relevant, while a very wide range of species 'balloon' through the air using silk threads [154]. The importance of these behaviours depends on the dispersal kernel implied by riding air currents.

2.4.3 The generative processes of Lévy flights

Understanding how power law dispersal kernels with $\beta < 3$ can be generated may give an indication of how widespread they are in nature. Additionally, our model assumes that all individuals undergo independent and identical dispersal, and it is useful to consider under what conditions leading to Lévy flights these assumptions might hold. I consider two types of Lévy flight generative behaviour - that predicted from empirical models of the world, and that suggested by more general mathematical rules.

Empirical models

Lévy flights have been suggested for several species that undergo passive aerial dispersal, as with wind borne seeds. This behaviour is thought to be related to properties of air flow and turbulence, which can lead to a kernel that is not exponentially bounded [155], that follows a Lévy flight [156], or can follow a power law or exponential depending on atmospheric conditions [157]. Power laws movements may also be created by convection in water [158].

Theoretical models

Many processes lead to a signature of Lévy flights in movement or dispersal data (see [141]). Many of these processes describe modifications to a standard model of movement, the correlated random walk, in which the direction of sequential jumps are non-independent.

Multiplicative noise
Generative processes leading to power law distributions often involve multiplicative growth [159, 160, 161]. In the context of movement, incorporating multiplicative noise into the damping velocity component of a Langevin equation describing Brownian motion leads to a power law stationary distribution [162, 163]. This finding has been related to a suggested Lévy walk movement pattern among certain cells and snails [164].

Autocorrelation
Continuous-time representations of correlated random walks lead to Lévy walk movement patterns on a time scale less than that of the autocorrelation [165]; at longer time scales, the power law is truncated. Practically, this observation relates as much to the time scale of sampling in empirical studies than to power law dispersal kernels. However, it is worth noting that sufficiently long term correlations can lead to persistent superdiffusion, as is caused by power-law kernels with $\beta < 3$ [84, pp.116-118].

Brownian landing patterns

Reynolds combines two simple results concerning Brownian motion - the Gaussian horizontal displacement of particles and the distribution of time intervals between first-passage events over a fixed location - to suggest that the spatial distance between successive landings events by flying or floating organisms will approximately follow a power law distribution with $\beta = 2$ [141]. This has intriguing implications for passive aerial and aquatic dispersal in that Lévy flights may be expected even in the absence of convective or turbulent flows.

Population heterogeneity

Heterogeneity in the dispersal kernels of individuals in a population can lead that population to have fat-tailed dispersal kernels, including power-law distributions [166].

Composite correlated random walks

The impression of a Lévy walk can be created when an animal moves according to multiple correlated random walks. This requires that common short steps implied by one behaviour are punctuated by long steps according to another [167]. Theoretically, an infinite series of exponential random walks can be constructed so as to approach a power-law kernel, known as a Weierstrass Lévy flight [92, 168]. Truncating the series such that walks occur according to a finite number of modes similarly truncates the emergent power-law kernel.

Re-orientation based on cues

Forward movement punctuated by reorientation triggered by encountering a scent cue can lead to a power-law distribution of step lengths if the cues are left by organisms making correlated random walks [169]. This is related to the fractal behaviour of the boundary of a correlated random walk.

2.4.4 Stretched exponential dispersal kernels and their generative processes

Stretched exponential kernels lead to interesting behaviour in some models of dispersal [46] and offer an alternative to truncated power laws for capturing moderately long-range dispersal. The stretched exponential distribution is, however, less often fitted to ecological data than a power-law. Exceptions are found in *Drosophila* data [170], wind-dispersed fungal spores [171], the tails of locust movements patterns [172], and in retention times for seeds attached to sheep ([173]; replicated in several sheep datasets, but not for other animals) and human shoes [174].

To my knowledge, mechanistic models leading to stretched exponential dispersal have not been considered in the direct context of biological movement. While the distribution has been found to characterise data from many different fields, no universal explanation of its ubiquity has been identified [175]. Some generative processes have been suggested, but without a specific biological motivation it is difficult to assess which findings might ultimately find application or analogy in movement ecology. As with power laws, stretched exponentials appear in the study of liquid flow, now in the context of the distribution of velocity changes at small spatial scales [176], though the relevance of this is not immediately clear. Three further observations appear of potential interest.

Firstly, stretched exponential distributions are used to describe the relaxation time of disordered systems. Mechanistic models include work on the properties of hierarchical constraints [177], the decay of particles diffusing or transported in the presence of randomly distributed traps [178, 179], and a range of other scenarios [180]. Several approaches rely on deriving a scale-free distribution of waiting times [180]. That diffusion models with traps are included in this framework (albeit not in the context of step-size) is of speculative interest.

Secondly, the extreme tails of the product of a finite number of independent identically distributed random variables drawn from a distribution $p(x) \propto \exp(-Cx^{\gamma})$ is described by a stretched exponential [181]. The importance of this result has been emphasised by [182]. The exponent of the stretched exponential increases quickly as $1/n$ with the number of random variables n, rapidly leading to very fat tails.

Finally, recent work on random walks on networks has found that the step size distribution from a start point to end point falls into one of three classes depending on the topology of the network - a finite number of possible step sizes, a stretched exponential distribution, or a power law [183]. Given that movement between two points might be modelled as a network describing movement probabilities between intermediate locations, this result may have broad relevance to movement patterns over heterogeneous environments.

2.4.5 From generative process to model

The above indicate that movement patterns appearing to follow a power-law or truncated power-law kernel, and perhaps dispersal patterns also, can be generated both by plausible animal behaviours and by models of air and water flow. In contrast, neither the prevalence of stretched exponential kernels nor their mechanistic derivation has been studied in detail. Focussing on the generative processes that create power law kernels, I note that different underlying mechanisms will break our model assumptions in different ways. To emphasise and explore the implications of this, I discuss a situation with clear biological relevance - the generation of fat-tailed kernels through population heterogeneity.

Interestingly, we might expect different causes of heterogeneity to yield different behaviours. Recall that each propagule moves with the same dispersal kernel in our model. If variation in individual dispersal kernels is heritable, the long-term variation in dispersal within the wave-front and system as a whole will depend on some mutation-movement-drift equilibrium. It is likely that the wave-front in particular will rapidly become genetically homogeneous, and dominated by lineages with a high rate of long-range dispersal. The fat-tail kernel observed at a population level due to individual heterogeneity will be difficult to maintain in this homogeneous front. Although our model is likely to provide an accurate representation of system behaviour at some time scales, a non-zero mutation rate affecting dispersal tendencies can lead to evolution in the travelling wave (faster organisms are more likely to escape to virgin territory).

A second possibility is that population heterogeneity is caused by age structure, as suggested in [141]. Age-structured dispersal could lead to rich behaviour. For example, in many models the leading edge of the wave of advance will have a skewed age-structure toward younger organisms.

In summary, the representation of long-range dispersal in our model is standard [46], but is a considerable simplification of real dispersal patterns. Often, these deviations will be of relevance to short-term behaviour. Long-time qualitative results, such as kernels without finite variance leading to an accelerating wave of advance and those that are not fat-tailed leading to a constant wave speed, are more likely to be more robust. Nevertheless, the specific generative function implied by the dispersal process of each species is important. Even if this can yield a kernel that can be approximated by one of the distributions we explored, it is important to assess whether representation using independent draws from identical kernels is suitable. Our work should be viewed as a guide to the implications of modelling choices and a qualitative indication of possible biological behaviour, rather than a quantitative characterisation of actual species invasions.

2.5 The role of simulation in testing mathematical models

To better assess the role of simulation in population biology modelling, it is useful to re-visit the modelling narrative of the work presented above. Briefly, the original idea of modelling the impact of long-range dispersal on species invasions in my thesis was suggested by my supervisor, Tim Sluckin, in the context of apparently rapid dispersal of humans along coastal routes during the Out of Africa migration (eg. [184, 185, 186]). The project follows on from, but is very different to, work modelling Paleoindian dispersals that he was involved in during the late 1990s [187]. Our first simulation models were similar to the stochastic model presented in the manuscript above (with $N = 1$), but in two dimensions and with a strip of space over which long-distance dispersal was possible so as to represent coastal migration. Although we obtained interesting results on the relative advantage afforded to our idealised populations dispersing more rapidly along the coast, our attention was drawn by the marked difference between our stochastic modelling results and those of Kot *at al* in their integro-difference model [46]. Following this, the project became less applied and more theoretical as we attempted to characterise this difference in more detail. Some time through this process we identified studies by previous workers who had considered stochastic long-range dispersal, such as Mollison in the 1970s [95]. Ultimately, we were able to use simulations to investigate in detail the implications of the mathematical approximation of dispersal applied by Kot *et al*.

It is especially interesting that, after relatively little recent work on the relationship between stochastic and deterministic models of dispersal, Hallatschek and Fisher published their more mathematically rigorous study focussing on the wave velocity implied by power-law dispersal kernels in a stochastic system [134]. The spreading regimes they identified corresponded well with those we suggest.

What are the implications of our modelling narrative and of the new analytic results for the role of simulations in population biology? I emphasise two points. Firstly, aspects of the difference between deterministic and stochastic models of dispersal were already explored, analytically and through simulations, by Mollison in the 1970s [95]. By using simulations, he was able to observe the patchy nature of occupation far ahead of the main wave, which is not easily visualised using equations alone. Thus, simulations - though not, as it happens, our simulations - helped to guide understanding of dispersal models, and relate their predictions back to the types of pattern that might be observed in nature.

Secondly, the flexibility of numerical simulations allowed us to probe a relatively wide variety of different dispersal models in the early stages of this project and identify a subject of especial interest. This flexibility also simplified investigation of model variations such as kernel truncation and the detailed implementation of dispersal and death. Overall, a major advantage of simple simulations is their ability to rapidly investigate diverse systems, giving suggestive (rather than comprehensive) results on each. As such, they are an excellent tool to flexibly test the assumptions of model design.

Chapter 3

Modelling animal markets: distributions of species and genes under dynamically biased migration

3.1 Chapter introduction and summary

In this chapter, I investigate the role of animal movement biases, such as those caused by the trading of domestic species between farmers in traditional agro-pastoral economies, on gene flow and selection. Using a simple model in which two subpopulations, one poor and one rich, exchange animals of two different types, I find that migration biases intended to represent trade can reverse the expected course of evolution - the animal type with highest natural growth rate may, under plausible parameter settings, go extinct in the system. Given the counter-intuitive nature of this finding, I discuss the evidence for similar patterns in the real world, and the appropriate methodological response to unexpected results from highly idealised models.

With respect to the methodological advantages of simulation in population biology, this chapter shows how models can be extended by breaking assumptions. The characterisation of migration as unbiased is widespread in population genetic modelling, and simulation offers a flexible tool through which to explore a highly idealised model of biased migration due to trade. As in Chapter 2, I assess the robustness of our modelling results by investigating structural variations of the model. I find that several different model designs generate behaviour that is unexpected based on intuition about evolutionary systems. Indeed, migration biases often have a strong influence on model dynamics, such that they may be an important feature of the evolutionary process whether they arise due to trade or other phenomena (e.g. the road that divides a habitat, the convergence of organisms to a mating site). Nevertheless, the details of model behaviour given biased migration depend on the details of model design, emphasising the importance of relating patterns observed in models to those observed in the real systems they are intended to represent. In this respect, models can guide our observations, highlighting patterns in nature that might otherwise have been missed.

The structure of this chapter is as follows. I begin by introducing the contemporary market for domestic animals and evidence for its antiquity. This supports the notion that the trading of domestic animals, which to my knowledge has not been modelled in an evolutionary context before, is a system of real world relevance. I then briefly address the historic treatment of migration, and migration biases, in population biology, before presenting a draft manuscript detailing our simple model of the trade in domestic animals. After this, I elaborate on the robustness of our findings and suggest several additional questions that might be studied at the potentially rich interface between genetics and economics. Finally, I relate our work back to the role of simulation in population biology models, arguing, as indicated above, that the flexible use of simulation can extend mathematical models and facilitate the modelling of novel systems.

3.2 The market for domestic animals

3.2.1 The contemporary trade in animal genetic material

Human decisions play a fundamental effect in guiding gene flow in domestic animals. Just as with the accidental spread of invasive species, in which our transport and logistic networks facilitate otherwise vanishingly unlikely dispersal events, the movement of domesticates follows routes of human migration, trade and contact. However, agency now plays a more active role. At the local scale of rural markets in the developing world, farmers choose which animal species, and which breeds and individuals, to buy and sell. At the international scale, domestic animals and their germplasm are the focus of globe-spanning markets, involving the long-range redistribution of breeds [188] and individual genetic lineages from specific breeds through artificial insemination [189, 190, 191]. Again, human decisions determine the animals, and genes, that are able to move.

The combined trade of bull semen and live cattle and pigs for breeding is estimated to be worth up to 1 billion dollars annually [190]. The focus of this market is on trade between developed countries, where artificial insemination using intensely selected animal lines [192], and the intensive farming it supports, is widespread. However, proportionally, live animal exports to developing countries have increased [190], while the prevalence of artificial insemination, often using semen from productive breeds originating in Europe, is also rising [189, 191]. This latter development is especially important, in that it fascilitates the transport of animal germplasm on a far greater scale.

That the international trade in animal genetic resources has a profound impact on the distribution of genetic variation is well known [188]. Highly selective breeding, and artificial insemination in particular, have reduced the genetic diversity of many popular breeds, with populations of several million animals often holding the genetic variation expected of a randomly mating herd of a hundred or less ([193], and references therein). It has been suggested that the indigenous, locally-adapted animal populations of developing countries are also at risk (eg. [194]). The question here is what exactly 'at risk' means - certainly, introducing Western genetic variation will initially increase genetic diversity, although this may break apart the associations between characteristic and advantageous traits that define local breeds. Nevertheless, it is clear that we are in a period of intense global homogenisation, that the flow of genetic material is selective, and that it is directed by the changing economic relationships between countries.

While this situation contrasts markedly with the local live animal markets that have served agricultural communities for thousands of years, the core principle - the subversion of animal agency in their movements by people - remains. The unit of decision is no longer large companies, governments, or intensive, highly managed farms, but the individual or family. Agents are driven by altogether different objectives - to liquidate wealth stores in the face of drought [195] and personal calamity (eg. [196]), or obtain animals for the associated social prestige (eg. [197]) and as insurance, as well as the obvious need to provide food, secondary animal products, transportation and labour.

In this chapter, I focus on the local domestic animal market, drawing inspiration from the traditional market for cattle in India. Such small-scale markets are common in many contemporary agro-pastoral societies (e.g. in India [198, 199] and in Africa [200, 201]). This, and several other lines of evidence from archaeology, linguistics and ancient texts, argue for a deep antiquity, discussed below. As such, an understanding of the theoretical impacts of markets on domesticate genetics, and indeed on the distribution of different domestic species, is relevant both to development economics and to our interpretation of the human-domesticate relationships of the past.

3.2.2 Literature and linguistics on the ancient livestock trade

The cattle economy has a rich history, which echoes in our language today. The word 'pecuniary' derives from the Latin for money, 'pecunia', which in turn has origins in 'pecus', a sheep or cow [202]. 'Fee' is related to the German word for cattle, 'Vieh', via Anglo Saxon 'feoh', while the word 'cattle' itself has a common origin with 'chattel' and 'capital', all relating to property and derived via Old French from Latin 'capitalis', 'of the head' [202]. Although head, 'caput', is used as a term for property in classical Latin [203], whether this forms part of a semantic loop linking back to 'head of cattle' is not entirely clear [203, 204]. The relationship between cattle, property and trade can also be seen in the Anglo-Saxon 'caepman' and 'caepscipa', meaning merchant and merchant ship respectively [205]. Value and cattle are similarly connected through the Welsh 'tlws'/'tlus', treasure or jewel, and Irish 'tlus', cattle, through the concept of valuable things [206].

Support for a long-standing association between livestock and wealth is also found in ancient textual sources. In Mesopotamia, the rent for agricultural land was paid as a combination of silver, grain and sheep or goats. The interpretation of the relevant tablets is complicated by a shift in meaning of the Sumerian word for 'goat', *máš'*, which came to mean 'tax' or 'rent', and ultimately 'interest', in the late 3rd millenium BC [207], itself an interesting development. Barter in cattle is mentioned in the Old Testament (*Genesis* 47.17), while in classical Greece, oxen are often used as the principal unit of value - for example, in Homer's *Iliad* (XXIII 700-708) a large three-legged pot is valued at 12 oxen and a female slave skilled in fine handiwork four. In the same epic, a son of Priam, king of Troy, was sold into slavery for 100 oxen, and then ransomed back for 300 [208]. The laws of Draco at Athens, dating to the 7th century BC, stated fines in cattle, but shortly after this came the first legal document explicitly allowing penalties measured in livestock to be paid in marked copper, the Roman *Lex Aternia Tarpeia* which dates to 454BC [209].

The use of cattle as units of value, and sometimes currency, also extended east to southwest Asia and India. In the *Vendidad* of the Zoroastran *Avesta*, which is composed of material of heterogeneous and uncertain dating (though very broadly between the first millenium BC and early first millenium AD), cattle are used in parallel with metalic money [209]. Similarly, in Vedic texts, 'go-puccha' ('tail of cattle', in contrast to 'head of cattle', above) is unit of value and non-metalic currency (Rig Veda iv. 24. 10; vii. 1. 5), while an object received in exchange for a cow is termed a 'gaupucchika' [210]. We return to the role of cattle in Indian history shortly.

Livestock have continued to have an economic function above mere productivity in more recent times. The standard measure of value in medieval Iceland was the 'kugildi', a medium-sized horned cow aged 3-10 winters, without blemishes, having given birth to fewer than 3 calves and providing milk [208]. And in early medieval Ireland, too, we find an emphasis on milk-producing cattle as a method of payment, as well as an indication of social and material wealth [211]. The traditionally elevated place of cattle in the cultural and belief systems of Nilotic peoples is well documented, with a frequent in role in bride-price payments [212], a phenomenon which itself no doubt leads to fascinating migration patterns and which has a deep history (eg. as ārsha marriage in the laws of the *Manu smṛti* (3.29) and before [210]). Inevitably, domestic animals also form part of pastoral bartering systems when economic conditions do not support monetary transactions (eg. [213]).

Livestock have long been used as a measure of value, a store of wealth, and a unit of exchange. The last two of these roles will have impacts on migration patterns. But domestic animals are also a focus of trade themselves. The accounts of individual private merchants in Sargonic (late 3rd millenium BC) Mesopotamia record of livestock transactions [214]. Large scale movements are also in evidence at this time, with one tablet noting the transfer of 3,500 sheep and goats between the cities of Umma to Girsu [214]. There is possible evidence from tablet ITT II 5845 that cattle and sheep were shipped through the now lost city of Agade [214] [5]. The New Testament also refers to the trade in livestock, for example in *Revelations* (18.13; commonly dated to the first century AD) when describing the impact of the fall of Babylon. Polybius (*Histories* IV 38.4) describes the maritime trade in cattle between Byzantium and Greece in the 2nd to 1st century BC, while Strabo's *Geography* (V 1.8) mentions, in 7BC, the inland cattle trade between Italian merchants and the tribes of the Danube [216]. Early direct references to the livestock trade can additionally be found from India, as in the *Pačhatantra* of the Gupta period (320-550AD) [217]. In the broadest terms, then, textual evidence implies the exchange of livestock, whether as barter for goods or as merchandise sold in more developed market systems, over at least the last 4000 years ago. We shall see that archaeology can push back the date for deliberate transportation of animals, if not indisputably for trade, another six millenia.

3.2.3 Domestic animal movements through bones and molecules

Before examining this evidence in detail, however, it is useful to consider what archaeology, literature and linguistics can tell us about trade. Despite biases and corruptions, textual sources have the potential to indicate the quality of contacts between people. Of especial relevance to our question, we can specifically identify cases in which animals were traded by mutual consent (as opposed to through duress), and sometimes the goods that travelled in each direction. Nevertheless, literacy has historically been a skill possessed by, and employed for, the elite, creating potential distortions in both the type and accuracy of information preserved. Perhaps more imporantly, documents tend to cluster around certain locations and time periods. The large-scale, complex societies most visible in the literary record promote long-distance trade, and indeed their collapse tends to be accompanied by, and probably form a feedback with, a reduction in trade [218]. Despite this, the economic lives of ordinary people will continue. Evidence from historical linguistics, and in particular the semantic shifts mentioned above, helps

[5] The original translation does not appear to indicate maritime activity, see de Genouillac [215], available online as of 08/2015

to bridge this gap, both because there may be less of a correlation between language change and societal complexity and because language change is less obviously biased toward the influence of elites. Our modelling will focus on the effects of small-scale, decentralised livestock trading, and our discussion of the archaeological evidence will be similarly directed.

When approaching archaeological data, the ease of interpretation for different forms of evidence varies. For example, the spread of a subsistence strategy focussed on livestock and agriculture is clear from zooarchaeological evidence due to ‘first finds’ of domestic bones at different sites. Early cases of early movement of animals by people include the transportation to Cyprus of cattle, sheep and goats around ten thousand years ago [219], or the diffusion of the Neolithic around the Near East and Anatolia [220]. However, later exchange is more difficult to detect and interpret. Tools exist - stable isotope analysis in particular, as well as ancient DNA and bone morphology - but even if one identifies non-local animals, the question of interpretation remains. Finds give a tangible insight into the past, but are spatially and temporally clustered, rendering attempts to weave together a digestible narrative, perhaps covering many hundreds of years, both challenging and dangerous. For our purposes, then, I will present available data on animal movements in societies where domesticates are firmly established.

During growth and development, bones and teeth incorporate an isotopic signature based on dietary intake - the type of food consumed and where it was grown. The signal from strontium isotopes ^{87}Sr/^{86}Sr derives from the geological makeup of a region, and is particularly useful in identifying changes in diet associated with animal movements. A heterogeneous pattern of animal migration is suggested. Early cattle migrants have been identified in Linearbandkeramik sites in Germany dating to the early European Neolithic (late 6th millenium BC) [221]. Some migrants have also been detected among sheep remains from Çatalhöyük (7400-6000BC, Turkey) [222], and among cattle from Neolithic France (3350-3050BC) [223]. Extensive animal movements appear to have occured in Neolithic western Sweden (mid-4th to mid-3rd millenium BC), where over 50% of cattle at studied sites were found to be non-local, far more than in either the human or domestic pig populations [224]. A simliar pattern, in which a vast majority of cattle appeared to be migrant as compared to a minority of pigs, was found in a Late Neolithic (2515-1460BC) site associated with feasting in Britain, with some cows having been transporated over 100km [225, 226]. Other animal remains have also been studied - interestingly, an ass that had been sacrificed in what is thought to be the merchant’s quarter of the Early Bronze Age (3600-2400BC) Israeli site of Tell es-Safi was found to be a migrant, with some indication that sheep in the same deposits were also non-local [227]. Evidence for animal movement unsurprisingly extends to non-Eurasian societies, such in South Africa (first millenium AD), where migrant cattle and sheep show profiles indicating extended periods of non-local occupation [228].

Direct evidence is available, then, for the movement of animals in established pastoral communities from before 5000BC. Given the transportation of domestic animals inherent in their first diffusion through southwest Asia at the onset of the Neolithic, there is no especial reason to assume such movements did not occur substantially prior to this date. The studies mentioned above do suggest, however, considerable inter-site variation in the distances of livestock movements, and also intra-site variation among different domestic species. The precise nature of these movements is unclear - temporal patterns in non-local isotope signatures argue for transhumance in some studies (eg. [221]) but against it in most (eg. [222, 227, 228, 225]). Although the role of trade in increasing animal movement distances has been specifically suggested based on an increase in the proportion of migrants and the distances travelled associated with the Roman period in Britain [229], and although trade and the movement of goods was clearly an important process in Neolithic society (e.g. Robb and Farr review trade in the Mediterranean [230]; evidence for maritime trade in the Persian Gulf from the sixth millenium BC [231]; the Near East trade in bitumen [232]; and widespread trade in obsidian [233]), the isotope data is consistent with, rather than proof of, ancient livestock exchange networks.

These varied evidences, then, suggest that markets for domesticates have existed for several thousand years. This translates to many hundred generations, depending on the domestic animal in question. Furthermore, the market in reproductive live animals, as well as other forms of human-mediated animal exchange such as dowry and bridewealth, exist to this day. There is strong support, then, for market forces playing a significant role in the current and future distribution of animal genetic diversity.

3.2.4 An overview of movement models

The literature on movement models is diverse, with work focussing on different temporal scales of movement in different species that may have quite variable biological mechanisms for, and patterns, of movement. As such, strict definitions of key terms (like movement, dispersal and migration) are likely appropriate in some contexts but not others, and for some academic fields but not others. I therefore do not give formal definitions. Nevertheless, I intend the terms ‘migration’ and ‘dispersal’ to describe movement that is likely permanent and, over the course of an individual’s lifespan, rare. In some cases, these terms can be considered as the total movement process taking a parent from its birthplace to the birthplace of its offspring. The term ‘movement’ will be used more broadly to mean either short- or long-term movements, but a ‘movement model’ is specifically a model or short-term day-to-day movements such as foraging or predator avoidance.

Pioneering work in modelling movement focussed on two rather different themes. Ravenstein's Laws of Migration, published in various forms in the 1890s (eg. [234]), but echoing his own work of 20 years earlier [235], attempted to lay down postulates concerning patterns of human migration, as understood from census data. The conclusions drawn, identifying a predominance of small distance movements, with larger movements to large population centres, anticipate the application of a gravity model of human migration (eg. [236]); a similar theme had earlier been considered by Monge on the subject of trade flows [237]. This model of migration in humans has been hugely influential, and both it and important alternatives (eg. [238], and recently [239]) are based on human migration being structured according to both distance and other landscape qualities. Although alternative descriptions based on simple random processes may capture aspects of modern human movements (the Lévy flights in [150]), models are usually based on a clear assumption that we can know more about human motility and migration than a random null hypothesis.

Quite the opposite approach has been taken in the study of non-human movement, with the starting point of the earliest models being the random walk [10, 240]. When the movement of individual animals over short time scales is the topic of study, directional correlations are often introduced (a correlated random walk, [241, 242]). Conversely, population genetic models and those of species invasions tend to focus on the effects of movement over a generational timescale. Short-term correlations in the direction of movement are no longer of interest, and often individual dispersal events are described only in terms of their effect on the population.

We have already discussed this representation in the context of species invasion models in Chapter 2. Various methods of representing the dispersal of many individuals can be used to model their population-level effects, from diffusion approximations [11, 85, 86] to integro-difference equations [243] to systems that fully or partially represent the stochasticity of random walks (eg. [95, 1]), each taking different assumptions.

The representation of migration in population genetics generally focusses on the idea of a population, and in particular one in which mating is random. As such, spatial representation tends to focus on a network of connected subpopulations rather than continuous occupation in space. Exceptions certainly exist - R. A. Fisher's wave of advance model was originally conceived in the context of the selective spread of a favoured allele [11], while J. B. S. Haldane's ideas about genetic clines [244] have proven highly influential in evolutionary ecology. Nevertheless, the earliest migration models [245, 13] focussed on the role of migration between partially isolated patches of occupation, and especially on the potential of migration to disrupt local adaptation [246]. Several approaches to connecting these patches have been suggested - from a single island linked to a much larger mainland [245] to a series of connected islands [13] or subpopulations, potentially with different environmental conditions [247, 248], to neighbouring populations interacting only as 'stepping stones' on a 1D or 2D lattice [249]. Indeed, the last of these is essentially a spatially discrete variant of some of the models of dispersal mentioned above, with a nearest-neighbour dispersal kernel. The essential assumptions behind many of these models is, again, that migration proceeds at a constant average rate between locations, with correlation in movement between individuals or over time ignored.

3.2.5 Human versus animal migration and the random walk

Before briefly discussing the representation of movement biases, such as those created by market forces, it is interesting to note the radically different starting points taken when modelling animal and human movement. In the case of humans, work began on the basis of detailed, though limited [235], datasets in the form of census surveys. This, combined with privileged insight into the motivations for human migration, meant that attention was focussed on the role of environment and agency in movements, contrasting sharply with the pioneering random walk models used to describe insect movement [10]. More recent work has identified Lévy flights (e.g. [150, 250, 251]; also see Chapter 2), as capturing elements of human movement. Although the distribution of movements may be well fitted by a power law, however, this is usually recognised as an emergent property based on the distribution of e.g. cities [250] and, perhaps, spatial properties of social networks [252].

The situation is reversed for species invasion and population genetic models, where a simple random walk has, historically, been the basic microscopic system from which models are then derived. This approach has yielded important theoretical results, some of which have been validated by data, and is widely used when modelling animal movement. Nevertheless, there is a recognition that both dispersal (e.g. Part 1 of [253]) and everyday movement [254] of animals are highly non-random. By this, I mean that the direction of travel and locations visited are determined by phenomena such as resource availability, habitat preferences, and community interactions. The first of these (though explicitly not the latter two) forms the basis of the classic theory of the ideal free distribution (IFD, [255]), which I will introduce later. Understanding all the contributing factors to the non-random component of animal movement is part of the newly-described discipline of movement ecology [254], which, prompted by technological advances such as GPS, aims to obtain an integrated understanding of the motivation, capability and stimuli involved in determining movement patterns.

Given that animal movement is extremely complex, the apparent consistency between theoretical results derived from simple models and some real world observations is interesting. An example is the constant rate of population spread predicted by many models of species invasion. There are multiple explanations for this pattern. For example, if movement direction biases tend to act on short scales of time and distance, then over larger scales they may still be well-approximated by models based on random walks.

From the other perspective, a single modelling result may be generated by quite a broad range of model specifications (e.g. [77]). Both these points are examples of equifinality - a single end state being reached through multiple processes, or system trajectories, or initial conditions. The former can be described as equifinality in the target system, while the latter is equifinality of the model. These are related, but different, in that the model is only a representation of the system. Practically, the situation leads to two difficulties. Firstly, there is a danger of over-interpreting the parameterisation of a model. Successfully describing the rate spread of an invasive species using a given diffusion constant says nothing about the diffusivity of individuals. Secondly, the model may break down in unpredictable ways as more complex questions are asked. The impact of long-range dispersal or Allee effects on species invasion models offers an example.

3.2.6 Representing movement biases in models

How one constructs of a model representing complex movement patterns depends on the time-scale of interest. For example, in species invasion models a single, direction-symmetric (isotropic) dispersal kernel applying to all individuals might be applied quite successfully to a population-level description of movement (as in integro-difference equations) even if individuals vary in their dispersal capabilities - when capability is not age-structured or heritable. The details of the target system are important here. I now offer some examples of how movement that is poorly represented by an isotropic random walk with homogenous individuals has been represented.

Remaining with species invasion models, the simplest approach has been to incorporate directional-asymmetry. When a diffusion equation is used to describe the spread of a population, an advection term may be introduced [256]. The equivalent in models that represent dispersal tendencies through a statistical distribution would be an anisotropic dispersal kernel (e.g. [257]). Such approaches are used to describe riverine dispersal (as in [257]), dispersal according to coastal currents (e.g. [258]), and the impact of wind direction on spore dispersal (e.g. [259]). Models have also been modified to incorporate structured dispersal, whereby the dispersal kernel of a population depends on the features such as the age of individuals from which it is composed [79].

In population genetic models, migration is usually represented as a single parameter describing the rate of gene flow between subpopulations. This, in turn, is determined by the movement of individuals. As we might expect, the focus is on heritable variation in migration, and, often, the evolution of migration rates under different circumstances. The simplest approach is to describe migration using multiple parameters, such that the rate of migration is different for different subsections of the population, depending on allelic state. An early (1973) example is the 2-subpopulation model of Balkau and Feldman [260], who mathematically investigated the fate of a migration-modifying mutation linked to a genetic variant favoured in one patch but negatively selected in the other. Unsurprisingly, they found that the mutation would rise in frequency if it reduced migration rate only. Work from the same period used simulations to explore more complex situations, including density-dependent migration between environmentally heterogeneous patches, with migration tendency represented either as a Mendelian or complex trait [261]. Models in which migration rate depends both on the direction and on allelic state have also been considered, as by Bull *et al* [262]. The evolutionary implications of heritable variation in migration rate is of relevance to our results, and this study, supporting the earlier work of Hastings [263], found that selection tends to act against migration when the environment varies over space. Other factors can promote migration, such as kin competition [264] or an environment that is both temporally and spatially variable [261].

The representation of genetically heritable migration tendencies is similar, and sometimes the same, as that of multiple species with different tendencies (e.g. [265, 266, 267]). However, there is often a more explicit recognition of the varied pressures on movement decisions. The IFD offers a classic example, deriving a simple mathematical result from a simple model of animal behaviour. Under certain conditions - animals are equal competitors for limited resources that do not change in quality over time, have complete information about resource quality, and move without costs and independently of other animals - this model suggests that the distribution of animals will tend to minimize resource competition [255, 268]. Once animals are distributed in such a manner, migration can only reduce the reproductive success of an individual and hence, under the stated assumptions, will evolve toward zero [263, 265]. McPeek and Holt, in a detailed simulation study, found that this is no longer guaranteed when unbalanced migration (in which migration rate depends on direction) is possible [269]. When locations have equal carrying capacity, unbalanced migration is selectively excluded, a result recreated in my own work. However, when several species with different directional biases exist, these can reach an equilibrium at frequencies depending on the details of these biases.

More complex models of animal movement that represent different aspects of the decision to move have also been considered. Payoff-biased migration has been explored in the context of social interactions described using game theory [270]. While this is similar in principle to the movement toward optimal resource availability in the IFD assumptions, the payoff in a location is now determined by the game strategies of its resident population. Other game-theoretic models have explored the complex effects of migration distance being heritable and movements occurring when individuals are confronted with adverse conditions in interactions with a dynamic social network (e.g. [271]). Even relatively simple modifications to random movement can lead to biological aggregations, such as density dependent movement speed, turning rate, or direction biases, and have been described using advection-reaction equations (for a mathematical review see [272]).

The population genetic and ecological models above suggest that incorporating migration biases into models of movement is not a recent innovation (e.g. [256, 260]). However, there has been increasing emphasis on the evolutionary importance of non-random or complex movement patterns [254], both in natural populations and models (e.g. reviewed in [273]). For example, biased migration and the selective forces acting specifically on migrants are considered important factors in maintaining local adaptation [274] and promoting long-range correlations between genetic variation and environment [275]. Modelling forms parts of this resurgence, with a model that our simulations extend [267] offering an example. There are increasing hints from modelling work that biased migration may play an imporatant role in diverse population systems (e.g. [266, 276]).

I now present a draft manuscript in which I model the impact of animal trading on spatial patterns of breed variation. For details on contributions of myself and co-authors to the manuscript, see the Acknowledgements at the start of this thesis.

3.3 On the counter-evolutionary effects of market mediated gene flow *(Unpublished manuscript)*

G. S. Jacobs, G. A. Kaiping and T. J. Sluckin[6]

3.3.1 Abstract

Human decisions have significant influence on the evolutionary trajectories of many other species. This influence is particularly strong in the case of domestic animals and cultivated plants. Here humans directly control both breeding opportunities and the environment to which individuals are exposed. More significant in an evolutionary context may be the fact that, through trade-related changes in ownership of domestic species, humans also influence their migration patterns. Here, we modify a simple two-allele, two-patch population genetic model to reflect aspects of local animal markets. We consider the implications of different wealth distributions in the two patches, and draw on auction theory to derive either static or dynamic migration parameters from the market process. When animal movements are unbiased by market forces, the evolution of the system state is strongly determined by selection on reproduction. Introducing biases to movement through our model of trade can break this expected behaviour, such that a negatively selected animal type is able to reach local or global fixation. This effect is prevalent in a range of models we explore, and in some cases has a greater effect on the equilibrium behaviour of the richer of the two patches. The general indication is that wealth inequality is likely to impact the distribution of species, breeds and genes in interesting and sometimes surprising ways. This may be relevant when designing development interventions involving imported animals or their germplasm, and also in our understanding of the evolution of breed distributions and transitions to livestock-owning subsistence strategies. More broadly, migration biases may play an important evolutionary role in contexts that do not involve human intervention.

Keywords: Market genetics, animal trade, animal market, asymmetric migration

[6] This author list will hopefully be extended, depending on the future course of this project, but reflects current contributions.

3.3.2 Introduction

The impact of human action on the movement patterns of other species is substantial. Much of this is through unintended consequences - the road that divides a habitat [277], the fishing that distorts local ecologies [278], the hitch-hiking of invasive species on our logistic networks [74]. But running parallel to these accidental effects is the deliberate transport of animals and plants, often through trade. Domestic species, and livestock in particular, are a primary focus, and the impact of large scale trade in germplasm from highly selected lineages can be seen in the predominance of certain breeds, and the substantial reduction in genetic diversity among these [193]. The exchange of animal genetic material is not new, however, and traditional live animal markets remain a common feature through much of the world.

The commercial network dictating animal movements has a significant role in the spread of infectious disease [279, 280, 281] and genetic variation [190, 188]. The structure of connectivity in the network of livstock movements has been used to assess the threat from epidemics, clarify the role of animal movements in these, and understand the impact of potetial interventions [281, 282, 283, 284, 285, 286, 287]. Similarly, when reproductive animals or their germplasm are traded, gene flow occurs through the market network. An understanding of this process clarifies the natural backdrop of domesticate genetic variation, but may also have important implications for development interventions that re-distribute animals or provide artificial insemination services. These are run by a wide range of governmental and charitable organisations [288, 289, 290, 291, 292], and, in addition to improving the livelihoods of farmers, often explicitly aim to increase the productivity of indigenous breeds. The spread of domestic animal populations, and of genetic material within these, is thus of considerable interest.

The movement of organisms has long been recognised as a critical dimension of ecological (e.g. [10]) and genetic [13] systems, and is often studied through computational or mathematical models. Although the factors impacting individual dispersal may be complex, and in some cases difficult to infer, it is common practice to characterise the average behaviour of many individuals and thereby make assessments of the population-level impact of migration patterns. Given this, we might suggest that the mechanism of movement is irrelevant, and that markets can be represented by a suitable footnote to a classic diffusion equation [11] or a bi-directional graph of populations linked by constant migration terms [293]. The problem is that human choice plays a significant role - the flow of animals is driven by economic factors, which are, in turn, impacted by consumer trends [294] and production technology [295, 296], as well as by climatic events [297, 298] or social phenomena such as conflict [299] at a more local scale. Furthermore, the animals selected for trade are an unstable subset of the population as a whole, determined by regional and personal preferences [300, 301].

Economic models, such as work on supply chains [302] or trade networks [303], can guide our understanding of the flow of goods through a system given different economic conditions. However, when trade is in reproductive animals rather than animal products, the flow of goods does not always fall into a natural channel from manufacturer to consumer, often an assumption of these models (though see, for example, [304]). More fundamentally, animals are heterogeneous goods that reproduce in a density-dependent manner, with heritable variation that allows for natural and artificial selection. The 'population of goods' is not a natural focus in economics. Nevertheless, it is clear that many concepts from microeconomic theory are of relevance to the description of the livestock trade, and auction theory plays a significant role in the work we present in this paper.

So far, we have emphasised the complexity of the livestock market. This is not to say that all these features must exist in a representation of domestic animal movement patterns, and judging the thorny trade-off between model complexity, interperability and realism is complicated and situation-dependent. However, asymmetric, dynamic migration of animals based on economic characteristics may be the rule rather than the exception. In this paper, we explore the role of markets in creating such conditions, describing how asymmetric non-constant gene flow can be derived from a simple representation of animal trade.

We investigate the evolutionary implications of trade using a classic two-patch, two-allele model from population genetics [247, 267]. This offers the simplest possible population topology and representation of animal types. Even here, market-mediated gene flow leads to equilibria that contrast dramatically with those expected from symmetric-migration models, with some parameterisations leading to the fixation of a breed type with globally inferior reproductive capacity. We conclude by discussing the relevance of this model, and in particular the critical parameters in it, in the context of the Indian cattle market, and consider evidence for our model dynamics among agro-pasotral communities.

3.3.2.1 The static and dynamic market filters

Central to this work is the question of how economic choices and constraints modify the movement of animals between herds. We can suggest multiple forces at work - environmental conditions distorting local supply and demand, differences in access to markets and transaction costs, legislative effects controlling which animals may be exchanged, locally specific preferences for animal breeds, and wealth inequality between herders are some examples. In an idealisation in which animals are taken to markets and then traded, such that human decision at the market determines the exchange of animals between herds, these features combine to serve as a 'filter' on migration.

When the impact of these human choices on migration are independent of the composition of the market in terms of animals, buyers and sellers, we describe the filter as *static*. Conversely, a more realistic representation of the market allows purchasing decisions to depend on the animals available and on competition. This means that the market-induced filter on migration is conditional on the market composition, and the filter is *dynamic*. With regard to modelling, in both cases biases caused by markets mean that animal movement can no longer be captured by a single, constant migration parameter that applies to all animal breeds and is equal in all directions. Instead, migration can incorporate breed-specific biases in the direction of movement. The static filter implies that the various migration parameters remain constant, while the dynamic filter implies that these parameters are updated as the system state changes.

The correct description of the market filter, which allows us to derive appropriate migration rates, is a critical step in the development of our model.

3.3.3 Methods

In this work, we explore the role of the market as a migration filter in determining the equilibrium occupation of two breed types in a two-patch model. Our system incorporates wealth inequality between the inhabitants of the two patches and differences in their breed preferences, as well as patch- and breed-specific reproductive capacity among animals. The two animal types are presented here as reproductively isolated breeds, but can equally be considered different animal species, or carriers of different alleles at a bi-allelic locus subject to selection and market-choice. We first describe a deterministic system of recurrence equations that represent this model, before finding equilibria and exploring dynamics numerically.

3.3.3.1 Model design

Our model is based on that of Bolnick and Otto [267], who extended the classic Levene model of migration between subpopulations [247] to include direction-dependent and genotype-dependent dispersal and explored special cases of that system. This model is general enough to represent the market as a static filter on migration, and, as our model builds on theirs, we give an overview of their model now.

The model consists of two habitat patches, here Patch 0 and Patch 1, with carrying capacities K_0 and K_1. These are fully occupied by two animal breeds, Breed A and Breed B. The system is updated according to a deterministic update rule. This consists of recurrence equations that, given the proportion $\kappa_i(t)$ of Breed κ animals in Patch i at time t, yield the new proportion $\kappa_i(t+1)$ at time $t+1$. The process involves calculating three quantities - the proportion after migration, $\kappa_i^m(t)$, selection, $\kappa_i^s(t)$, and regulation $\kappa_i^r(t)$. The census of breed occupation in each patch occurs after population regulation, so $\kappa_i(t+1) = \kappa_i^r(t)$. Breed- and patch-dependent migration is first applied,

$$A_i^m(t) = \frac{\sum_{j\in 0,1} K_j \pi_{ji}^A (1-c_{ji}) A_j(t)}{\sum_{\kappa\in A,B} \sum_{j\in 0,1} K_j \pi_{ji}^\kappa (1-c_{ji}) \kappa_j(t)}, \tag{3.1}$$

where π_{ji}^κ is the migration rate of Breed κ from Patch j to Patch i and c_{ji} is the mortality associated with any migration from Patch j to Patch i. Selection then occurs,

$$A_i^s(t) = \frac{s_i^A A_i^m(t)}{\sum_{\kappa\in A,B} s_i^\kappa \kappa_i^m(t)}, \tag{3.2}$$

where s_i^κ represents the fitness of Breed κ in Patch i. Finally, population regulation is applied, which is the last update step,

$$A_i(t+1) = A_i^r(t) = \frac{K_i A_i^s(t)}{K_i \sum_{\kappa\in A,B} \kappa_i^s(t)}. \tag{3.3}$$

This form of population regulation is rapid, such that total occupation in each patch is set to the carrying capacity with no regard for the implied rate of population growth or mortality. Such a simplification allows us to ignore current population size and focus instead on the relative occupation of each breed in a patch. If we also assume that selection doesn't change the total population size, population regulation could equivalently be applied following dispersal.

In this work, we focus on the role of migration in this system. Accepting that the actual population dynamics of livestock are likely to be more complex, we therefore take several assumptions. Regarding model parameterisation, we assume that the carrying capacities in the two patches are equal, $K_0 = K_1 = K$, and that mortality in the trading process is negligible, $c_{ij} = c_{ji} = 0$. We also assume that the selection stage represents the combined effects of natural selection due to environmental differences between the two patches and artificial selection due to local breeding control. Finally, as stated, population regulation is assumed to be rapid (though see Appendix 2, §3.3.7).

This model is sufficient to describe a static market filter, and we now discuss how such a filter might be inferred from the trading process.

3.3.3.2 The market algorithm

Our chosen update rules, Eqs. (3.1, 3.2, 3.3), do not explicitly represent wealth and the breed preferences of herders in each patch. Instead, we incorporate these important economic factors into our calculation of the four migration parameters, π_{ij}^κ. We consider each patch to have a disposable wealth distribution, $W(i)$, and adopt a convention whereby $P(\xi_{W(0)} > \xi_{W(1)}) > 0.5$, where $\xi_W(i)$ is a random variable drawn from distribution $W(i)$, representing the disposable wealth of a buyer from Patch 0. This convention means that buyers from Patch 0, on average, can outbid those from Patch 1.

Using local Indian cattle markets as our general guide, see §3.3.5, we consider a market to be a regular event occurring at a publicly known location. In real markets, livestock are brought to a market by their owners or professional animal traders from the surrounding area. Farmers and traders also attend in order to buy animals. Through the course of the market, trading takes place, which may occur through intermediaries who determine price in exchange for a sometimes complex commission system including payments by one or both trading parties. Animals that not sold return to their owners, and may be taken to other local markets. In the Indian cattle market, most animals that are traded are non-reproductive bullocks used for draught labour, but reproductive bulls and females are also sold; these are the animals that we are primarily interested in.

In our idealisation of this system, then, we identify three stages:

1. A set of animals are chosen and taken to the market by farmers from each village, and buyers also travel from each village
2. Animals are bought and sold.
3. Animals that are not sold return to their home village, while those that are sold return to the village of the buyer.

We are interested here in the role of wealth and preferences in distorting the outcome of the market, as described by these stages.

Human choice is involved at three stages - the choice of animals taken to market, the decision of buyers to attend market, and the trading process itself. Critical to our modelling of these choices is the idea of wealth as a route to avoiding undesirable decisions, such as selling a preferred animal, and promoting favoured outcomes, such successfully competing for a desired animal. For simplicity, we assume identical breed preferences for all herders in a patch, with ρ_i^κ indicating the preference for Breed κ in Patch i. Given this, our simplest model of the market would suggest a bias in migration such that a wealthier district is subject to greater inward migration from its preferred breed and greater emigration of it's less preferred breed. In the simplest representation of this, a static filter is created, in which the four constant migration parameters of Eq. (3.1) are sufficient to describe animal movements, Fig. 3.1.

Deriving a static market filter

Consider a situation in which Breed A is more reproductively successful in both patches,

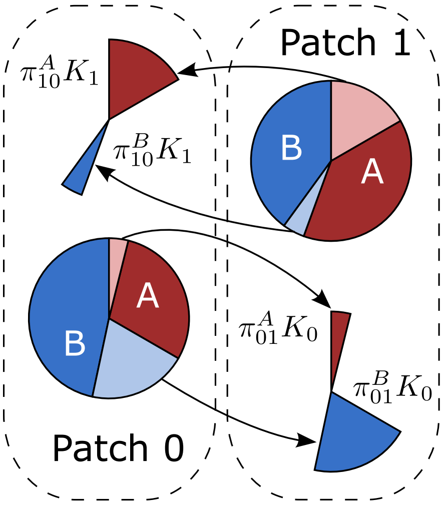


FIGURE 3.1: A single migration step given a static filter, in which there are fixed breed-specific and direction-specific migration parameters, π_{ji}^{κ}, which are often unequal, describing animal movement patterns. Relative occupation of Breeds A and B is indicated by red and blue shading respectively, with the circles showing patch occupation before migration and the areas enclosed by dashed lines indicating post-migration occupation. Lighter shading is used to denote animals that move during the migration step. While the importance of patch-specific carrying capacities is emphasised in this figure, in our models $K_0 = K_1 = K$.

and the order of purchase preference for herders from both patches is Breed A, then Breed B, then no cow at all, $\rho_{0,1}^{A} > \rho_{0,1}^{B} > \rho_{0,1}^{\cancel{A,B}}$. Given that Patch 0 is wealthier, we expect a bias of movement of Breed A to Patch 0 and, correspondingly, a bias in movement of Breed B to Patch 1. The strength of this bias depends on the level of wealth inequality between the two patches. A logical static filter result would be a symmetric migration bias - and this is precisely the migration model investigated by Bolnick and Otto in the context of genotype dependent migration to favourable environments. The only difference between their model and the one proposed here is that Breed A is globally and equally selected for, $s_i^A = s_j^A > s_i^B = s_j^B$, rather than biased migration acting to increase movement only toward selectively favourable environments as in [267]. This difference is nevertheless expected to cause very different system behaviour. Thus, we have a system of symmetrically distorted high and low migration rates,

$$m_H = \pi_{01}^{B} = \pi_{10}^{A} = \bar{m} + \frac{d_m}{2}, \tag{3.4a}$$

$$m_L = \pi_{01}^{A} = \pi_{10}^{B} = \bar{m} - \frac{d_m}{2}, \tag{3.4b}$$

$$\bar{m} = \frac{m_H + m_L}{2}, \tag{3.4c}$$

$$d_m = m_H - m_L \geq 0. \tag{3.4d}$$

This offers our first example of a static filter, the symmetric Static Filter 1.

Static Filter 1 has the significant advantage of simply representing a plausible outcome of animal trading, with the wealth difference between the two patches scaled using d_m. However, the disconnect between these equations and the detailed trading process obscure the conditions necessary for them to hold. We can clarify matters by considering the type of market algorithm required to create such a model. Assuming the carrying capacity is the same in both patches, $K = K_0 = K_1$, a possible scenario runs as follows:

1. $2\bar{m}K$ random animals are bought to market by each population. Note that this gives the constraints $\bar{m} \leq 0.5$, $|\delta_m| \leq 2$.

2. The globally preferred Breed A animals are sold first, with each individual submitting their bid as drawn from the disposable wealth distribution of their home patch, $W(0)$ or $W(1)$. To obtain the correct migration behaviour for rich Patch 0, we solve

$$K\big(m_H A_1(t) - m_L A_0(t)\big) = \\ P(\xi_{W(0)} > \xi_{W(1)})\, 2\bar{m}K\big(A_0(t) + A_1(t)\big) - 2\bar{m}KA_0(t), \quad (3.5)$$

   yielding $P(\xi_{W(0)} > \xi_{W(1)}) = \dfrac{m_H}{2\bar{m}}$. The corresponding probability of a herder from the poorer Patch 1 making the winning bid is $P(\xi_{W(0)} < \xi_{W(1)}) = \dfrac{m_L}{2\bar{m}}$.

3. Bidding now commences on the cheaper Breed B animals. As individuals from rich Patch 0 are able to outbid those from Patch 1 for these too, adaptive bidding behaviour is needed to maintain the symmetric filter described above. Two examples of such bidding rules are:

   (a) Total demand from Patch 0, here the number of animals that population seeks to obtain, is set to $D_0 = K\Big(\bar{m} + \dfrac{\delta m\big(A_1(t) - B_1(t)\big)}{2}\Big)$, and rich individuals stop bidding when this demand is satisfied.

   (b) There is a bias amongst rich Patch 0 buyers against bidding for worse quality animals, such that while they win the same proportion of bids, the probability of submitting a bid is $\dfrac{m_L}{m_H}$.

   In the above cases, demand of the poor population must be sufficient for them to obtain the remaining animals.

4. Animals are re-distributed according to winning bids and the system proceeds to the selection phase.

In Step 3, adaptive bidding can also take the form of restrictions in the prices offered for Breed B animals. Such behaviours may be consistent if they represent either perceived differences in the utility gained from the two Breeds by Patch 0 herders, or delayed bidding in an effort to obtain the preferred breed type in a future market event. However, they are quite specific, and we merely note that bidding scenarios can be constructed where the symmetric filter might approximate trading behaviour.

To further explore the range of behaviours supported by static filters, we suggest three more models. The first is a simple alteration to the above symmetric filter, with a reduced core migration rate for Breed A to represent an aversion to selling preferred animals,

$$\pi_{01}^{A} = \frac{\bar{m_A}}{\bar{m}}(m_L) \tag{3.6a}$$

$$\pi_{10}^{A} = \frac{\bar{m_A}}{\bar{m}}(m_H) \tag{3.6b}$$

$$\pi_{01}^{B} = m_H, \tag{3.6c}$$

$$\pi_{10}^{B} = m_L, \tag{3.6d}$$

where $\bar{m_A}$ is the average migration rate of Breed A and $\bar{m_A} < \bar{m}$. We term this hoarding behaviour Static Filter 2. The two other models we expore involve a market in which only one breed is subject to competitive bidding, with the other migrating randomly. In the case of Static Filter 3, which acts on Breed A only,

$$\pi_{01}^{A} = m_L \tag{3.7a}$$

$$\pi_{10}^{A} = m_H \tag{3.7b}$$

$$\pi_{01}^{B} = \pi_{10}^{B} = \bar{m}. \tag{3.7c}$$

The converse, Static Filter 4, acts on Breed B only,

$$\pi_{01}^{A} = \pi_{10}^{A} = \bar{m} \tag{3.8a}$$

$$\pi_{01}^{B} = m_H \tag{3.8b}$$

$$\pi_{10}^{B} = m_L, \tag{3.8c}$$

with wealth allowing Patch 0 individuals to avoid purchasing lower-quality Breed 1 animals. These two models are both simpler than the symmetric Static Filter 1, above, and, in that there is no need to invoke adaptive bidding behaviours, more widely applicable.

While the above models capture a wide range of possible migration distortions due to market effects, their mathematical simplicity comes at the cost of imposing potentially unrealistic constraints on bidding behaviour. Movement biases depend solely on the wealth difference between patches, with supply and demand not explicitly included in the model. A dynamic filter, in which the migration parameters are updated over time according to supply and demand, better reflects economic theory. In order to derive the relationship between migration parameters and the composition of the market, it is necessary to briefly explore the role of wealth in the bidding process.

Wealth in auctions

We have already noted that wealth offers flexibility to avoid forced animal sales and beat competitors in bidding for a desired breed. This latter point - that success in obtaining a finite, equally desired good is correlated with disposable wealth - is a preferable quality of auction mechanisms. This forms the basis of our dynamic market filter model. Specifically, we assume that, given a finite set of animals at the market and a finite number of bidders with equal preferences but different entirely disposable incomes, the buyers are able to choose their animal in wealth order, until supply or demand are exhausted. We assume here that sellers only wish to maximise their animal sale value, and that the price offered is a premium to a universal minimum animal value such that any bid is better than none (or, equivalently, that sellers have no utility associated with keeping their animals). These assumptions can be used to derive a simple model of the movement of animals, and also to calculate the prices at which they are sold, see below and 3.3.6.

Deriving a dynamic market filter

A dynamic market filter is created when decisions regarding the trading of animals are conditional on the market composition, and, specifically, on the levels of supply and demand for each breed. Unlike the symmetric Static Filter 1, where supply is represented but demand is scales so as to maintain constant migration rates, we now propose a model in which demand is incorporated by including buyers as agents. This is both more realistic, and provides a clear scenario for which a corresponding dynamic market filter can be derived:

1. Each animal in each population is taken to market with probability $\bar{m}$
2. b_0 and b_1 buyers arrive at the market from Patch 0 and Patch 1 respectively, with disposable wealth drawn from wealth distributions $W(0)$ and $W(1)$.
3. Buyers are ordered in wealth from richest to poorest, and animals are ordered from most to least desirable. Within sets of animals of equal desirability, and within sets of buyers with each wealth, ordering is random, such that both sellers and buyers have equal preference for transactions with individuals from the home or neighbouring patch. Animals and buyers are matched such that wealthier buyers receive preferred animals. Animals are redistributed to new patches accordingly, with animals that are not bought returned to their home patch.

This is a stochastic algorithm that takes advantage of the role of wealth in auctions of multiple heterogenous goods as proposed above. An illustrated example of the outcome of this algorithm when $b_0 = b_1$, as in all subsequent modelling, and, by convention, $P(\xi_{W(0)} > \xi_{W(1)}) > 0.5$, is shown as Fig. 3.2.

There are several options to determine migration parameters from this algorithm. The most accurate approach is to simulate each auction stochastically. This raises some difficulties in capturing equilibrium behaviour, as it is necessary to perform many replicates and to make a subjective choice as to the definition of system equilibrium. An alternative is to propose deterministic approximations that capture the fundamental behaviour predicted by the algorithm. The aim would be to determine the number of successful

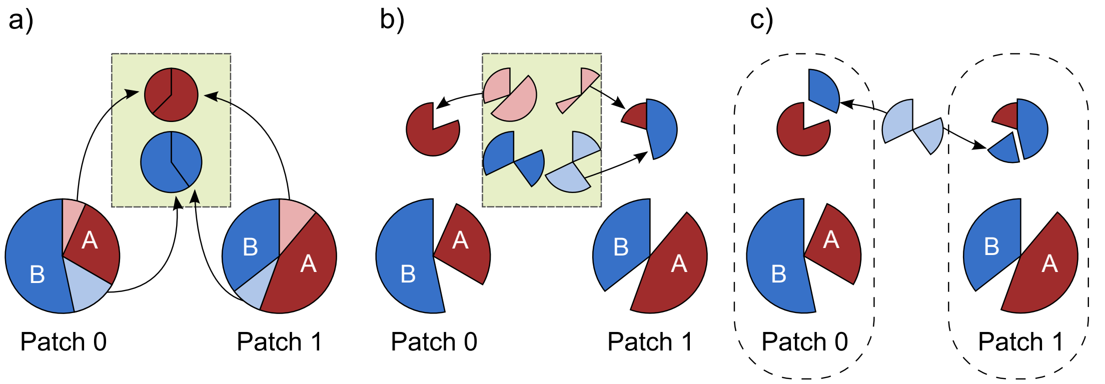


FIGURE 3.2: A single migration step given a dynamic filter following the auction algorithm described in the main text. As before, Breed A is indicated in red and Breed B in blue, and animals that move during a market stage are indicated by lighter shading. The market, consisting of animals that have not yet been assigned buyers, is indicated by the green rectangle, and Breed A is preferred in both patches, $\rho^{A}_{0,1} > \rho^{B}_{0,1} > \cancel{\rho^{A,B}_{0,1}}$. a) $\bar{m}P(\kappa|i)K_i$ animals move to the market, which consists of a total of $s_{A,B} = \bar{m}(K_0+K_1)$ animals. b) Animals of Breed A are first sold according to the number of buyers and wealth distributions of the two patches, followed by Breed B animals. In this case, all buyers from Patch 0 obtain the preferred Breed A animals, with unsold Breed A animals are assigned to Patch 1 buyers. Remaining Patch 1 buyers are assigned Breed B animals. As there is sufficient supply to satisfy all buyers, $s_{A,B} \geq 2b$, b animals are bought by each patch. In all cases, buyers show no preference for either local or non-local animals, such that the origin of each animal has no impact on its assignment to a buyer. c) Any animals that remain unsold due to excess supply are returned to their home site. Dashed lines indicate final occupants of each patch.

bids being made by wealthier Patch 0 buyers given b_0 and b_1, the number of animals present in the market, and the wealth distributions $W(0)$ and $W(1)$. This would have the substantial advantage of facilitating mathematical analysis of auction behaviour. However, the cost is accuracy, and in this work we are primarily in obtaining an unbiased and detailed exploration system behaviour. We therefore take an intermediate approach, offering further comments on the problem in Appendix 1, §3.3.6. We note in passing that the problem involves finding the size of ordered sets of order statistics, and that a more robust approach is suggested by recent work in the statistical literature [305].

Simulating auction behaviour

Instead of stochastic simulation of the auction at every migration update step, we chose to determine the average outcome of each relevant auction in advance of our main system modelling (as described by the update Eqs. (3.1 - 3.3)). We simulated bidding order behaviour for 50,000 auctions in which equal numbers of buyers from each patch, $b = b_0 = b_1$, submitted a bid drawn from the wealth distributions $W(0)$ and $W(1)$, calculating for each replicate the average number of bids from Patch 0 above the n^{th} highest bid, $\omega_0^{n,b}$. This can be used to retrieve y_i^{κ}, the absolute number of Breed κ animals won at the auction by Patch i buyers. For example, if there are b buyers and n Breed A animals at the market, and Patch 0 buyers prefer to obtain Breed A, then

$y_0^A = \omega_0^{n,b}$ if $b > n$ and $y_0^A = b$ if $b \leq n$. An important methodological point is that in taking the average number of successful bids for the finite good, we allow the trade of non-integer numbers of animals. Our value choice of K (which determines overall supply at the market) and b, then, are somewhat arbitrary, with their relative values being more important than their absolute values.

In our Dynamic Filter 1, selection is for Breed A in both patches and all buyers have preferences $\rho_{0,1}^A > \rho_{0,1}^B > \rho_{0,1}^{\cancel{AB}}$. Assuming, by convention, that $P(\xi_{W(0)} > \xi_{W(1)}) > 0.5$, animals here move according to

$$y_0^{A,B} = \min\Big(\omega_0^{2\bar{m}K,b},\ b\Big), \tag{3.9a}$$

$$y_0^A = \min\Big(\omega_0^{\bar{m}K\left(A_0(t)+A_1(t)\right),b},\ b\Big), \tag{3.9b}$$

$$y_0^B = y_0^{A,B} - y_0^A, \tag{3.9c}$$

$$y_1^{A,B} = \min\Big(2\bar{m}K - y_0^{A,B},\ b\Big), \tag{3.9d}$$

$$y_1^A = \min\Big(\bar{m}K\big(A_0(t) + A_1(t)\big) - y_0^A,\ b\Big), \tag{3.9e}$$

$$y_1^B = y_1^{A,B} - y_1^A. \tag{3.9f}$$

To obtain the correct system of equations for each auction step, the number of relevant animals must be substituted into the $\omega_0^{n,b}$ terms. Note that we do not include unsold animals in these equations, but that, as detailed below, we do not need this information to calculate migration rates. We also explore a second example, Dynamic Filter 2, in which selection favours Breed A in Patch 0 and Breed B in Patch 1. Breed preferences are now $\rho_0^A > \rho_0^B > \rho_0^{\cancel{AB}}$ and $\rho_1^B > \rho_1^A > \rho_1^{\cancel{AB}}$. In this case, animal movements are described by

$$y_0^{A,B} = \min\Big(\omega_0^{2\bar{m}K,b},\ b\Big), \tag{3.10a}$$

$$y_0^A = \min\Big(\bar{m}K\big(A_0(t) + A_1(t)\big),\ y_0^{A,B}\Big), \tag{3.10b}$$

$$y_0^B = \min\Big(y_0^{A,B} - y_0^A, \bar{m}K\big(B_0(t) + B_1(t)\big) - y_1^B\Big) \tag{3.10c}$$

$$y_1^{A,B} = \min\Big(2\bar{m}K - y_0^{A,B},\ b\Big), \tag{3.10d}$$

$$y_1^B = \min\Big(\bar{m}K\big(B_0(t) + B_1(t)\big),\ y_1^{A,B}\Big), \tag{3.10e}$$

$$y_1^A = \min\Big(y_1^{A,B} - y_1^B, \bar{m}K\big(A_0(t) + A_1(t)\big) - y_0^A\Big). \tag{3.10f}$$

The next step is to determine migration parameters, π_{ij}^{κ}, from the absolute number of animals moving. The correct normalisation for migrating animals takes into account the proportion of Breed κ animals originating at the source patch i rather than locally, and the total breed occupation in Patch i. When $i \neq j$,

$$\begin{aligned}\pi_{ij}^{\kappa} &= \Big(\frac{\bar{m}\kappa_i(t)K}{\bar{m}\kappa_i(t)K + \bar{m}\kappa_j(t)K}\, y_j^{\kappa}\Big)/\kappa_i(t)K \\ &= \frac{y_j^{\kappa}}{K(\kappa_i(t)+\kappa_j(t))}.\end{aligned} \tag{3.11}$$

To obtain the proportion of stationary animals, we simply solve $\pi_{ii}^{\kappa} = 1 - \pi_{ij}^{\kappa}$. Substituting migration parameters from Eq. (3.11) into the deterministic model described by the update Eqs. (3.1-3.3) yields very similar results to a stochastic model in which selection and population regulation are applied according to Eqs. (3.2 3.3) but the market algorithm described above is simulated stochastically (Fig. 3.10).

3.3.3.3 Simulation modelling

We have now defined four static migration filters designed to provide a coarse representation of plausible market-induced migration biases, and also two dynamic filters that capture a more realistic market algorithm. In each case, we have been able to suggest the migration parameters, π_{ij}^{κ}, needed to appropriately modify the genotype-dependent migration model described by Bolnick and Otto [267]. Our simulation work involved investigating the impact of markets in biasing animal movement patterns by iterating the update rules of this deterministic model, Eqs. (3.1, 3.2, 3.3), to obtain the static distribution of breed occupation at each patch. In the main text of the paper we explore symmetric initial conditions $\kappa_i(0) = 0.5$, with a wider range of initial conditions presented in Appendix 4 (§3.3.9results are qualitatively unchanged). Each simulation consisted of 500 iterations, although far fewer generations were usually needed to approximate the static state across most of the parameter spaces explored.

We focus on the role of core migration rate, $\bar{m}$, market-induced migration bias (δ_m in the static filter models, ΔW and b for the dynamic filter models), and selection strength, s, in determining the equilibrium breed occupation in the two patches, and retrieve phase-space diagrams based on these parameters (Figs. 3.3, 3.4, 3.6 and 3.7).

To further explore the behaviour of these market filters, we plot the temporal evolution of breed occupation in the two patches and patterns of trading between them, Figs. 3.5 and 3.8. This highlights the importance of migration in causing unequal patch occupation, and hence influencing the local effect of population regulation. We therefore assess the way in which local birth and death rates are influenced by migration at system equilibrium, Fig. 3.9. As overly large death or birth rates may be implausible in real systems, this gives an indication of the plausibility of results under different model parametrisations. We explore alternative implementations of population regulation and sensitivity to initial conditions in Appendices 2 and 4.

3.3.4 Results

We present our results for the static market filters, followed by those for dynamic market filters. In each case, we find notable deviations from system behaviour expected under unbiased migration.

3.3.4.1 Static Market Filters

We first assessed the implications of a static migration bias on a model in which selection favours Breed A in both patches, $s^A_{0,1} = 1$ and $s^B_{0,1} = 1 - s$, $0 < s \leq 1$. In the absence or migration, or when migration is not biased, a stable equilibrium in which Breed A is fixed is expected. As shown in Figure 3.3, applying a market model using Static Filter 1 can break this behaviour. A highly unexpected pattern emerges whereby moderately strong migration bias can lead Breed B to reach high equilibrium frequency in 'poorer' Patch 1, while Breed A may not reach fixation in 'richer' Patch 0. The normal and expected course of evolution can be disrupted by this form of migration bias.

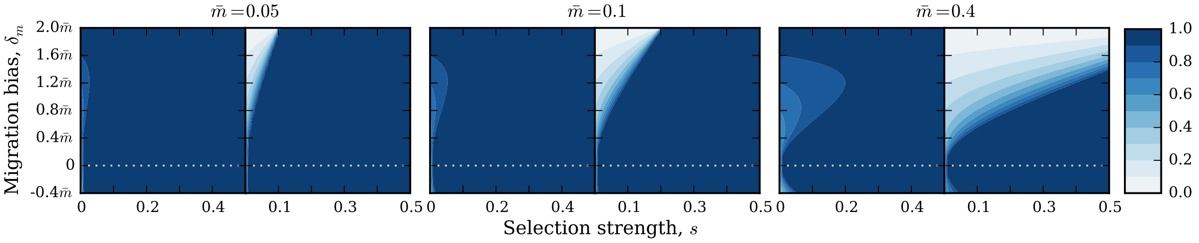


FIGURE 3.3: Equilibrium occupation given Static Filter 1 when selection is against Breed B in both patches. Each pair of plots shows Breed A occupation in Patch 0, $A_0(t)$, and Patch 1, $A_1(t)$, on the left and right respectively. A 40*40 evenly spaced grid of selection and migration bias parameter values was sampled using 500 iterations of update rules Eqs. (3.1, 3.2, 3.3), with migration parameters according to Eq. (3.4). Equilibrium behaviour when migration is unbiased corresponds to $\delta_m = 0$.

As discussed above, Static Filter 1 is a highly specific representation of the market. We compare our results for this model when $\bar{m} = 0.1$ to Static Filters 2, 3 and 4 in Figure 3.4. In three of the four Static Filter variants, similar anomalous behaviour is apparent. Static Filter 2 reduces the strength of the migration bias affecting Breed A, and reduces the size of the area of parameter space in which negatively selected Breed B fixes. Placing a migration bias on Breed A only, Static Filter 3, can cause Breed A to reach a lower equilibrium in 'richer' Patch 0, such that wealth differences promote a disadvantageous outcome for both patches. Static Filter 4 describes a situation in which Breed A migrates in an unbiased way but movement of Breed B is biased, and leads Breed A to become fixed whenever $s > 0$; the equilibrium expected under unbiased migration is restored.

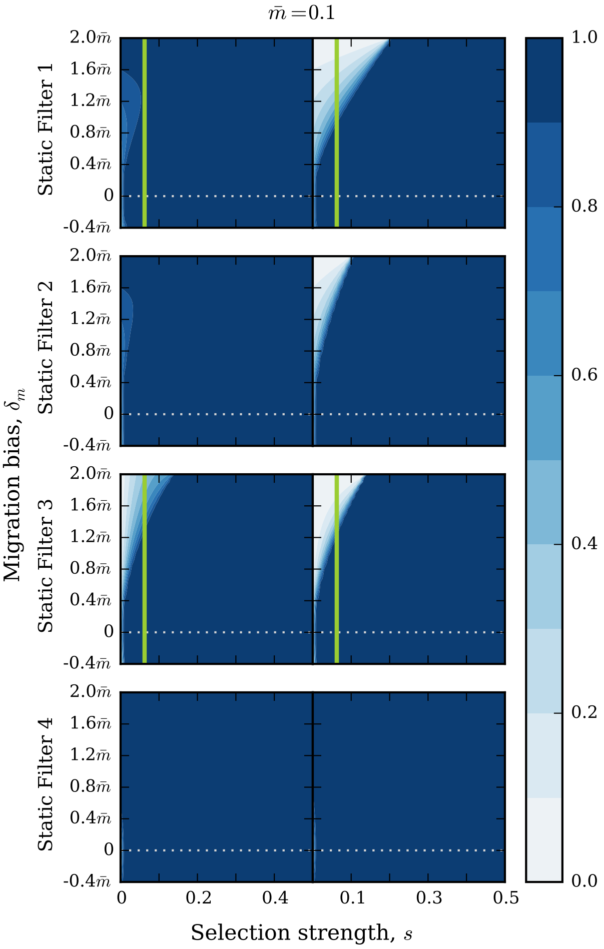


FIGURE 3.4: Equilibrium occupation given Static Filters 1 through 4, see Eqs. (3.4, 3.6, 3.7, 3.8), and selection against Breed B in both patches. Plots were generated as described in Fig. 3.3. The green vertical lines mark behaviour when $s = 0.0625$, the parameterisation shown to illustrate evolution of the system state over time in Fig. 3.5 below.

These results are initially surprising, but are anticipated by the modelling of Ngoc *et al* [266], who find that fast asymmetric migration in a Lotka-Volterra model, which is very similar to the system we investigate, can strongly impact model behaviour. Our observations can be further illuminated by observing the temporal evolution of patch occupation and migration behaviour, shown for Static Filters 1 and 3 in Figure 3.5. The rate and manner of approach to equilibrium depends on δ_m. Focussing on patch occupation, long-time behaviour sees a roughly exponential approach to equilibrium. At short times, which for $s = 0.0625$ corresponds to $t \lesssim 10$, a transient trend away from the equilibrium is sometimes apparent, suggesting that the system dynamics are being determined by multiple opposing processes.

To understand the equilibrium distortions we observe, it is easiest to consider the case of $\delta_m = 2.0\bar{m}$, under which migration is solely one-directional for animals that are subject to the migration filter. Given Static Filter 1, this corresponds $\pi_{01}^{B} = \pi_{10}^{A} = 2\bar{m}$ and $\pi_{01}^{A} = \pi_{10}^{B} = 0$. As selection is against Breed B in both patches, Patch 0 must fix for

Breed A at least as fast as if it subject to selection but no migration (Fig. 3.5, top-left plot, solid line). Patch 1 will only fix as Breed A if the effect of selection overpowers that of outward migration, and indeed the regime shift occurs at $s = 2\bar{m}$ (Fig. 3.3). When the average migration rate of Breed A is halved, Static Filter 2, the transition of equilibrium from $A_0(\infty) = 1, A_1(\infty) = 0$ to $A_{0,1}(\infty) = 1$ for $\delta_m = 2.0\bar{m}$ is correspondingly shifted to $s = \bar{m}$.

When $\delta_m < 2\bar{m}$ and $\bar{m} > 0$, $\pi_{10}^B > 0$ and the equilibrium frequency of Breed A in Patch 0 must be less than 1 whenever Patch 1 contains any Breed B animals. In our Static Filter 1 model, the impact of this back-migration tends to be low. Scenarios in which migration forces Patch 1 to high Breed B frequency correspond to those in which δ_m is large and π_{10}^B is therefore small. However, applying a directional migration bias to Breed A only, Static Filter 3, breaks this dynamic, Figs. 3.4, 3.5. Under such conditions, both patches may exhibit low-Breed A equilibria. Essentially, the effect of biased migration in suppressing selection in Patch 1 is sufficient to cause emigration into Patch 0 to be dominated by Breed B animals, which in turn can over-power selection in Patch 0. This narrative can be seen, for example, in Fig. 3.5 (Static Filter 3, $\delta_m = 2$, solid black line; also see Fig. 3.24), and is also relevant to our understanding of unusual behaviour in the dynamic migration models.

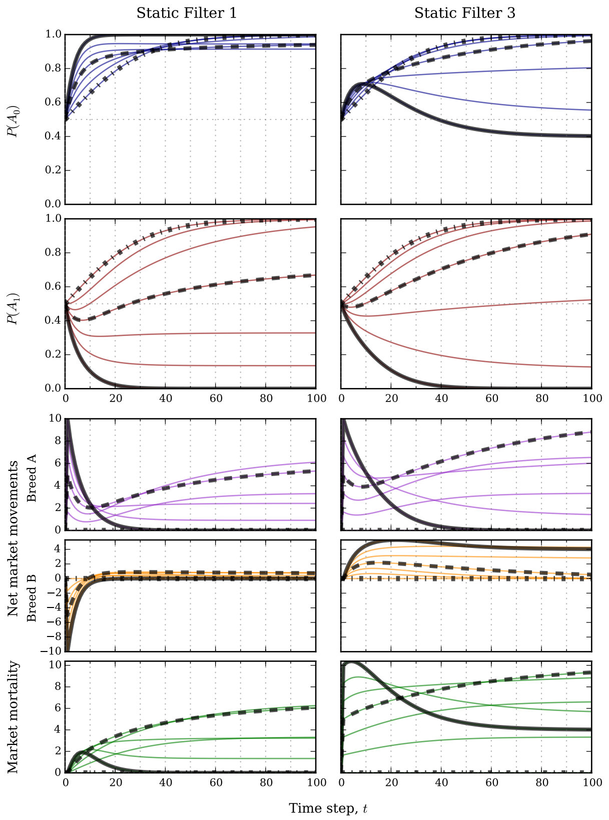


Figure 3.5: Temporal evolution over 100 time-steps for two systems incorporating a static migration bias, Static Filter 1 (left) and Static Filter 3 (right), when selection is in favour of Breed A in both patches and $s = 0.0625$. The core mase migration rate, $\bar{m} = 0.1$, and each line represents a migration bias, δm, sampled evenly in the interval $0 \leq \delta_m \leq 2$. Three cases are higlighted using black lines - no migration bias ($\delta_m = 0$, dot-dash), moderate migration bias ($\delta_m = 1$, dashed) and strong migration bias ($\delta_m = 2$, solid). The top two rows show, respectively, the changing proportion of Breed A in Patch 0 and 1 over time. The bottom three rows focus on net movements of the two breed types between patches (with positive movements indicating migration from Patch 0 to Patch 1, and negative movements the converse), and on the imbalance in patch occupation created by these movements - which leads to market mortality.

We can gain further insight into this narrative and the role of biased migration in modifying patch occupation by considering market behaviour through time, shown in the lower plots of Fig. 3.5. In particular, this highlights the critical role of population regulation in allowing migration biases to change the global frequency of breeds. Net migration out of a patch leads to excess population growth for those animals that remain (and reduced growth for emigrants). The extent of market induced mortality - which is just the net movement of animals in a generation, and hence the imbalance in total patch occupation created by migration - gives an indication of how strong this effect is. Note that high-δ_m, which often leads to lower Breed A occupation equilibria in these models, does not necessarily correspond to high market-induced mortality. A clear example is Static Filter 1 with $s = 0.0625$ and $\delta_m = 2.0\bar{m}$, which has equilibrium occupation $A_0(t) = 1$ and $A_1(t) = 0$, resulting, by Eq. (3.4), to no migration at all.

3.3.4.2 Dynamic Market Filters

We have argued that dynamic migration filters are better able to represent the market process. Interestingly, the dynamic filters that we explore also lead to strongly anomalous equilibrium behaviour as compared to expectations under unbiased migration. Indeed, the outcome for the wealthier Patch 0 is often be worse, with lower high-quality

Breed A occupation, than is seen given a static migration filter.

While we explore the systems using normally distributed wealth, we emphasise that the ordering of bids is the only important factor in retrieving migration parameters, such that identical results will arise using log-normal wealth distributions.

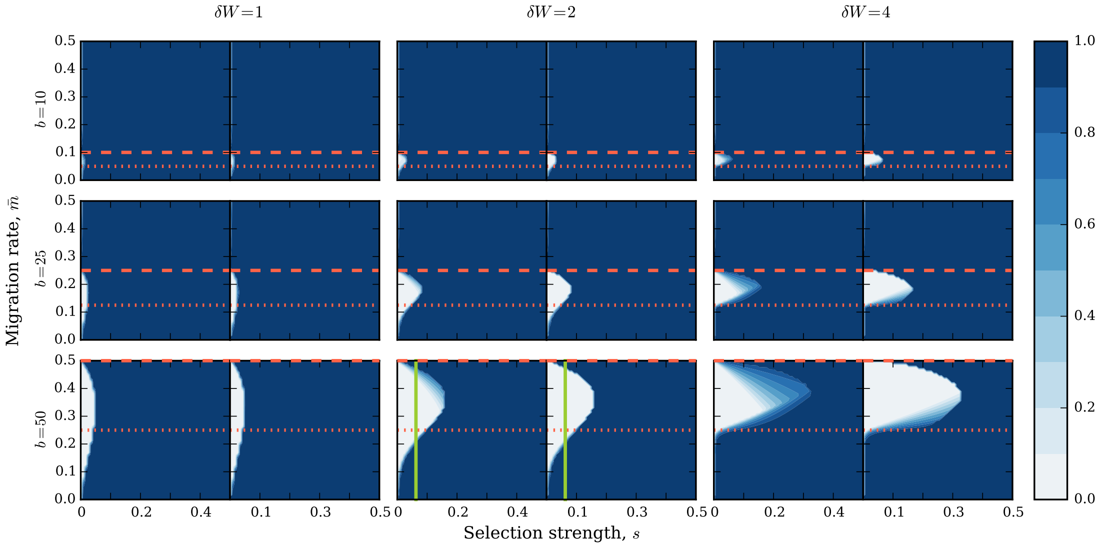


FIGURE 3.6: Equilibrium Breed A occupation given Dynamic Filter 1, with $s_(0,1)^A = 1$ and $s_(0,1)^B = 1 - s$. Three wealth differences are shown, with $W(1) = N(0,1)$ and $W(0) = N(\delta W, 1)$, such that Patch 0 buyers are wealthiest in the right most column. The number of buyers coming to market from each patch also varies, with least buyers in the top row of plots and most in the bottom row. The red dotted line indicates the migration parameter at which supply equals demand $2b = 2\bar{m}K$, and the dahed line $b = 2\bar{m}K$. As in Fig. 3.4, the green line indicates the $s = 0.0625$ regime explored in Fig. 3.8.

When selection is against Breed B in both patches, and both patches contribute the same number of buyers, b, with purchasing preferences $\rho_{0,1}(A) > \rho_{0,1}(B) > \rho_{0,1}(\cancel{AB})$, we have Dynamic Filter 1. Using Eqs. (3.9, 3.11), we explore the impact of core migration rate, selection, number of buyers and wealth difference in Figure 3.6. When supply exceeds demand, $\bar{m}K > b$, both patches fix for Breed A. However, when demand is greater than supply the exact opposite can occur, with negatively-selected Breed B rising to high equilibrium frequency or fixation in both patches. This is behaviour constrasts starkly with models incorporating unbiased migration and supports our observations based on the static filters, Figs. 3.3 and 3.4 - specifically, that biased migration due to market effects may counteract artificial or natural selection pressures, and lead to evolutionarily anomalous outcomes.

When selection is for Breed A in Patch 0 and Breed B in Patch 1, we might naively expect market forces to promote the sorting of animals into their respectively favoured patches. In this case, preferences are $\rho_0(A) > \rho_0(B) > \rho_0(\cancel{AB})$ for Patch 0 and $\rho_1(B) > \rho_1(A) > \rho_1(\cancel{AB})$ for Patch 1, and we have Dynamic Filter 2. Following the theme of our previous results, Figure 3.7 shows that the opposite can occur. Again, when supply exceeds demand, Breed A fixes in Patch 0 and Breed B fixed in Patch 1. When demand is greater than supply, we find that the richer Patch 0 often fixed for Breed B, which is locally negatively selected. This model provides an interesting example of wealth differences leading to a less positive outcome for the richer patch, but an optimum result for the poorer patch, a phenomenon that appears as soon as demand exeeds supply and and is in evidence even for relatively low wealth inequality.

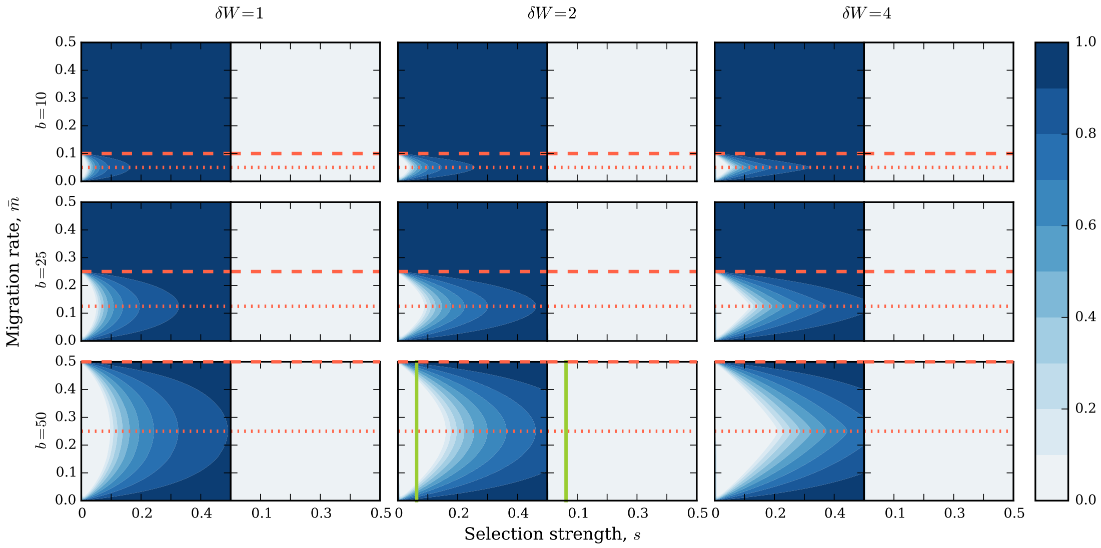


FIGURE 3.7: Equilibrium occupation given Dynamic Filter 2, with $s_0^A = s_1^B = 1$ and $s_0^B = s_1^A = 1 - s$. See caption of Fig. 3.6 for details.

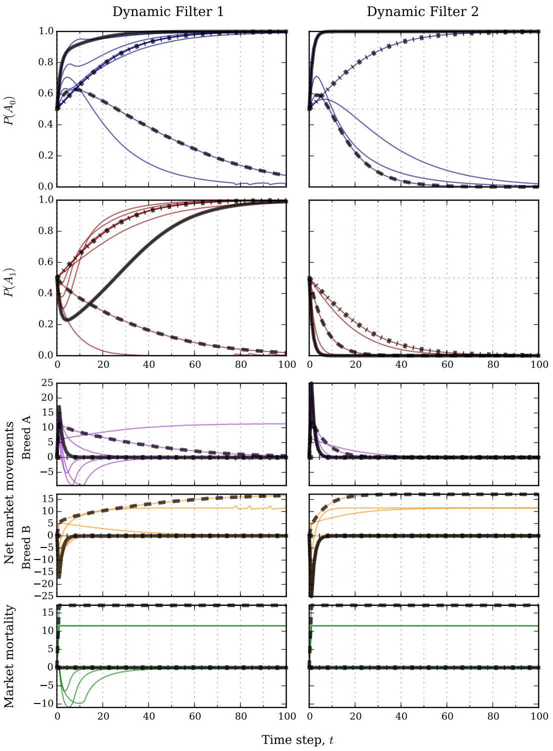


Figure 3.8: Temporal evolution over 100 time-steps for the two dynmic migration filter systems, Dynamic Filter 1 (left) and Dynamic Filter 2 (right). For Dynamic Filter 1, selection favours Breed A in both patches, while for Dynamic Filter 2 selection favours Breed A in Patch 0 and Breed B in Patch 1. In both cases $s = 0.0625$, and each there are 50 buyers from each patch. Each line corresponds to a different migration rate, sampled uniformly in the interval $0 \leq \bar{m} \leq 0.5$. Now, black lines indicate no migration ($\bar{m} = 0$, dash-dot), moderate migration ($\bar{m} = 0.25$, dashed) and high migration ($\bar{m} = 0.5$). See the caption of Fig. 3.5 for further details.

As before, these behaviours can be investigated by observing the time-evolution of the system state, Fig. 3.8. A pattern that is similar in spirit to that noted in our static filter models is apparent. Wealthy buyers from Patch 0 purchase the desirable Patch 1 Breed A animals, leading to or supporting the increasing Breed B frequency in Patch 1. As Breed A becomes rare in Patch 1, most animals that it contributes to the market are Breed B, and now wealthy buyers from Patch 0 purchase these, driving down the local frequency of the higher-quality and preferred Breed A animals.

Turning to the details of our auction algorithm, it is clear that a core reason for this is the strict ordering of breed preferences, whereby wealthy Patch 0 buyers always prefer to purchase a low-quality animal over no animal at all. Given the critical role of population regulation in allowing migration to distort global breed frequencies, this behaviour can be seen as a Tragedy of the Commons. Despite the increased pressure on limited local resources, and the correspondingly increased probability that personally owned animals will die, selfish buyers nevertheless seek to obtain any animal available, even if this will lower the quality of the local stock.

3.3.4.3 Market mortality and initial conditions

In all cases, model behaviour is not radically altered by using alternative initial conditions, Appendix 4 (§3.3.9), although some differences are apparent in the Dynamic Filter 1 model where selection is globally equal. We do not prove robustness to initial conditions in any of the models, though our numerical investigations suggest dependence cannot be large (excluding special cases such as the unstable fixation of negatively selected species).

We have argued above that market-induced mortality due to population regulation plays a critical role in model behaviour. To investigate this further, we determined the equilibrium death rate in Patch 0 due to migration given Static Filters 1 and 3, and each of the two dynamic filters, Fig. 3.9. The patterns observed for the dynamic filter models were indistinguishable, and we therefore show results for only Dynamic Filter 1. In the market models represented by the static filters, the combination of high levels of selection and large migration bias are associated with a high rate of migration-induced mortality. This suggests that such systems are not stable, with probable population extinction in Patch 1. However, mortality in the region of the parameter space showing most interesting equilibrium behaviour was not excessive, suggesting that the results we observe cannot be discounted on these grounds. In the dynamic market filter, market-induced mortality was largest when $\bar{m}N = b/2$ such that demand was exactly double supply. This is expected, as under such conditions richer Patch 0 has sufficient buyers to submit a bid against a poorer Patch 1 buyer on every animal, which in turn maximises the trade imbalance that leads to market induced mortality. Mortality is high, but not always extremely so, throughout the parameter space leading to most interesting system behaviour, see Figs. 3.6 and 3.7.

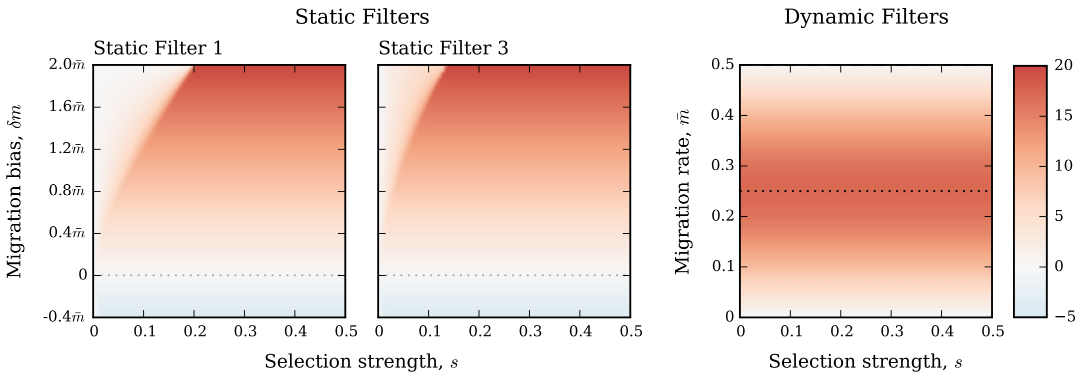


FIGURE 3.9: Extent of net trade imbalance at occupation equilibrium under various migration filters, which leads to market-induced mortality. For the static filter models, core migration rate $\bar{m} = 0.1$. For the dynamic filter model, number of buyers $b = 50$ and wealth difference $\delta W = 2$. In all cases, carrying capacity in each patch is $K = 100$. The trade imbalance is positive when more animals immigrate into Patch 0 than emigrate out of it.

The ultimate cause of this migration-associated mortality is our approach to population regulation. In a regulation step, the total population in each patch is merely normalised to 1. This implies a potentially very large birth or death rate, which in turn implies that markets are rare events. We re-implemented the Static Filter models to incorporate an alternative model of population regulation through the Beverton-Holt [306] equation, which allows population filling to differ from the carrying capacity K and can be used to include less dramatic density dependent growth or mortality,

$$n_i(t+1) = \frac{n_i(t) R_0 K}{K + \big(R_0 - 1\big) n_i(t)}, \tag{3.12}$$

where $n_i(t)$ is the population size in Patch i at time t and R_0 is the maximum growth rate. As our results did not qualitatively change, they are included in Appendix 2, §3.3.7. Briefly, we found that the region of the parameter space leading to equilibria unexpected under unbiased migration depends on the maximum growth rate, with low R_0 reducing the size of this region. Furthermore, using Eq. (3.12) to relax the strict population size constraint applied in our standard model of population regulation, Eq. (3.3), lead to a considerable reduction in migration-induced mortality as well as an imbalance between population sizes, with the poorer patch containing fewer cows. We also simulated the Dynamic Filter 1 and Dynamic Filter 2 models with Beverton-Holt growth, confirming that the previously observed system behaviours can remain, depending again on the maximum growth rate R_0.

3.3.5 Discussion

In the above work, we have shown a variety of possible system behaviours that might be caused by unconventional biases to migration rates. We have derived these migration rate biases from various representations of the market process. Some of these behaviours strongly contradict evolutionary expectations under unbiased migration. Our work does not show that such situations should arise in the real world, but does highlight the importance of properly considering migration biases, both whether these result from economic forces or other factors.

Main Findings

We explored a system in which carrying capacity was equal in both patches and there was no migration-associated mortality. Our main findings for such systems can be summarised as follows:

1. When selection favours one animal type in both locations, migration biases can lead to the fixation of the inferior animal type in one (Fig. 3.3) or both (Fig.

3.6) locations. The evolution of the system involves the richer location buying many desirable animals from the poorer location, and then, potentially, buying undesirable animals from the poorer location as desirable ones become unavailable.

2. When selection favours different animal types in the two locations, market-induced migration biases as represented by Dynamic Filter 2 can lead to the fixation of the locally worse animal type in the richer location (Fig. 3.7). Economically, this is related to individually selfish behaviour, in which rich buyers prefer to purchase any animal rather than none, even if this leads to increased local mortality - a tragedy of the commons.

3. Despite the potential for the migration filters implemented to create these two system distortions, usually the majority of the parameter space explored yields expected results whereby the fitter breed type rises to high frequency in each patch, Figs. 3.4, 3.6, 3.7.

The fundamental implication is that migration biases can lead to system occupation equilibria that are very different or even opposite to those expected given unbiased migration, a result that echoes some previous work [266]. Furthermore, in our Dynamic Filter models even minor wealth differences may be able to drive a system in which animals are traded to a state that is universally undesired. To understand whether such scenarios might exist in the real world, it is useful to discuss aspects of actual animal markets. For this, we turn to the case of the Indian market for cattle.

The contemporary Indian market for cows

Wealth and goods in the rural Indian economy flow through a network consisting of tens of thousands of local markets - estimates range 21,000 to 47,000 [307, 308], and perhaps 70,000 [309]. The vast majority of these are *haats*, periodic markets held in larger villages that often occur weekly [307]. Animals may be traded at general rural markets, but the principle route is through specialist livestock markets. One report suggests that 1,300 of these were recorded in 11 states [198], though a count of 868 in Uttar Pradesh alone [199] may imply a somewhat larger total. Exchange through social contacts is also common [310]. The volume of livestock markets varies considerably, with reports suggesting some markets hosting thousands of transactions per day [199], although the largest cattle market in Maharastra is thought to process 65-70,000 transactions a year [311]. A more standard range for local markets may be an annual total of 8 to 35 thousand [312].

The probability of cattle being bought or sold is substantially impacted by climatic conditions [313, 314]. Nevertheless, some estimates of annual market participation have been published or are easily inferred from studies - 10% [314], under 26.3% in ICRISAT data and 10.1% in REDS data (estimates are elevated due to double-counting sales within villages) [310], perhaps 9% [315] (assuming 80% of arrivals are ultimately sold and the markets serve around 1.8 million cattle in Raipur District, [316]), 21% [312]. Trading often focusses on castrated males, although this depends on local agricultural practice (54.8% [312], 70% [315], 68.9% [313], 49% in ICRISAT data but 14% in REDS data [310]). Clearly transactions involving reproductive animals are of primary relevance to long-term population evolution, although as our work highlights crowding effects can also be important in population models.

The structure of a trade between farmers can follow one of several channels, and essentially involves four actors - buying and selling farmers, cattle traders, and brokers [315, 317, 318, 311, 319, 320]. Cattle are either brought to the market by the farmers themselves, or are bought by cattle traders who visit villages. Animals are then sold to a buyer at the livestock market, usually directly to the farmer but sometimes through cattle traders. Any transactions that occur at the market may be mediated by brokers, who negotiate a mutually acceptable price in exchange for commission [321]. Price negotiation may be through hidden bids using the *hatta* method, and as alternatives to brokers public bids by buyers or auctions may occur [318]. More middlemen and the involvement of brokers tends to increase the spread between buyer and seller price, and various market and transportation costs are incurred by parties who attend the market. The rate of successful sales to livestock arrivals varies (80% [318], 23.2% [313], 64-75% [322]), and animals that are not sold may be taken to subsequent market events or neighbouring markets [318].

This brief summary clearly demonstrates the complexity of the market system. Not only are multiple transaction channels used, but their popularity varies between markets (eg. compare [315] to [319]) and also within markets according to the type of animal being traded (eg. [320]). Different cattle types - bullocks and cross-bred milch cows, for example - may also travel different distances to market, again depending transaction channel [318].

Despite these complications, we can suggest that some parameters explored in our models are at least plausible. For example, the probability of a female cow going to market might indeed be around $\bar{m} = 0.1$ over a year, and the variation in supply and demand based on climate suggests that demand may, at times, exceed supply, as is important for anomalous behaviour given Dynamic Filters 1 and 2. Although not all animals are sold on a given market day, this does not preclude the possibility that farmers prefer to purchase worse quality animals over no animal at all, which is the effect of excess demand in our models and likely a driving factor of the evolutionary trajectories they describe. Roughly estimating the generation time of cattle at five years would suggest that our standard approach to population regulation, Eq. (3.3), implies relatively slow market events. However, applying the smoother Beverton-Holt population growth model does not break our fundamental results (see Appendix 2, §3.3.7). In conclusion, then, while we are well aware that more complex representations of the market, or of income, or of animal choice will broaden our understanding of the dynamics we identify, a pattern of evolution in which the preferred and positively selected animal type does not reach fixation appears possible.

Limitations and extensions of our model

The subject of markets and gene flow is a rich one, and we feel that progress in clarifying the implications of trade in the distribution of genetic variation might be made in a number of directions. Before discussing these, we focus on the limitations and extensions of our specific model. By necessity, our initial investigation of gene flow in an economic context is a simplification. Behaviour of the implicit human agents in the model is rigid, with well-defined and stable preferences for purchasing different animals. In reality, individuals will have heterogeneous demands for animal types depending on their subsistence strategy and personal preferences. Furthermore, they are likely to learn from their experience with different animals, and an adaptive representation of the decision process would more accurately reflect human behaviour.

Perhaps the most dubious economic assumption we take is that supply and demand are exogenous and inflexible. Our model does not allow for more agents to become cattle producers when demand is high or for agents to leave the cattle trading business when demand is low. Nevertheless, becoming a livestock herder will usually involve entry costs, time, experience and, in some cases, may be an occupation that tends to be inherited. Furthermore, the markets we aim to describe are small-scale, with correspondingly limited supply of labour. As such, modelling the trading process as an idealised, flexible market is also questionable. It will ultimately be necessary to use data to quantitatively guide our models of animal movement as opposed to the qualitative discussion we offer.

A second economic assumption of our models is the vastly simplified representation of auctions, with disposable wealth perfectly correlated with the order in which individuals can choose their preferred animal to purchase. The true complexity of animal markets can be astounding, as hinted at in the example of Indian cattle markets, above. The simplest extension to our model would be to add stochastic noise to the auction process, such that the order of choice is no longer stricly according to wealth. However, more detailed representation of the interaction between the various market agents would be needed to assess what sort of stochasticity is appropriate and whether systematic biases to bidding order are likely.

Our final economic point relates to the role of animals in the subsistence of herders, and specifically the connection between disposable income and animal ownership. In our models, these are not linked. Although it depends on the assumptions used, we would usually expect models in which disposable wealth is an increasing function of herd size to promote to wealth condensation, increasing wealth inequality between and within patches. While the distribution of cattle ownership has been suggested to approach a power-law among East African pastoralists ([323], cited in [324]), less extreme log-normal wealth distributions have been suggested for several agricultural societies (eg. [325]). Ultimately, distinguishing between such fat-tailed distributions in data requires care [119], but the important point is that wealth inequality can be substantial, with the probable implication that present resources promote the acquisition of future resources. We note that income is often approximated by a log-normal distribution [326], and if this is the case then the behaviour of our dynamic filter models (which rely on ordering of bids rather than absolute bids) over a given generation should hold. If the wealth distributions of patches diverge, we would expect the parameter space over which undesirable evolutionary outcomes emerge to become larger (eg. see Fig. 3.6).

There is a strong historic connection between livestock, investment and wealth - associations between the words for livestock and wealth are evident in many languages [202, 205, 206], but a particularly illustrative example here is the the shift in meaning in of the Sumerian word for 'goat', *máš'*, which came to mean 'tax' or 'rent', and ultimately 'interest', in the late 3rd millenium BC [207]. A critical difference between these phenomena is ecological carrying capacity, and our results cannot be simply extrapolated to describe eg. trading of investments, which are most simply modelled as growing exponentially. As mentioned before, our results are, qualitatively, robust to the application of a more realistic Beverton-Holt model of population growth, see Appendix 2 (§3.3.7). Other representations, such as the swapping of animals between patches, avoid migration-induced mortality entirely, however, and do not display the anomalous behaviours identified in this study.

The final obvious limitation of our model is the use of two patches and two breeds. The network of potential animal movements is clearly far more complex than two isolated interacting villages or herds, and the implications of this should be explored. Alternative approaches could involve explicitly describing a possibly dynamic social network, as in several agent-based models of trade from the archaeology literature [324, 327], of through the use of observations of animal movements, as is often applied in livestock epidemiology models (egs. [285, 287, 281, 286, 284, 282]). Furthermore, the use of two animal types does not represent the range of domestic animals available to farmers. The spread of complex genetic traits - such as those related to milk and meat production - through economically linked populations is of particular practical interest, and exploring models that better describe such a system is an important future challenge.

Models and reality

A natural question arises as to whether there are concrete examples of human-induced migration biases among livestock leading to reversals in the selection-driven course of evolution, and, if so, whether these have been driven by trade. Alternatively, what data or situations might afford the identification of such a phenomenon? The challenge is two-fold. Firstly, it is difficult to determine the expected distribution of species or breeds when migration is unbiased. Secondly, even if migration biases are observed or inferred, there is a complication in teasing out the impact of these biases as compared to other effects in what may be quite complex and dynamic systems.

Given these difficulties, progress is best made by returning to both the core structure of our models and the core narrative implied by their results. In terms of fundamental features of the systems we explore, the potential for breed and directional biases in migration patterns is critical, as is the possibility of migrating animals impacting herd structure through population regulation. High and relatively inflexible demand is also important when the market is represented by the dynamic migration filters we explore. The most interesting narratives involve two processes - biased emigration reducing the number of locally preferred animals among poorer herders, and high immigration of low-quality animals diluting high-quality local stock in wealthier sectors of society. If the distribution of animals more closely reflects our model systems close to equilibrium, then it may be difficult to observe such trades. Rather, we might expect either to encounter a situation whereby there is a deficiency of desirable animals among groups of herders where a desired exogenous supply is present but unaccessable, or a deficiency among all groups of an animal type that is desirable, locally viable and, when introduced, goes extinct.

The wide range of agro-pastoralist groups, following different marketing and livestock ownership strategies, make it difficult to draw general conclusions about local animal trading. Nevertheless, it is possible to make relevant observations. Among surveyed villages of Dagota pastoralists in Tanzania, households with fewer cattle sold proportionally more of them, though less in absolute terms, and at lower prices [328]. Sales were made to provide income for grain purchases, and it was considered possible that the quality of animals being sold from households with few animals was lower. Although this general pattern may support our model, in which desired animals flow to the wealthy and the poorer group are forced to make sales, more information on the reason for pricing patterns would be needed to draw unambiguous conclusions. Anecdotal reports from development interventions involving livestock distribution suggest that the poorest recipients have sometimes had to sell animals upon receipt [329].

Among Ethiopian smallholders, wealth, as well as recent livestock mortality, increased the probability of herders being net livestock buyers [330], while owners of larger herds were more likely to focus on sales [331, 330] and were also more likely to make at least one purchase [330]. Overall, market participation was lower for households with small herds [331, 330]. The impression here, then, is that wealth facilitates purchases, supporting one aspect of our model. Herders with more animals are also more integrated into the market system, and while they buy more animals - these being for reproductive purposes - they also sell more. The details of market behaviour appear complex - for example, those with less land also tended to sell rather than buy animals, possibly reflecting differences in subsistence strategy within the sample. The pattern of owners with larger herds selling more animals was also observed among among Zambian farmers [332].

Overall, biases in the direction of animal transfers associated with wealth appear likely. In some cases, characteristic geographic movement patterns among different breeds through market systems have also been observed [313, 333], as well as local preferences in the breeds being bought by farmers for incorporation into herds [334, 335, 300], such that breed and directional migration biases can occur.

In our models, price variation between animal types is an implicit outcome of the auction process and a necessary driver of migration biases. Different species of livestock yield markedly different market prices, but price premiums associated with specific breeds are also commonly reported [318, 315, 336, 300, 334, 335, 337]. Such price information is potentially useful in giving a qualitative indication of local preferences and supply/demand dynamics. Price may be determined by a range of factors, but in the context of our model it is interesting to note examples in which a more expensive and/or preferred animal breed forms the minority of locally kept livestock (non-Zebu cattle breeds in [300]; West African Sheep [334]; preference for sheep or mixed herd composition among goat herders [338]). The situation in India is interesting, in that perhaps 80% of cattle are considered non-descript *desi* animals [339], but these animals fetch lower prices at market [318, 315, 336]. Such patterns do not confirm the presence of the dynamics seen in our models, but would be expected under some scenarios, and do indicate that poor farmers are unable to keep or breed their ideal animals. Cases of more common animals being highly valued also exist (eg. [335]).

The sale of livestock to obtain cash is a common strategy when faced with one-off bills or resource stress. As a result, drought can be a major factor in influencing commercial destocking [195], with response often correlated with household wealth [340, 341, 342]. In creating anomalous but predictable supply shocks, and the potential for a post-drought spike in demand associated with restocking, drought conditions offer an interesting view into animal movements given a market in a perturbed state. The impact of market systems in spatially redistributing animals sold due to drought or economic stress has been observed [313, 340, 333] and may be breed-specific. We are not aware of cases whereby trade-related movements have been directly linked to increased population pressure in neighbouring, non-drought districts; indeed, the nature of density-dependent population growth in some livestock producing systems is debated [343]. Nevertheless, the temporary re-location of migratory herds due to drought has been reported to cause high mortality in a region outside the drought zone [344].

The role of markets in rebuilding stock after drought varies. East African pastoralists have tended to rely more on reproduction in the surviving herd or on animal exchange through social contacts [345, 346], though purchases do play some role [298, 346]. In agricultural communities, differences in engagement with the market during and after drought may be associated with wealth [347]. Animal prices often collapse during drought and increase during recovery (eg. [298]), suggesting a relatively weak supply of animals after drought events. This lack of supply coupled with a need to restock has lead to difficulty in obtaining the preferred Boran breed of cattle among Ethiopian pastoralists, leading to genetic dilution of local stock [348, 349], which is one pattern that can appear in our models. A more artificial case of excess demand leading to the import of low-quality non-local animals comes from a restocking intervention aimed at helping Eritrean refugees, which involved large purchases from traders who either couldn't source animals locally or chose not to do so in order to maximise profits [350]. In further support of wealth-determined trade patterns, many of the refugee families had to sell a portion of their new herds to buy food.

The above examples provide evidence for a substantial role of the wealth and breed preferences of herders in influencing animal movement through trade. In some cases, stated animal or breed preferences and the animals owned do not match, suggesting that wealth is assortative in terms of herd composition. Furthermore, we have identified examples in which specific animal breeds move directionally through markets, and highlighted specific cases in which temporary excess demand has lead to the import of non-local animals and, in once case, breed dilution. Overall, it would appear likely that wealth is a factor in influencing the animals available to a herder on the market, and hence herd composition. However, direct evidence for the second narrative of our model - in which rich herders ultimately own a less-desirable animal type due to the low market supply that is, importantly, originally created by wealth differences - is not conclusive. Data on the ability of livestock markets to satisfy breed and species demand for different wealth categories would help to clarify this. Our models suggest that situations in which demand for animals is high - such as after drought or conflict, or potentially during restocking or breed introduction programs - have greatest potential for increased competition for animals between groups, and hence a more substantial role of wealth differences in their movements. We note in passing that perhaps the most interesting animal migration of all - the early spread of different livestock species after their domestication - may provide an exceptional example of just such a high-demand scenario.

Beyond markets

We conclude by returning to the core attribute of our model - breed and directional biases in migration. More than anything, this study shows that such biases, especially when they are dynamic, can have profound and surprising impacts on model behaviour. Such perversion of the normal course of evolution, as dominated by the classic forces of natural selection and population growth, may be of considerable importance in the spatial distribution of species and genes. It is possible that these effects go far beyond human influence through trade, or human influence at all.

Supplementary Information

On the counter-evolutionary effects of market mediated gene flow

G. S. Jacobs, G. A. Kaiping and T. J. Sluckin

3.3.6 Appendix 1: Auctions and bidding order

A central requirement when modelling the market as a Dynamic Filter was estimating the migration implications of competitive bidding for a limited number of animals. This involved making a simple approximation of what may be, in reality, quite a complex process. In our example of the Indian cattle market, various different methods of bidding, including secret and public bids, are used, and middlemen play a complex role in matching buyers and sellers (see references in main text, §3.3.5). Although we hope to explore this system in more detail in the future, the work presented here focusses on theoretical results. It is nevertheless useful to highlight some of the implicit assumptions made in our representation of auctions.

We first recall that our version of auction behaviour was based on the principle of ordering bidders according to wealth and then allowing them to choose which animal to buy in that order. This closely reflects desirable auction behaviour, in that sellers are implicitly receiving high bids for the best animals.

In the context of real auction behaviour in the Indian cattle market, we require that no stage of the bidding process - from observing cattle to commissioning middle-men - introduces a bias to who wins an animal, other than a sorting through wealth. Thus, we can imagine wealthier bidders getting more information on the auction such that they are able to choose from a greater subset of animals, or quirks of middle-men commissioning leading them to obtain their preferred breed type. However, we do not allow middlemen to favour specific contacts, or for sellers to give preferential treatment to buyers from their patch, or for information on animals to be biased according to local breed knowledge. The fact that aspects of the bidding process and animal qualities may be hidden is again assumed not to lead to wealth-independent directional biases. Any complexities of the auction process must be absorbed by the rate of migration to market, $\bar{m}$, or the wealth distributions, $W(0)$ and $W(1)$. For example, $W(0)$ and $W(1)$ could be re-interpreted as the marginal utility associated with wealth, or the extent of social capital available in obtaining preferred animals, or the acuity with which a group is able to correctly assess the quality of animals - although care would need to be taken regarding any additional assumptions doing so might introduce.

In addition to these limits on the details of matching buyers with sellers, we assume that the specifics of the auction method are not critically important to the outcome. The main point here is that prices are ignored, and that while bidding strategies might impact the amount paid for an animal, they do not disturb the wealth-ordering behaviour of the auction. While auction algorithms could be constructed to disrupt this core principle, we consider it to capture the essence of what auctions are trying to achieve.

The question of who wins which animal, then, is essentially one of order statistics given bidders from two different wealth distributions. For example, when preferences are $\rho_{0,1}^{A} > \rho_{0,1}^{B} > \rho_{0,1}^{\cancel{A,B}}$, such that Breed A is most desirable in both patches, we want to determine how many of the top $\bar{m}K\big(A_0(t) + A_1(t)\big)$ bids, and of the top $2\bar{m}K$ bids, fall to buyers from each population. This is used to calculate the migration terms, $\pi_{i,j}^{\kappa}$ in Eq. (3.1), using Eq. (3.11). We explored several approaches:

1. Stochastic simulation of the bidding process at each event, such that b random numbers were drawn from each wealth distribution, $W(0)$ and $W(1)$, and ordered then assigned animals appropriately.
2. Stochastic simulation of the average outcome of the bidding process. Here, the auction process was simulated as above 50,000 times, and the average ordering of bids recorded. We then deterministically use this average behaviour to determine $\omega_0^{n,b}$, i.e. the number of individuals from Patch 0, given b buyers from each patch, to win the n^{th} highest bid or above.
3. Using a deterministic approximation of the ordering behaviour given b and the wealth distributions, as described below.

The results of the three methods for our model given Dynamic Filters 1 and 2 are shown in Fig 3.10. As the second approach - stochastic simulation of average auction behaviour only - demonstrated good accuracy while dramatically speeding up simulations and avoiding complications in assessing convergence to equilibrium, we followed this method in the main work.

Making a deterministic approximation of auction behaviour

To suggest possible deterministic approximations of bidding order behaviour when $b_0 = b_1 = b$, we first used stochastic simulations to visualise the ordering of bids from each patch (see Fig. 3.11, left). We focussed on wealth distributions $W(0) = N(\Delta W, 1)$ and $W(1) = N(0, 1)$, with $\Delta W \geq 0$, such that Patch 0 is at least as wealthy as Patch 1. Two limiting behaviours are implied. When $\Delta W = 0$ there is a relationship

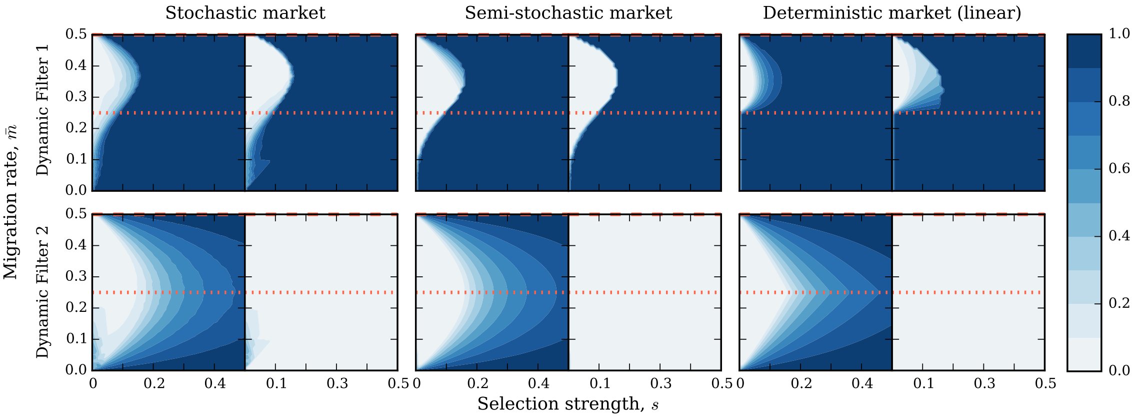


FIGURE 3.10: Comparing equilibrium behaviour given a market represented by Dynamic Filters 1 and 2 using different approaches to calculating migration parameters, $\pi^{\kappa}_{i,j}$, with $b = 50$ and $\delta W = 2$. The parameter space of selection coefficient, s, and migration rate, $\bar{m}$, was evenly sampled as a 40*40 grid in the ranges $[0, 0.5]$. In the left-most plots, stochastic auctions were simulated by explicitly drawing the wealth of b buyers from each distribution $W(0)$ and $W(1)$, placing these in wealth order, and then distributing available animals at the market according to buyer preference in wealth order. For each parameter combination, we performed ten replicates of 500 generations and averaged model results. The middle column of plots show a semi-stochastic approach in which average auction behaviour was determined using 50,000 simulated stochastic auctions, stored, and then used to deterministically determine auction results. The right-most plots show system equilibrium when a piecewise linear approximation of auction behaviour, see Eq. (3.17) and Fig. 3.11, was applied. We chose to use the semi-stochastic approach in the body of this work.

$$\omega_0^{n,b} \approx -0.5(2b - n) + b \tag{3.13}$$
$$= \frac{n}{2}. \tag{3.14}$$

This simply shows that each patch has an equal probability of being a buyer for any animal, and therefore that half of the bids over a given cut-off are from each patch. The second limiting behaviour occurs when the difference in wealth distributions between the two populations is so large that the top b bids are all made by buyers from the richer Patch 0. This is represented by a piecewise linear relationship, with

$$\omega_0^{n,b} \approx b \qquad \text{when } n > b, \tag{3.15}$$
$$\omega_0^{n,b} \approx n \qquad \text{when } n < b. \tag{3.16}$$

We found the equation

$$\omega_0^{n,b} \approx \alpha\frac{n}{2} + \beta \tag{3.17}$$

to interpolate between these regimes, with $\alpha = [1, 1+\delta W]$ and $\beta = [\delta W(b - \frac{n}{2}), 0]$ under conditions $[n > b, n < b]$. δW is a measure of the wealth difference between the two normal wealth distributions, and corresponds to the probability that a random buyer from the richer Patch 0 will make a higher bid than a buyer from the poorer Patch 1, minus the converse. We obtain this from the difference distribution of the two normal distributions $N(\mu(0), \theta(0))$ and $N(\mu(1), \theta(1))$, which gives $W(0) - W(1) = N(\mu(0) - \mu(1), \theta(0) + \theta(1))$, according to $\delta W = 1.0 - 2F_{W(0)-W(1)}(0)$ with $F_X(z)$ the cumulative distribution function at z for distribution X. We emphasise that this piecewise linear function is only an approximation of ordering behaviour, but that there is quite close correspondence between Eq. (3.17) and fully simulated stochastic auctions (Fig. 3.11).

We note that this model always given an approximation of $\omega_0^{n,b}$ that is less than or equal to the true value. Using non-linear fitting, as opposed to any analytic reasoning, we found the equation

$$\omega_0^{n,b} \approx \min(\frac{-\delta W}{2b} x^2 + (\delta W - 0.5) + b, b) \qquad \text{where } x = 2b - n \tag{3.18}$$

to provide a reasonable upper-bound to ordering behaviour. We do not explore the problem of bid order given buyers from two distributions here, but note that recent work has considered what is essentially a very similar problem and may offer better approximations than those suggested above [305].

3.3.7 Appendix 2: The Beverton-Holt model of population regulation

In the main text, we identified the role of population regulation as a critical factor in leading biased migration to alter breed distributions. In the model of Bolnick and Otto [267], regulation simply involves normalising total occupation in a patch to 1 (or K if population size, rather than the proportions of different breeds, is being tracked). If occupation is far from that equilibrium, this can imply very rapid population growth or very high mortality. We therefore implemented a model using the Beverton-Holt growth equation [306], Eq. (3.12). This is a discrete-time equation describing logistic growth, and is parametrised using carrying capacity K and a maximum growth rate, R_0, corresponding to the rate of population growth under ideal conditions. When $R_0 > 1$, the population will grow if $0 < n < K$ and shrink if $n > K$.

We show behaviour of our Static Filter models with $K = 100$ and $R_0 = 2.0$ (Fig. 3.12) or 1.25 (Fig. 3.13). The main effect of using the Beverton-Holt growth model on the phase space we explore is to compress the region that leads Breed B to fix in Patch 1. The manner in which this occurs is interesting, in that the greatest impact is on the (often) least relevant high-migration bias high-selection difference regime. We show the impact on mortality caused by biased migration and the relative population sizes of the two patches at equilibrium, this time with $R_0 = 0.5$, in Fig. 3.14. There is a pronounced

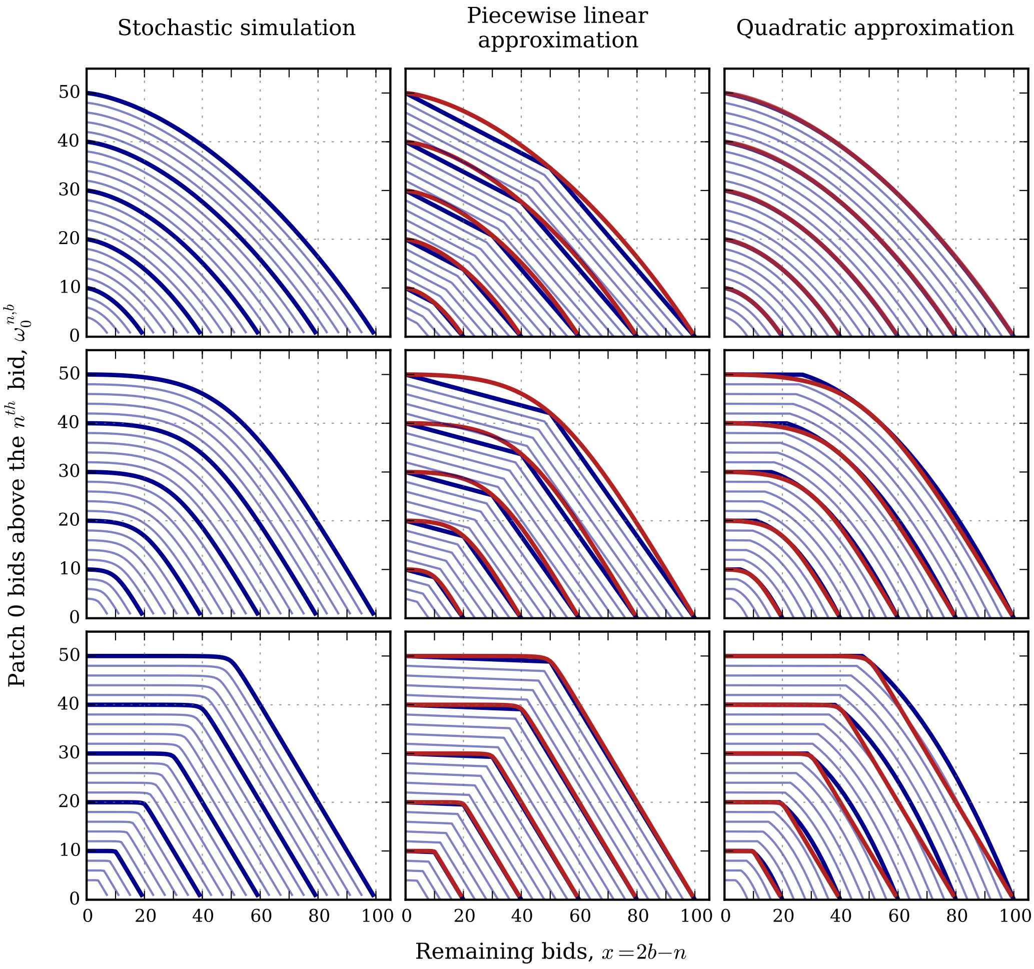


Figure 3.11: Average stochastic bid-order behaviour (50,000 replicates) as compared to simple piecewise linear and quadratic approximations using Eqs. (3.17, 3.18). Each line represents a number of buyers, $b = [2, 4, 6, ..., 48, 50]$, involved in the auction from each patch. Buyers from Patch 1 have wealth sampled from $W(1) = N(1, 0)$, while those from Patch 0 have wealth sampled from $W(0) = [N(1, 1), N(2, 1), N(4, 1)]$ in the top, middle, and bottom rows respectively. These correspond to weath differences $\delta W = 0.383, 0.683$ and 0.954. For each b, the x-axis describes the number of bids remaining after n bids, and ranges from 0 to $2b$. In the case of the approximations of auction behaviour, the average stochastic behaviour is indicated in red for representative values of b.

reduction in mortality caused by biased migration as compared to the system we explore in the main text, see Fig. 3.9. Intriguingly, strong migration biases ,which are intended to represent large wealth differences, either lead to a situation in which the population size of both patches is similar but the proportion of Breed A is different (top-left of the phase space, low s) or to a situation in which both patches are fixed for Breed A but have dramatic differences in herd size (top-right, moderate to high s).

We also investigated the behaviour of Dynamic Model 1 given the Beverton-Holt model of population regulation. We again considered a system with carrying capacity $K = 100$

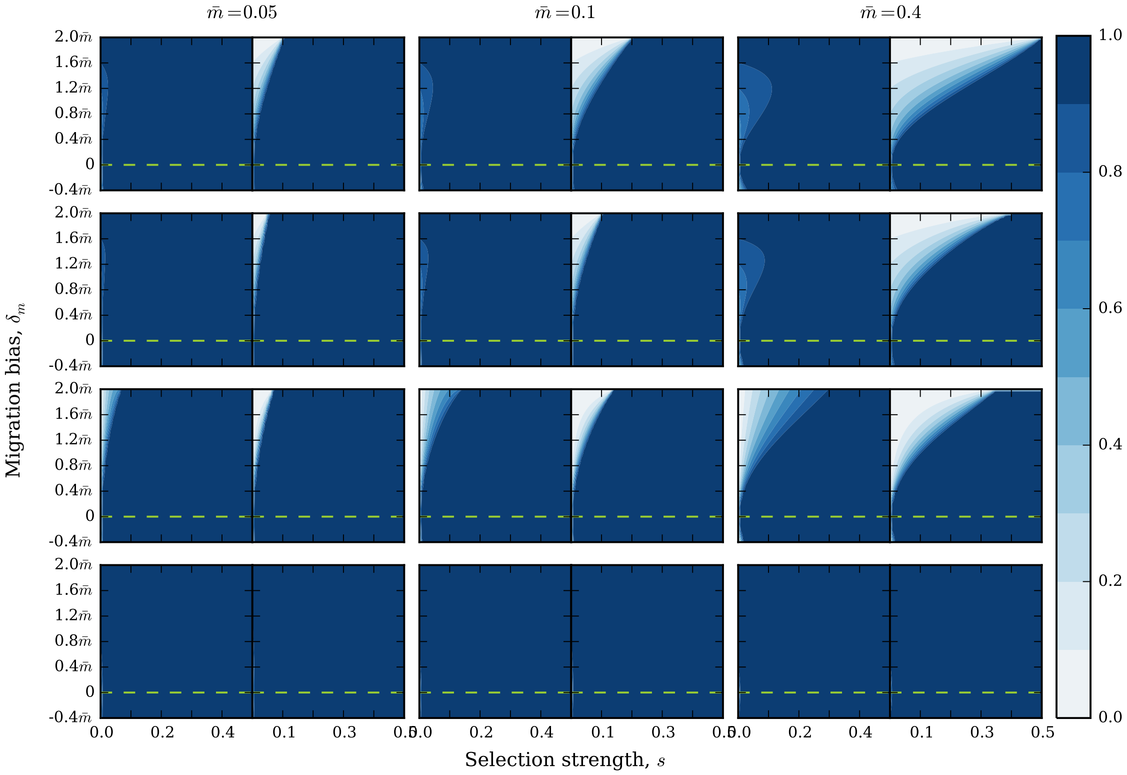


FIGURE 3.12: Equilibrium occupation behaviour given a static migration filter and selection against Breed B, using a Beverton-Holt growth function with maximum growth rate $R_0 = 2.0$. The rows correspond to four different types of biased migration filter, see caption of Figure 3.3 for details.

and either $R_0 = 2.0$, Fig. 3.15, or 1.25, Fig. 3.16. When population growth was high, the system equilibrium behaviour was extremely similar to that observed using our standard model of population regulation, see Fig. 3.6. The unusual behaviour apparent in the lower-right subplot, when $\delta W = 4$, $b = 50$ and $s \approx 0.3$, $\bar{m} \approx 0.4$, did not appear to be a numerical error, and arose after a rather complicated population size and composition trajectory, emphasising the complex interactions between a variable market supply (due to population size changes) of evolving composition coupled with constant demand. The population size in each patch when $R_0 = 2$ at equilibrium is shown in Fig. 3.16. The scale emphasises over-filling ($n > K$). As buyers from Patch 0 are better able to purchase animals, there is a relative increase in Patch 0 population size compared to Patch 1.

When we applied the Beverton-Holt growth function with $R_0 = 1.25$, the system equilibria changed over much of the parameter space, especially when the wealth difference was high. Under such circumstances, the impact of buyers from Patch 0 purchasing all or most animals arriving at the market from Patch 1 was sufficient to overpower local population growth in Patch 1, leading to very low Patch 1 population sizes. In such conditions, Patch 1 no longer contributed to the market, but could also import very

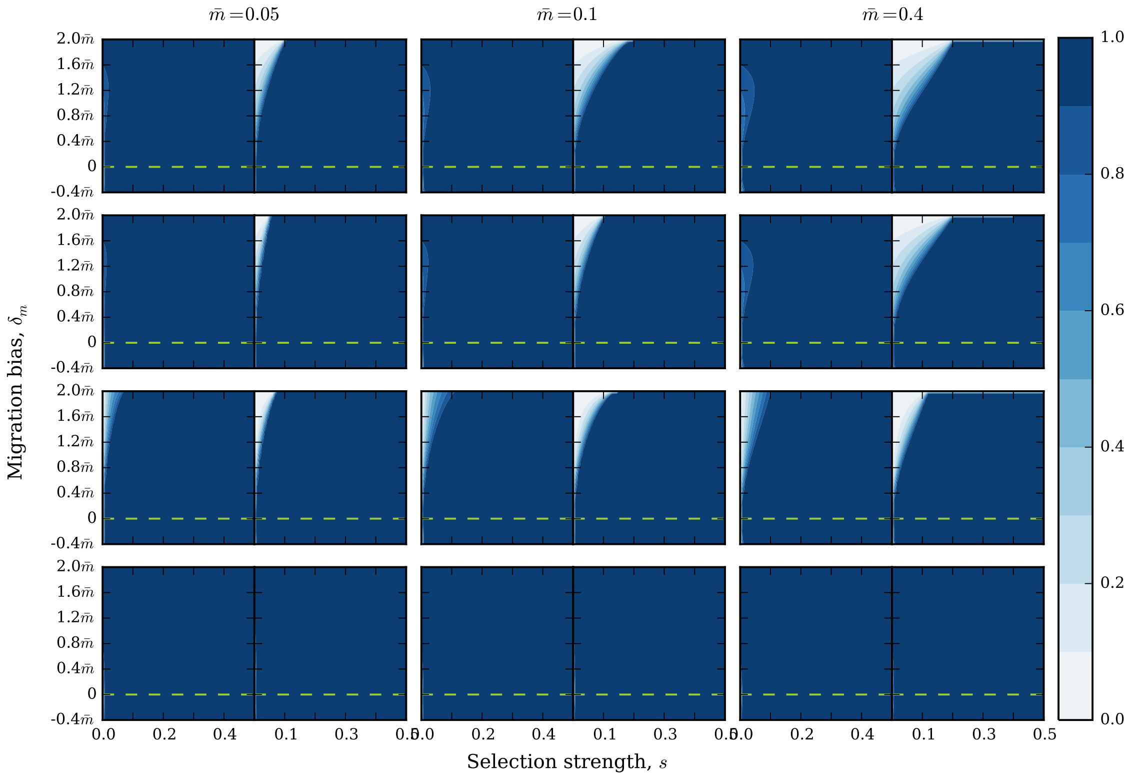


FIGURE 3.13: For details see 3.12. Now the Beverton-Holt growth function has a maximum growth rate of $R_0 = 1.25$.

few animals from Patch 0 due to wealth differences - essentially, Patch 0 approaches a situation where it evolves as an isolated population. Note that under such conditions, local extinction of animals in Patch 1 would be likely if demographic stochasticity was also modelled. Regions of the parameter space where the equilibrium population size of Patch 1 is less than 10 are shaded in grey in Figs. 3.17 and 3.18.

More broadly, the impact of lower growth on the equilibrium system state under a given selection and migration regime was similar to that observed for the static filter models, Figs. 3.12 and 3.13. Specifically, the regime corresponding to evolutionarily unexpected high Breed B frequency becomes compressed, such that a smaller degree of selection against Breed B (lower s) is able to overpower the effect of biased migration and yield a high Breed A frequency equilibrium.

Applying the Beverton-Holt growth equation to our Dynamic Filter 2 model resulted in broadly comparable effects - very limited impact on equilibrium patch occupation as compared to the standard model of population regulation when growth was fast ($R_0 = 2$, Figs. 3.19 and 3.20) and more pronounced impact when growth was slow ($R_0 = 1.25$, Figs. 3.21 and 3.22). As was the case with Dynamic Filter 1, for moderate wealth differences (eg. $\delta W \leq 2$) at least the essential characteristics of system equilibrium

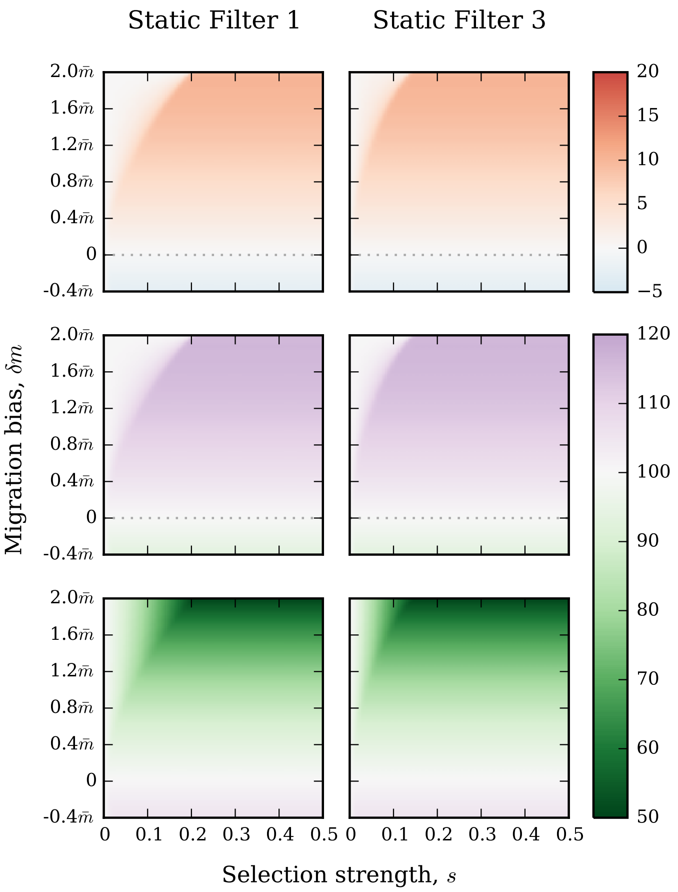


FIGURE 3.14: Mortality caused by migration (top row) according to a static filter with $\bar{m} = 0.1$ and $K = 100$, as well as population size in Patch 0 (second row) and Patch 1 (bottom row), when a Beverton-Holt growth function with maximum growth rate $R_0 = 1.5$ replaces the population regulation model used in the main text. The system is approximately at equilibrium, having run for 500 time steps from initial conditions $A_0(t) = A_1(t) = 0.5$.

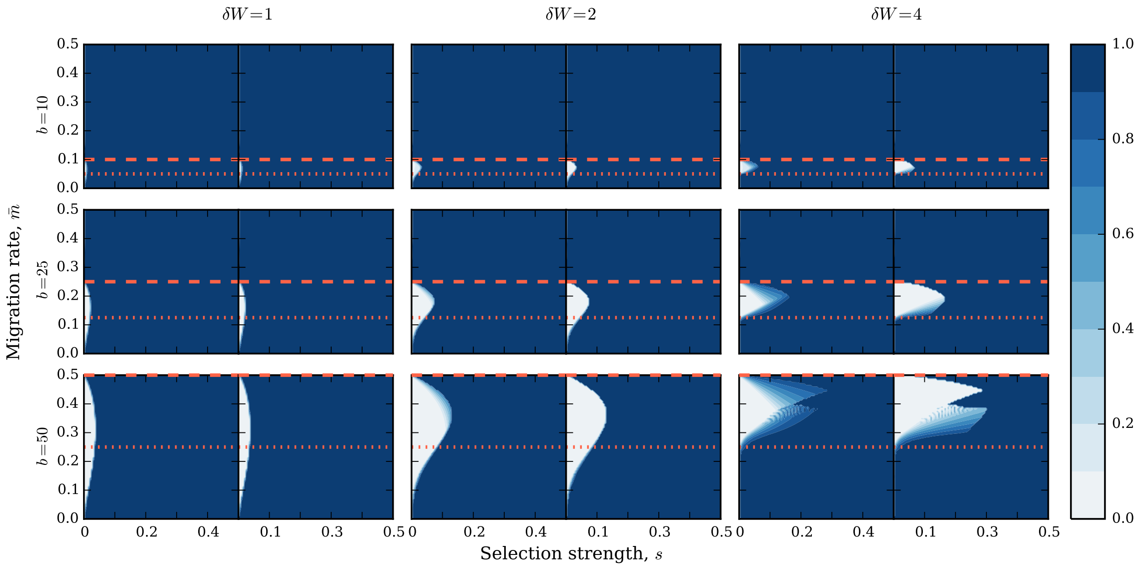


FIGURE 3.15: Equilibrium occupation behaviour given migration according Dynamic Filter 1 and selection against Breed B in both patches, with a Beverton-Holt growth function with maximum growth rate $R_0 = 2$ replacing the population regulation used in the main text. See caption of Fig. 3.6 for more details on the construction of the plots.

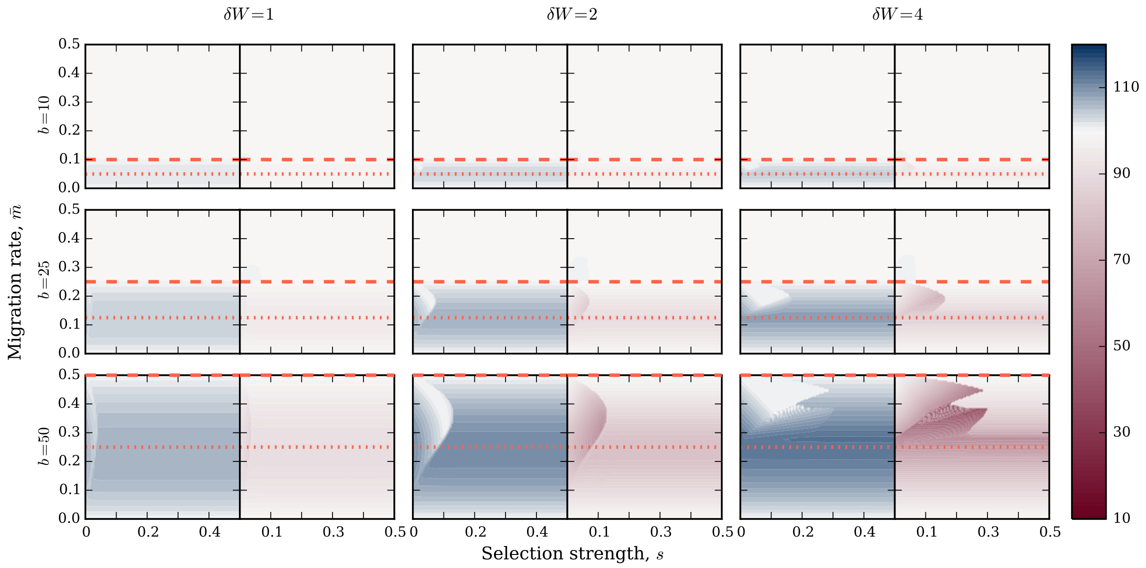


FIGURE 3.16: Equilibrium population size when migration is according to Dynamic Filter 1 and selection is against Breed B in both patches, with a Beverton-Holt growth function with maximum growth rate $R_0 = 2$ replacing the population regulation used in the main text. Note the asymmetric colour scale emphasising population over-filling, with white corresponding to $n = K = 100$. See caption of Fig. 3.6 for more details on the construction of the plots.

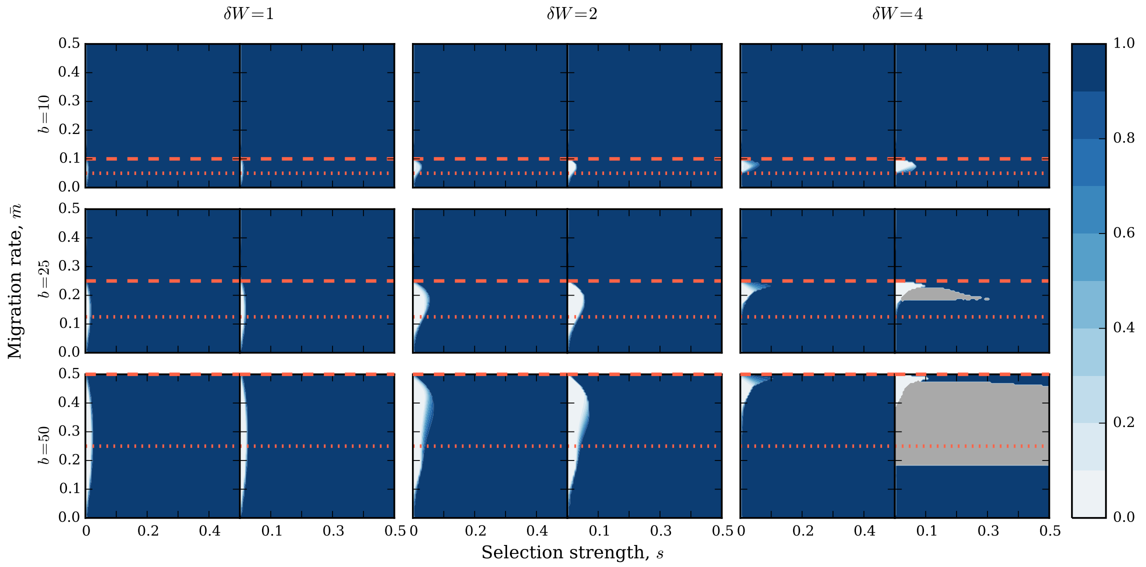


FIGURE 3.17: As caption of Fig. 3.15, but with the Beverton-Holt maximum growth rate of $R_0 = 1.25$.

behaviour when selection is low and demand just exceeds supply were unaffected by the new growth function.

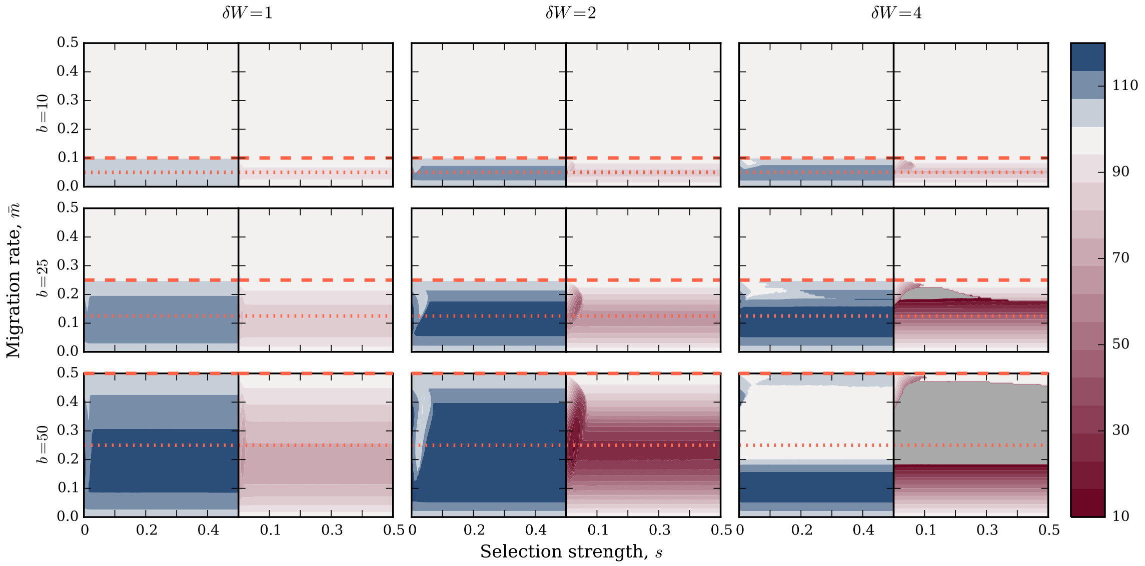


FIGURE 3.18: As caption of Fig. 3.16, but with the Beverton-Holt maximum growth rate of $R_0 = 1.25$.

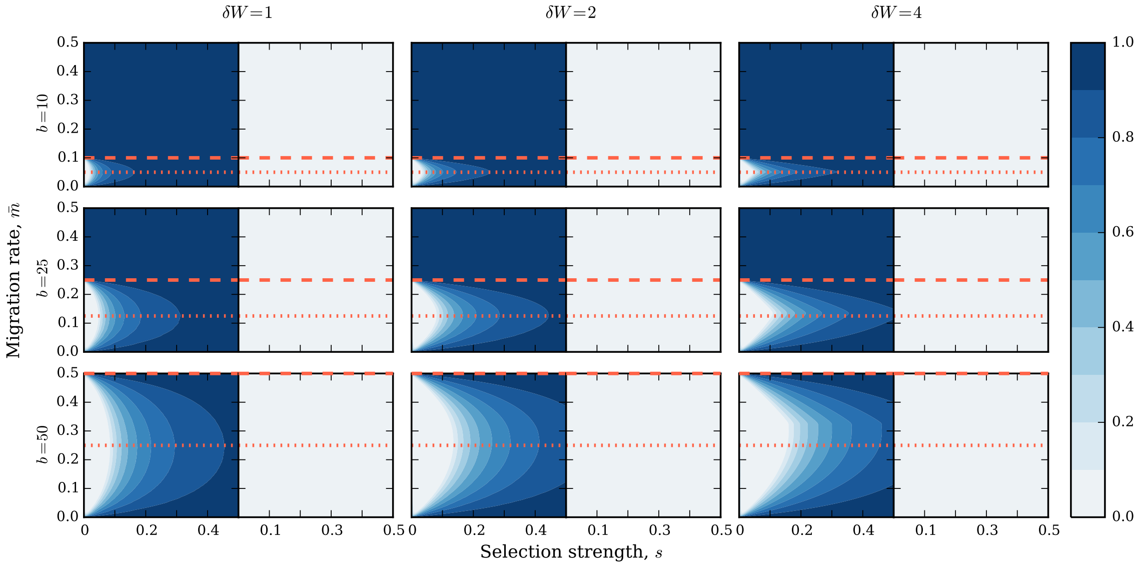


FIGURE 3.19: Equilibrium occupation behaviour given Dynamic Filter 2 migration and selection against Breed B in Patch 0 and Breed A in Patch 1, with a Beverton-Holt growth function with maximum growth rate $R_0 = 2.0$ replacing the population regulation used in the main text. Regions of the parameter space that lead to an equilibrium local population of less than 10 are shaded grey. See caption of Fig. 3.6 for more details on the construction of the plots.

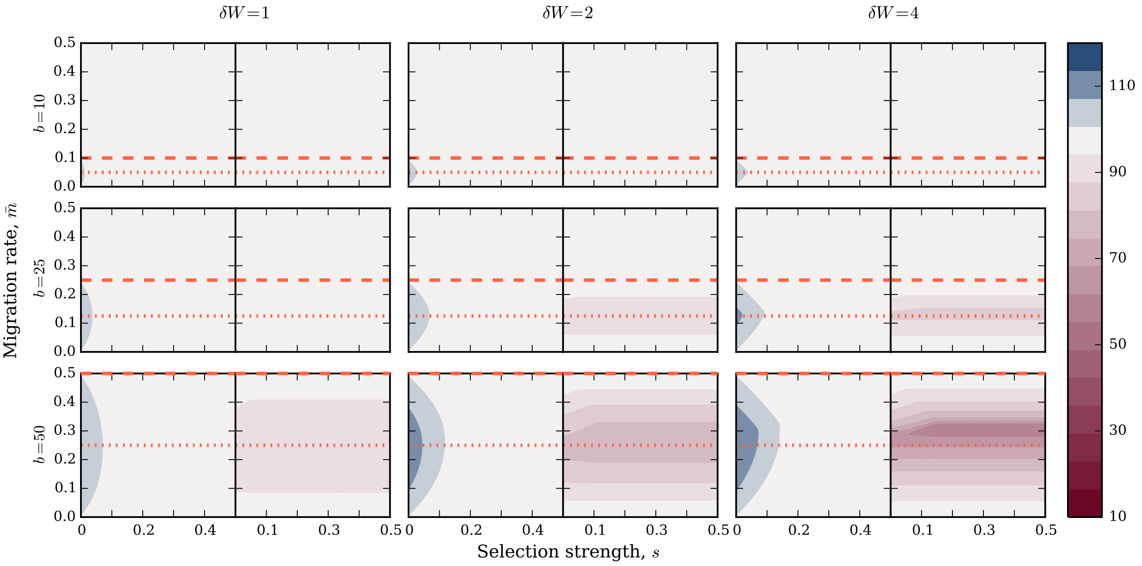


FIGURE 3.20: Equilibrium population size when migration is according to Dynamic Filter 2 and selection is against Breed B in Patch 0 and Breed A in Patch 1, with a Beverton-Holt growth function with maximum growth rate $R_0 = 2.0$ replacing the population regulation used in the main text. Note the asymmetric colour scale emphasising population over-filling, with white corresponding to $n = K = 100$ and regions of the parameter space that lead to local equilibrium population size under 10 shaded grey. See caption of Fig. 3.6 for more details on the construction of the plots.

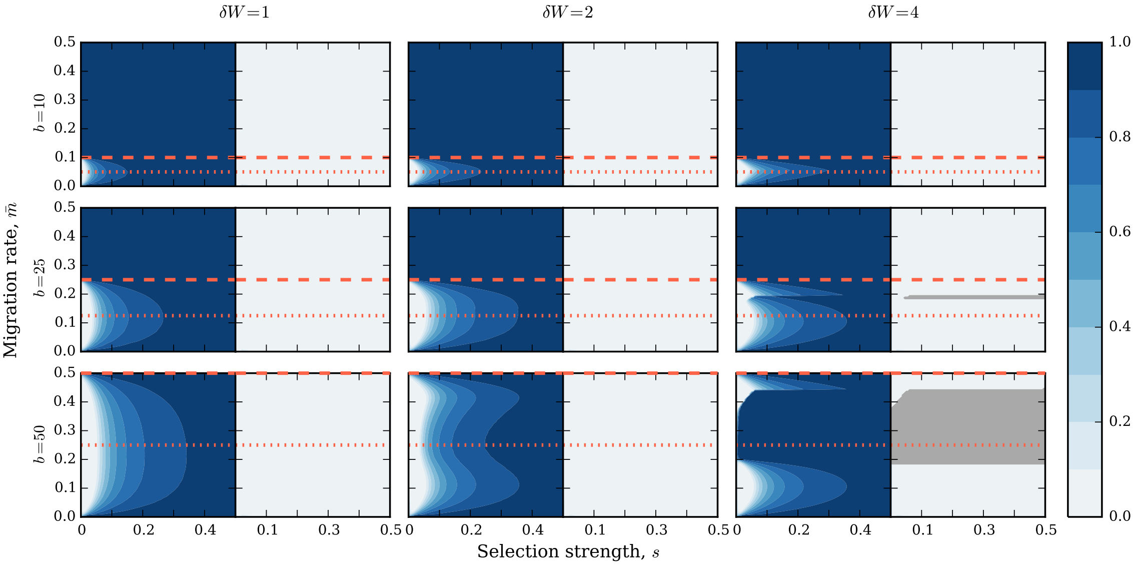


FIGURE 3.21: As caption of Fig. 3.19, but with the Beverton-Holt maximum growth rate of $R_0 = 1.25$

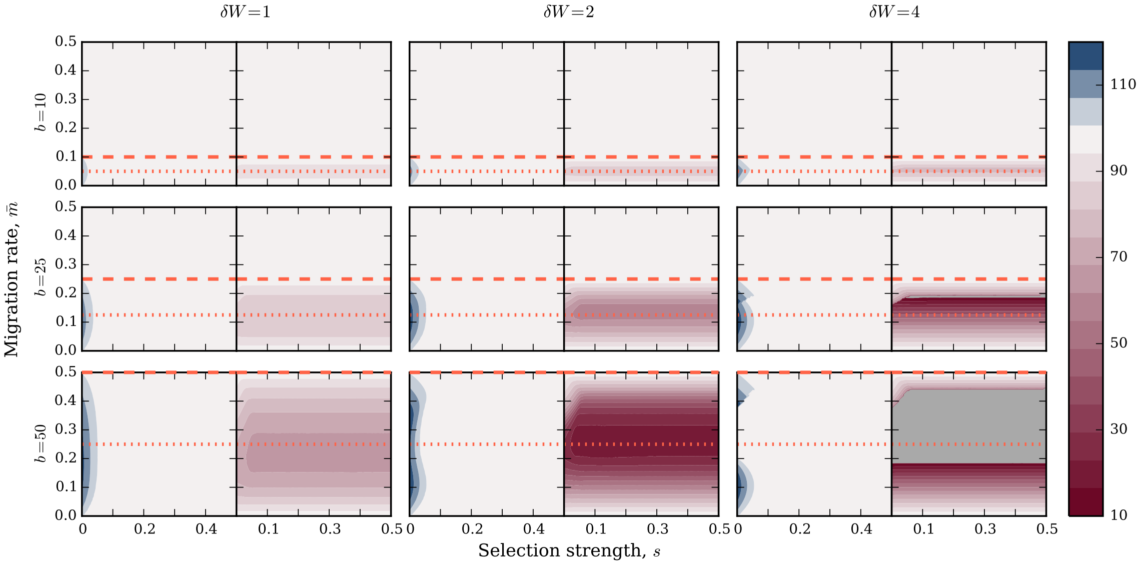


FIGURE 3.22: As caption of Fig. 3.20, but with the Beverton-Holt maximum growth rate of $R_0 = 1.25$

3.3.8 Appendix 3: A static market working with or against selection

For completeness, we have also characterised the impact of a our static migration filters when the migration biases either support or oppose the assortment of breeds to their more favourable habitats. The case of migration supporting the movement of breeds (species) to their selectively favoured environments was explored by Otto and Bolnick [267], and indeed our results are identical to theirs (Figure 3.23, first row).

The impact of placing the filter on one species only is of some interest, however, in that a single invading trait that impacts migration would usually be expected to do so in carriers only. We note that increasing migration bias will actually lead to a less-adapted equilibrium than unbiased migration, even when the bias supports movement of one species to it's selectively favourable patch. This is clearly apparent in these plots when migration is high, $\bar{m} = 0.4$, for Static Filter 3 (see left plot, corresponding to Patch 0) and Static Filter 4 (see right plot, corresponding to Patch 1). This result is interesting, suggesting that the evolution of migration biases based on environmental cues which are not density dependent (i.e. a particular habitat is preferred irrespective of local population filling) may be unlikely. Taking Static Filter 3 as an example, we recall that migration of Breed A from Patch 1 to Patch 0 is fast and Breed A movement in the opposite direction is slow, $\pi_{01}^{A} = m_L$ while $\pi_{10}^{A} = m_H$, and that Breed B moves at a constant rate in both directions $\pi_{01}^{B} = \pi_{10}^{B} = \bar{m}$. An example evolutionary trajectory of this model is shown in Fig 3.24. The evolutionary narrative is similar to that described in the main text, with the migration biases leading to rapid growth of Breed B in Patch 1, which is then able to flood Patch 0 with Breed B animals and counteract their negative selection.

For completeness, we also briefly characterised a model in which the migration filter opposes selection. This shows extremely rich behaviour, Fig. 3.25. Note that in this model we have reversed selection rather than the direction of the filter, such that Breed B is favoured in Patch 0 and Breed A in Patch 1. An expected pattern that is easily observed again involves the case of migration bias only effecting one breed. In such circumstances, that breed will tend to go extinct (as occurs, given moderately low s, for Breed A given Static Filter 3, and for Breed B given Static Filter 4). The behaviour of a simple symmetrical migration bias, Static Filter 1, is also intuitive, with the bias tending to oppose selection and reverse the expected equilibrium occupation.

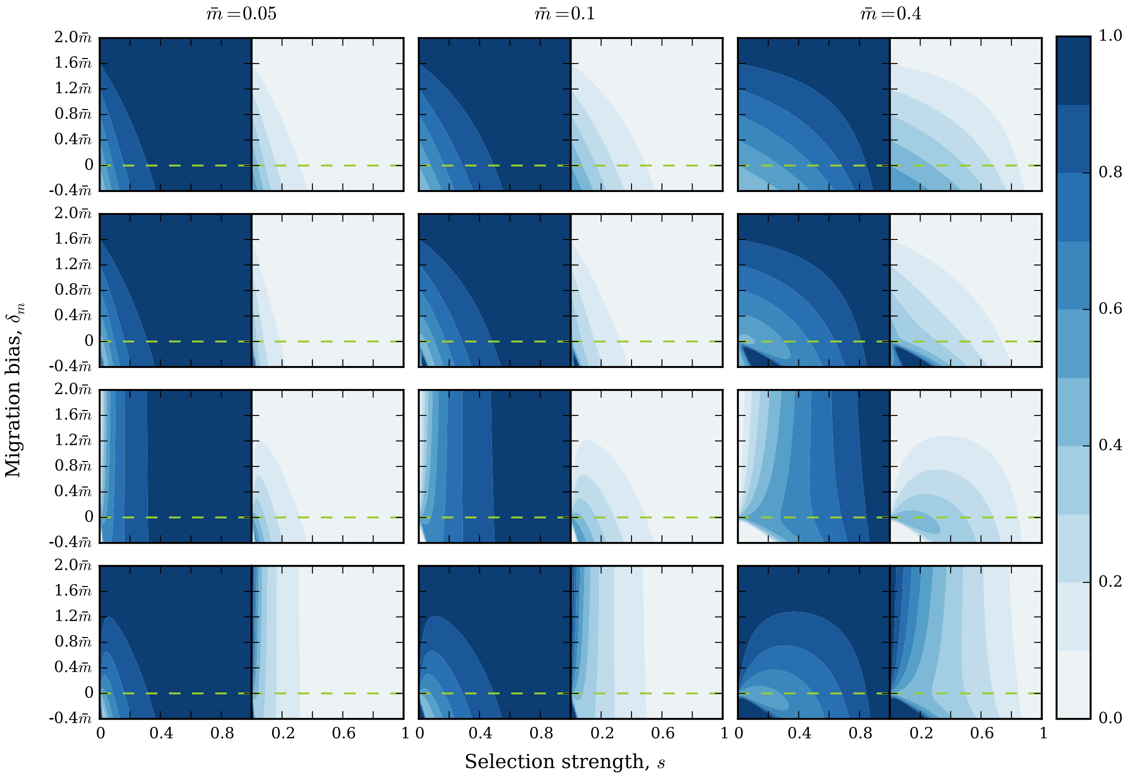


FIGURE 3.23: Equilibrium Breed A occupation given a static migration filter that promotes the movement of breeds to their selectively favoured habitats. Patch 0 occupation $A_0(t)$ and Patch 1 occupation $A_1(t)$ are on the left and right of each pair of plots. We explore the parameter space of selection strength, $s_0^B = s_1^A = 1 - s$, $s_0^A = s_1^B = 1$, with $0 < s \leq 0.5$, and strength of migration bias. Results from three core migration rates, $\bar{m} = 0.05, 0.1, 0.4$, and four different migration filter models are shown. From top to bottom, rows show Static Filter 1, Static Filter 2, Static Filter 3 and Static Filter 4, as described by Eqs. (3.4-3.8) The parameter space was evenly sampled as a 40 by 40 grid, with the model allowed to run for 500 generations with each set of parameter values.

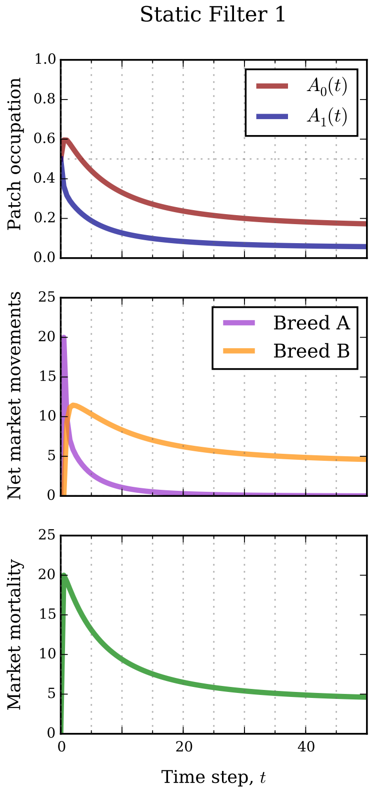


FIGURE 3.24: The trajectory of system evolution given selection favouring Breed A in Patch 0 and Breed B in Patch 1, $s_0^B = s_1^A = 0.95$, $s_0^A = s_1^B = 1$, with Static Filter 3 promoting the movement of animals to their selectively favoured patches with $\bar{m} = 0.4$ and $\delta_m = \bar{m}$. Under this regime, Patch 0 surprisingly reaches an equilbrium with low Breed A occupation. From top to bottom, the plots show relative patch occupation, net movement to Patch 0 due to migration, and the market-induced mortality (here, excess mortality in Patch 0 and excess growth in Patch 1).

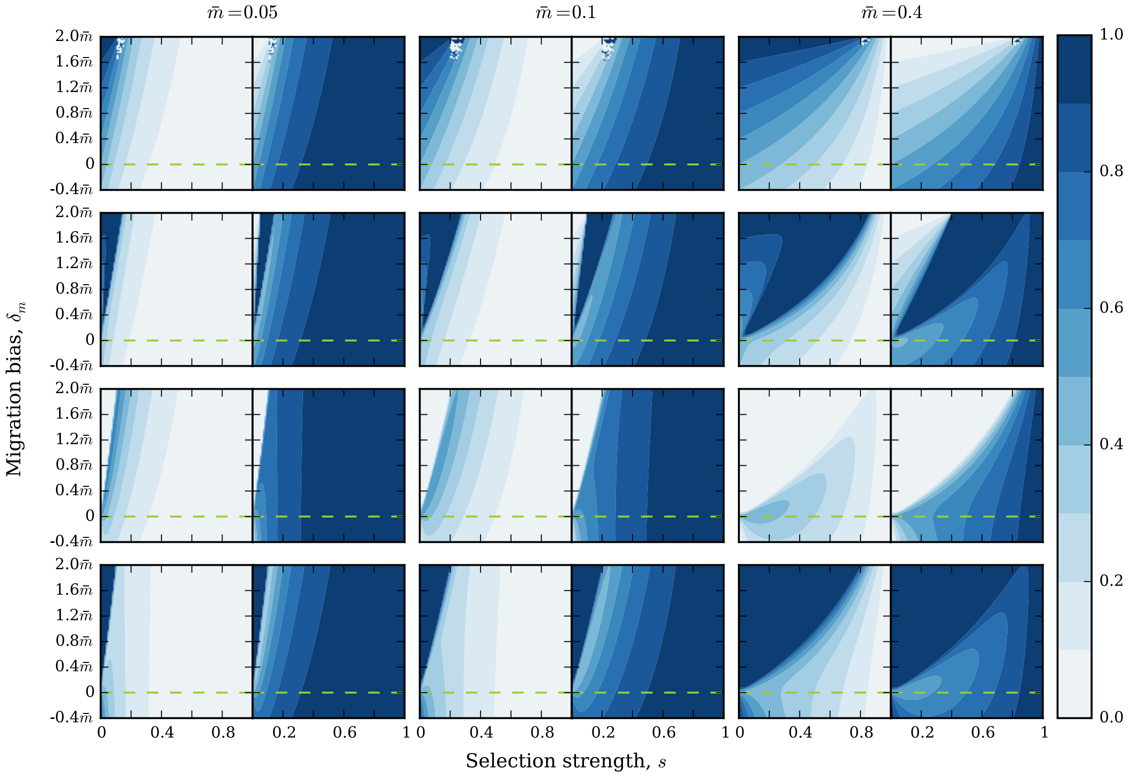


FIGURE 3.25: Equilibrium Breed A occupation given a static migration filter that acts against the movement of breeds to their selectively favoured habitats. The selection strength, $s_0^A = s_1^B = 1 - s$, $s_0^B = s_1^A = 1$, with $0 < s \leq 0.5$. Other details are as Fig. 3.23.

3.3.9 Appendix 4: Robustness to initial conditions

To explore sensitivity to initial conditions, we re-ran our numerical simulations with the starting population either as $[A_0(0), A_1(0)] = [0.99, 0.99]$, $[A_0(0), A_1(0)] = [0.01, 0.01]$, $[A_0(0), A_1(0)] = [0.01, 0.99]$ or $[A_0(0), A_1(0)] = [0.99, 0.01]$. Our results for the static market model (Figs. 3.3 and 3.4) and the dynamic model with opposing selection regimes (Fig. 3.7) were virtually identical to those shown in the main text, and we do not re-draw them here. However, we did observed differences in the behaviour of the Dynamic Filter 1 model. Results for high ($[A_0(0), A_1(0)] = [0.99, 0.99]$) and low ($[A_0(0), A_1(0)] = [0.01, 0.01]$) initial Breed A occupation are shown in Figs. 3.26 and 3.27 respectively, while those for imbalanced initial occupation regimes are illustrated by Fig. 3.28 ($[A_0(0), A_1(0)] = [0.01, 0.99]$) and Fig. 3.29 ($[A_0(0), A_1(0)] = [0.99, 0.01]$).

The parameter space corresponding to fixation of Breed B shrinks when Breed A starts at high frequency, Fig. 3.27. Nevertheless, different initial conditions do not entirely disrupt the anomalous equilibrium behaviour, with globally low-fitness Breed B fixing, of this model. In particular, there is still an high-frequency of low-fitness Breed B at equilibrium when selection is low and demand just exceeds supply, which is the regime that is most plausible economically and, depending on whether the breeds are taken to represent the same or different species, biologically.

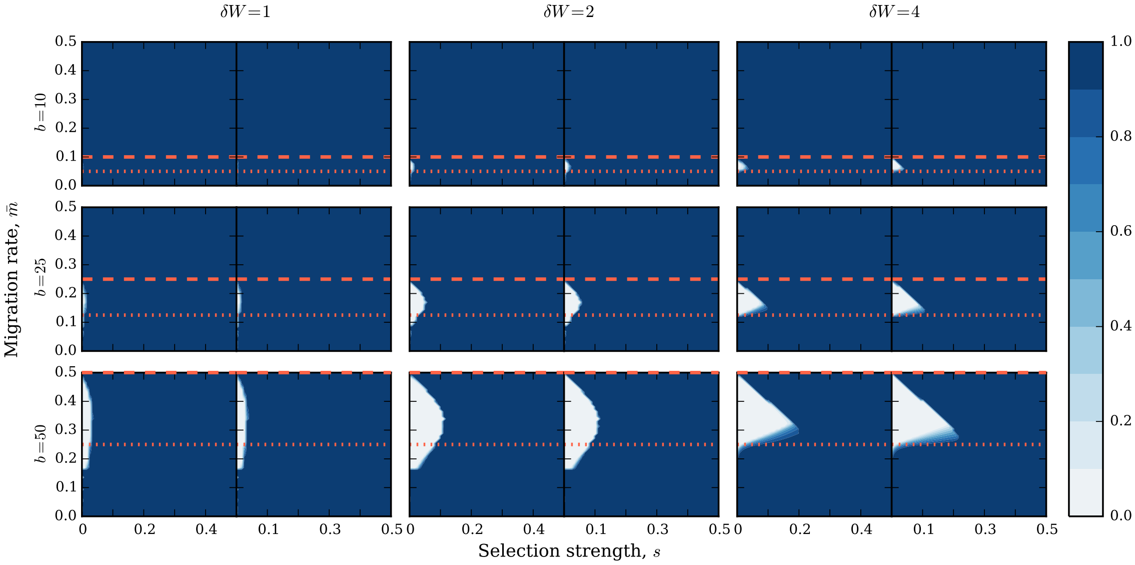


FIGURE 3.26: Equilibrium patch occupation given Dynamic Filter 1 and selection against Breed B, with initial conditions $[A_0(0), A_1(0)] = [0.99, 0.99]$. See the caption of Fig. 3.6 for further details, and to compare system equilibrium behaviour.

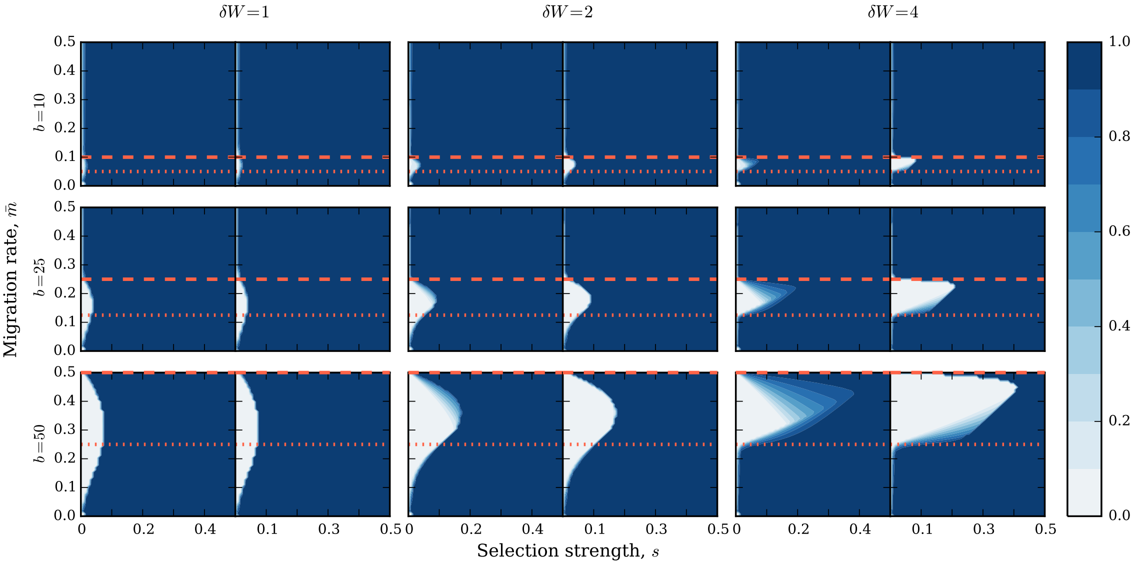


FIGURE 3.27: Equilibrium patch occupation given Dynamic Filter 1 and selection against Breed B, with initial conditions $[A_0(0), A_1(0)] = [0.01, 0.01]$. See the caption of Fig. 3.6 for further details, and to compare system equilibrium behaviour.

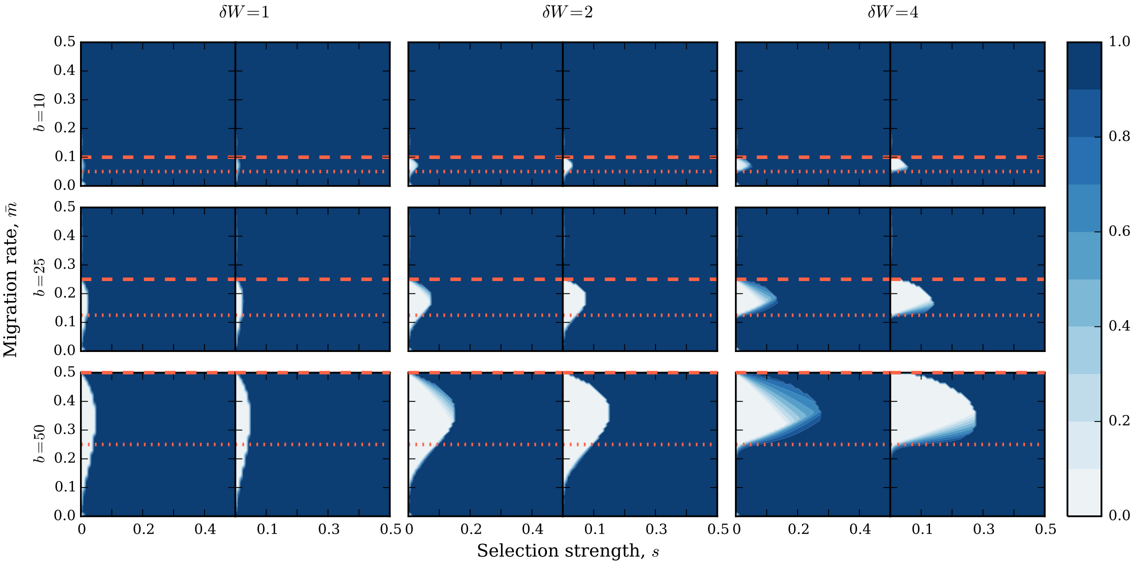


FIGURE 3.28: Equilibrium patch occupation given Dynamic Filter 1 and selection against Breed B, with initial conditions $[A_0(0), A_1(0)] = [0.01, 0.99]$. See the caption of Fig. 3.6 for further details, and to compare system equilibrium behaviour.

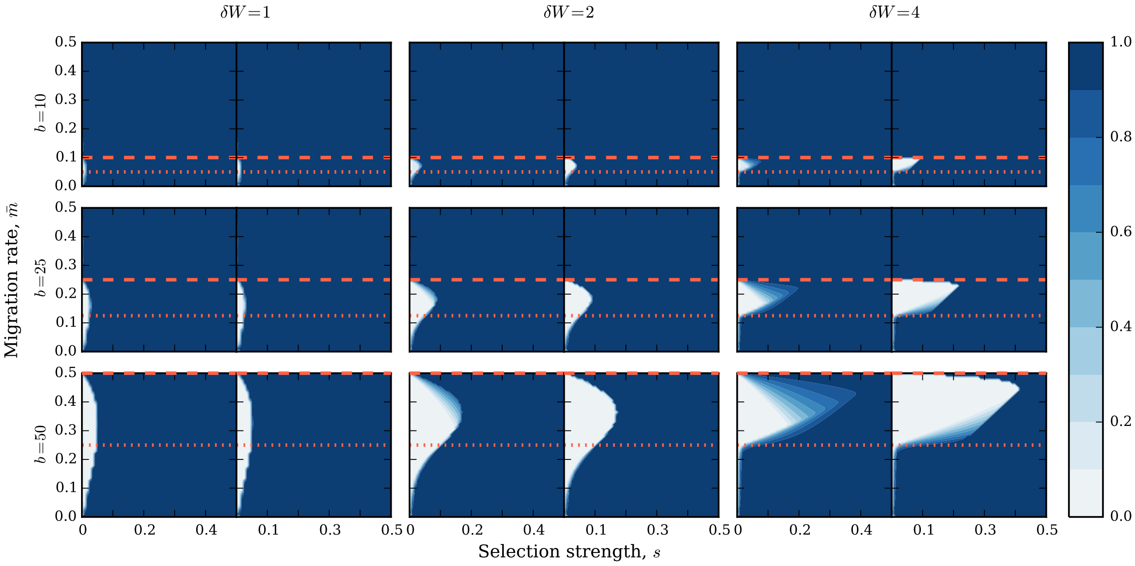


FIGURE 3.29: Equilibrium patch occupation given Dynamic Filter 1 and selection against Breed B, with initial conditions $[A_0(0), A_1(0)] = [0.99, 0.01]$. See the caption of Fig. 3.6 for further details, and to compare system equilibrium behaviour.

3.4 Some mathematical comments on the manuscript

In the manuscript presented above, we use simulations to explore a representation of domestic animals flowing through a simple market system. We did not make any attempt to mathematically analyse the model [7], and indeed I do not do so now. Rather, I note that it is posisble to emphasise important properties of the model through simple algebraic manipulation, and proceed by presenting some basic findings. I do not suggest that these mathematical points are novel - the system we explore is very simple, as are the questions I am attempting to clarify - but do suggest that they may help to elucidate the behaviour of the system or guide further analysis.

For clarity, I begin by briefly re-stating our model system, quoting almost directly form the manuscript above. The model consists of two habitat patches, Patch 0 and Patch 1, with carrying capacities K_0 and K_1. These are occupied by two animal breeds, Breed A and Breed B. The system is updated according to a determinsitic update rule consisting of three recurrence equations, which, given the proportion of Breed κ in Patch i at time t, $0 \leq \kappa_i(t) \leq 1$, calculates the value of $\kappa_i(t+1)$. The three recurrence equations are applied, in order, to calculate the proportion of Breed κ in Patch i after migration between patches, $\kappa_i^m(t)$, selection in the patch, $\kappa_i^s(t)$, and then population regulation, $\kappa_i^r(t)$, with the occupation of Patch i at $t+1$ equal to its occupation after the regulation step of time t, $\kappa_i(t+1) = \kappa_i^r(t)$. The first recurrence equation, corresponding to migration, is,

$$A_i^m(t) = \frac{\sum_{j\in 0,1} K_j \pi_{ji}^A (1-c_{ji}) A_j(t)}{\sum_{\kappa\in A,B} \sum_{j\in 0,1} K_j \pi_{ji}^\kappa (1-c_{ji}) \kappa_j(t)}, \qquad \text{(3.1 revisited)}$$

where π_{ji}^κ is the migration rate of Breed κ from Patch j to Patch i and c_{ji} is the mortality associated with any migration from Patch j to Patch i. Selection then occurs,

$$A_i^s(t) = \frac{s_i^A A_i^m(t)}{\sum_{\kappa\in A,B} s_i^\kappa \kappa_i^m(t)}, \qquad \text{(3.2 revisited)}$$

where s_i^κ represents the fitness of Breed κ in Patch i. Finally, population regulation is applied, which is the last update step,

$$A_i(t+1) = A_i^r(t) = \frac{K_i A_i^s(t)}{K_i \sum_{\kappa\in A,B} \kappa_i^s(t)}. \qquad \text{(3.3 revisited)}$$

As stated in the manuscript, population regulation simply normalises the current occupation of the site to K, which allows us to follow the proportion of the different breeds rather than their population sizes but can imply very high population growth or mortality in a generation.

[7] It may be that my more mathematically able co-authors attempt to characterise the system more fully in the manuscript we ultimately submit, or in other work.

I will focus on the combined effect of migration and population regulation, as I have argued that the interaction between these processes is critical in producing the 'counter-evolutionary' effects observed in our modelling. I will first assume that selection does not change the total population size in either patch. Under such conditions, the update rule will still correctly describe the system I simulated if the equations are re-ordered such that population regulation immediately follows the migration step. The combined effect of migration and population regulation on the total proportion of Breed A in the entire system (both patches) is

$$A_{0,1}^{m,r}(t) = \frac{A_0^m(t)K_0 + A_1^m(t)K_1}{K_0 + K_1}. \tag{3.19}$$

Assuming that the carrying capacity is the same in both patches, $K_0 = K_1 = K$, and that there is no mortality associated with migration, $c_{ij} = 0$, this can be written out in full as Eq. 3.2 revisited

$$\begin{aligned}
A_{0,1}^{m,r}(t) = &\frac{A_0(t)\pi_{00}^A + A_1(t)\pi_{10}^A}{2\big(A_0(t)\pi_{00}^A + (1 - A_0(t))\pi_{00}^B + A_1(t)\pi_{10}^A + (1 - A_1(t))\pi_{10}^B\big)} + \\
&\frac{A_1(t)\pi_{11}^A + A_0(t)\pi_{01}^A}{2\big(A_1(t)\pi_{11}^A + (1 - A_1(t))\pi_{11}^B + A_0(t)\pi_{01}^A + (1 - A_0(t))\pi_{01}^B\big)} \\
= &\frac{A_0(t)\pi_{00}^A + A_1(t)\pi_{10}^A}{2\big(A_0(t)(\pi_{00}^A - \pi_{00}^B) + A_1(t)(\pi_{10}^A - \pi_{10}^B) + \pi_{00}^B + \pi_{10}^B\big)} + \\
&\frac{A_1(t)\pi_{11}^A + A_0(t)\pi_{01}^A}{2\big(A_1(t)(\pi_{11}^A - \pi_{11}^B) + A_0(t)(\pi_{01}^A - \pi_{01}^B) + \pi_{11}^B + \pi_{01}^B\big)},
\end{aligned}$$

such that, noting $\pi_{i,i}^{\kappa} = 1 - \pi_{ij}^{\kappa}$,

$$\begin{aligned}
A_{0,1}^{m,r}(t) = &\frac{A_0(t) - A_0(t)\pi_{01}^A + A_1(t)\pi_{10}^A}{2\big(A_0(t)(\pi_{01}^B - \pi_{01}^A) + A_1(t)(\pi_{10}^A - \pi_{10}^B) + 1 - \pi_{01}^B + \pi_{10}^B\big)} + \\
&\frac{A_1(t) - A_1(t)\pi_{10}^A + A_0(t)\pi_{01}^A}{2\big(A_1(t)(\pi_{10}^B - \pi_{10}^A) + A_0(t)(\pi_{01}^A - \pi_{01}^B) + 1 - \pi_{10}^B + \pi_{01}^B\big)}.
\end{aligned} \tag{3.20}$$

Collecting terms as $S = \pi_{10}^A - \pi_{10}^B$, $T = \pi_{01}^A - \pi_{01}^B$ and $U = \pi_{01}^B - \pi_{10}^B$, we have

$$\begin{aligned}
A_{0,1}^{m,r}(t) = &\frac{A_0(t) - A_0(t)\pi_{01}^A + A_1(t)\pi_{10}^A}{2\big(A_0(t)(-T) + A_1(t)(S) + 1 - U\big)} + \\
&\frac{A_1(t) - A_1(t)\pi_{10}^A + A_0(t)\pi_{01}^A}{2\big(A_1(t)(-S) + A_0(t)(T) + 1 + U\big)}.
\end{aligned}$$

This corresponds to

$$A_{0,1}^{m,r}(t)=\frac{\big(A_0(t)T-A_1(t)S+U\big)\big(A_0(t)-2A_0(t)\pi_{01}^{A}-A_1(t)+2A_1(t)\pi_{10}^{A}\big)+A_0(t)+A_1(t)}{-2\big(A_0(t)T-A_1(t)S+U-1\big)\big(A_0(t)T-A_1(t)S+U+1\big)}. \tag{3.21}$$

This equation can be used to determine the combined effects of migration and regulation on global frequency of Breed A,

$$\Delta A_{0,1}^{m,r}(t)=A_{0,1}^{m,r}(t)-A_{0,1}(t), \tag{3.22}$$

where $\Delta A_{0,1}^{m,r}(t)$ is the change in the global proportion of Breed A animals due to the migration and regulation recurrence equations.

The change in global Breed A proportion due to selection is

$$\Delta A_{0,1}^{s}(t)=\frac{A_0^{m,r}(t)}{2\big(A_0^{m,r}(t)+s_0^B(1-A_0^{m,r}(t))\big)}+\frac{A_1^{m,r}(t)}{2\big(A_1^{m,r}(t)+s_1^B(1-A_1^{m,r}(t))\big)}-A_{0,1}^{m,r}(t) \tag{3.23}$$

and could be compared with Eq. (3.22) to assess the relative impact of migration and regulation or selection at a given system state, although I do not pursue this here.

Another quantity of interest is the market-induced mortality. This is simply the imbalance in population filling caused by a migration step, and can be calculated as

$$\begin{aligned} n_0^m(t)-n_0(t)&=K\Big(A_0(t)\pi_{00}^{A}+A_1(t)\pi_{10}^{A}+(1-A_0(t))\pi_{00}^{B}+(1-A_1(t))\pi_{10}^{B}\Big)-K \\ &=K(A_0(t)T-A_1(t)S+U), \end{aligned} \tag{3.24}$$

where $n_0(t)$ represents the total number of animals in Patch 0 at time t and $n_0^m(t)$ being the total number of animals in Patch 0 after the migration step. The corresponding excess mortality or growth in Patch 1, ignoring selection, is the negative of Eq. (3.24) to conserve total animal numbers.

I will briefly make a few observations regarding Eq. (3.21), assuming that there is no selection; under these circumstances, it is the only recurrence equation needed to update the system. Firstly, in a situation where migration is unbiased such that all migration parameters are equal, $\pi_{ij}^{\kappa}=\pi^{A,B}$, the terms $U=S=T=0$, and we simply have

$$A_{0,1}^{m}(t)=\frac{A_0(t)+A_1(t)}{2}. \tag{3.25}$$

Whatever the pattern of Breed A occupation is, the global frequency cannot change due to migration under these circumstances. Secondly, when $\pi_{ij}^A=\pi_{ji}^A=\pi^A$ and $\pi_{ij}^B=\pi_{ji}^B=\pi^B$, such that migration is biased by breed but not directionally biased, $U=0$ and the difference in the migration rates of the two breeds is $S=T=\pi_A-\pi_B=V$. We define the difference in Breed A occupation $A_0(t) - A_1(t) = \delta_A(t)$, giving

$$A^m_{0,1}(t) = \frac{-V(2\pi_A - 1)\big(\delta_A(t)\big)^2 + A_0(t) + A_1(t)}{-2V^2\big(\delta_A(t)\big) + 2}. \tag{3.26}$$

The change in occupation is

$$\begin{aligned} A^m_{0,1}(t) - A_{0,1}(t) &= \frac{-V(2\pi_A - 1)\delta_A(t)^2 + A_0(t) + A_1(t)}{-2V^2\delta_A(t) + 2} - \frac{A_0(t) + A_1(t)}{2} \\ &= \frac{\delta_A(t)^2 V(1 - 2\pi^A) + \delta_A(t)^2 V^2\big(A_0(t) + A_1(t)\big)}{2 - 2\delta_A(t)^2 V^2} \end{aligned} \tag{3.27}$$

The denominator is positive, or zero when $V = 1$ and $\delta A(t) = 1$. Assuming that $V \neq 1$, we can obtain the conditions under which the frequency of Breed A increases in a given generation by determining when the numerator is positive,

$$\delta_A(t)^2 V(-2\pi^A + V(A_0(t) + A_1(t)) + 1) > 0 \tag{3.28}$$

and, consequently, the relevant conditions depend on the sign of V,

$$1 + V\big(A_0(t) + A_1(t)\big) > 2\pi^A \qquad \text{when } V > 0 \text{ and } \delta_A \neq 0 \tag{3.29}$$

$$1 + V\big(A_0(t) + A_1(t)\big) < 2\pi^A \qquad \text{when } V < 0 \text{ and } \delta_A \neq 0. \tag{3.30}$$

As migration occurs between the two patches, however, we expect them to become more alike under neutrality, such that δ_A with approach 0. Nevertheless, in numerical simulations I have observed that the global change in frequency can be substantial before the system equilibriates. An extreme example occurs with initial conditions $[A_0(t), A_1(t)] = [1.0, 0.0]$, with parameters $\pi_B = 0$ and $0 < \pi_A < 1$. The inability of the Breed B animals in Patch 1 to invade Patch 0 means that they are subject to constant invasion pressure and, ultimately, go extinct. Perhaps a more useful example for judging the practical effect of this behaviour is offered by initial conditions $[A_0(t), A_1(t)] = [1.0, 0.0]$, with parameters $\pi_B = 0.05$ and $\pi_A = 0.1$. This leads to an equilibrium state of $A_0(t) = A_1(t) \approx 0.579$, such that, under neutrality, there has been an increase in the population of fast-migrating Breed A of nearly 16%.

Whether these sort of behaviours should be considered artefacts of the system or relevant to real ecological systems - for example, immediately after the removal of a barrier to migration - probably depends on the relative rate of migration and population growth/-mortality, as population regulation is exceptionally rapid in this system. I draw attention to this phenomenon largely because it appears of interest to our model behaviour, and do not suggest that it offers valuable novel understanding of the target system.

Several supplementary observations I have made during numerical simulations may warrant further attention. Firstly, a breed with migration an arbitrary bias, such that

$\pi^A_{01} = \pi^A_{10} = \pi^B_{01} \neq \pi^B_{10}$, will tend to become extinct; the same is true for a symmetrical bias of the form $\pi^A_{01} = \pi^A_{10} = \pi^A$, $\pi^B_{01} = \pi^A + \delta^b_m$, $\pi^B_{10} = \pi^A - \delta^b_m$. A similar pattern was observed in the wide-ranging simulations of McPeek and Holt [269] in 1992; although Ngoc *et al* do not emphasise this pattern in their mathematical analysis of a Lotka-Volterra system resembling our Static Filter models [266], it appears to occur there also. More broadly, it is clear that arbitrary migration biases lead to arbitrary clustering of similar individuals, which will increase competition with conspecifics, raising both an individual's mortality and that of its species type. Work on the IFD (Ideal Free Distribution [255], introduced above), which argues that animal clustering should follow the distribution of resources, is also relevant, in that species which deviate from the IFD are expected to have a below-optimal population growth rate.

A more interesting pattern is apparent in Appendix 3 of the above manuscript. There, we drew attention to the system behaviour implied by the Static Filter 3 migration model ($\pi^A_{01} = \bar{m} + \frac{\delta_m}{2}$, $\pi^A_{10} = \bar{m} - \frac{\delta_m}{2}$, $\pi^B_{01} = \pi^B_{10} = \bar{m}$) acting to promote the movement of animals to their selectively favoured patch, with $s^A_0 = s^B_1 = 1$ and $s^A_1 = s^B_0 = 1 - s$. Even this model can lead to breed distributions that are, surprisingly, non-optimal. This contradicts the findings of Bolnick and Otto [267], who used a model that corresponds to our Static Filter 1 to investigate the potential of migration biases to promote sorting of animals to their selectively favoured habitats. The model of Bolnick and Otto only considers a situation in which both animal types have migration biases, and as a novel genetically coded migration bias is likely to impact carriers only our model may be more realistic.

3.5 Representing the market

The simulation model presented in Chapter 2 is built upon a mature literature, while Chapter 4 follows a standard protocol to test a selection statistic using the widely used coalescent simulation framework. As there is much less modelling work on the impact of animal trade on genetic variation, there is an added responsibility to consider the accuracy of our model as a representation.

3.5.1 The structure of our model

Our model can be deconstructed into several components - the different animal types, the different locations, and the representation of population regulation, selection and market-induced migration biases. The use of two homogeneous animal types, two different patches and constant selection over time are idealisations designed to simplify model interpretation and generate initial results on the possible effects of migration biases. I do not discuss these in detail, but note that relaxing any of these constraints would be a useful extension to the model. Having already discussed the importance of population regulation, and explored the impact of using Beverton-Holt population growth, in the manuscript, I now focus on our representation of the market.

In our market models, we either take a top-down approach by proposing that wealth allows individuals to obtain animals they desire more often, and to sell those animal less often, giving our Static Market filters, or a bottom-up approach, describing a microscopic model of auction behaviour. This provides our Dynamic Market filters, and is a powerful approach, in that some aspects of the model are quite general - for example, our description of the auction algorithm through the ordering of choice according to wealth (an idea originally suggested by Gereon Kaiping, a co-author and fellow PhD student) is an elegant approach to representing the outcome of efficient (see [351]) auctions. Nevertheless, we do not use a market representation drawn directly from the micro-economics literature, and our representation of preferences, and supply and demand, are relatively rigid and unconventional.

3.5.2 Some alternative representations of the market

A standard route toward describing the interaction of agents in economics is to assign each of them a utility function and assume that they act so as to optimise these functions. For example, in our case the utility functions would be a function of how many animals of each breed type are owned, and potentially of wealth. Two examples of how these functions might be used are as follows. Firstly, one could attempt to derive the relative demand for each breed in each patch, and the maximum supply of animals, and use these to determine the population-level movement patterns of movement. Alternatively, one could track the ownership of different animals and implement trades by comparing the utility functions and patterns of wealth and ownership of pairs of individuals, with trades realised based on mutually beneficial (utility-increasing) transactions.

One useful tool for representing the latter is offered by the Edgeworth Box (e.g. [352]). Here, we consider a barter economy with two goods (Breed A and Breed B), and a utility function for two potentially trading parties. This system can be illustrated as in Fig. 3.30, with the x-axis representing Good 1 and the y-axis Good 2. As there is a fixed total amount of goods (the total owned by the two parties), any point in the box can be used to describe an allocation of the two goods. The initial allocation of goods is position (E_1, E_2), with the endowment of consumer C consisting of a combination of E_C^{G1} and E_C^{G2} units of Good 1 and 2 respectively. The aim is to determine the optimum final allocation of goods given an initial endowment, which could be used in our model to calculate migration rates. It is intuitive to proceed in the analysis of this system graphically. The first step is to use the utility function of the two parties to draw a curve of constant utility for each passing through the initial allocation. This is called an *indifference curve*. The region enclosed by these two lines indicates the set of possible

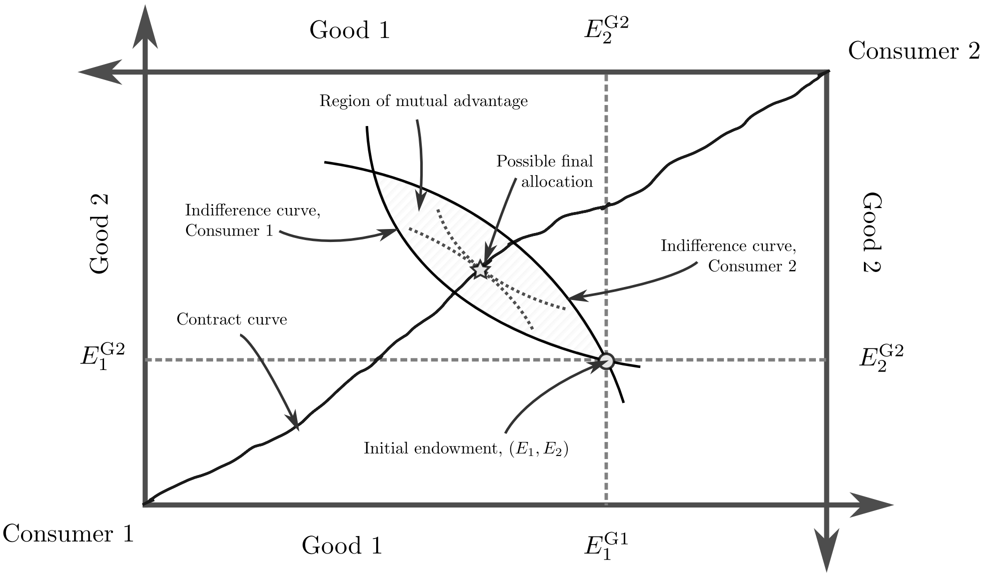

Good 1
E_2^{G2}
Consumer 2
Region of mutual advantage
Possible final allocation
Indifference curve, Consumer 1
Good 2
Indifference curve, Consumer 2
Good 2
Contract curve
E_1^{G2}
E_2^{G2}
Initial endowment, (E_1, E_2)
Consumer 1
Good 1
E_1^{G1}

Figure 3.30: A diagram of the Edgeworth Box, with important features labelled; see main text for details.

allocations that are mutually beneficial for both owners. The question becomes which of these trades should be followed. The optimal distribution of animals (after which no further mutually beneficial trades are possible) is described by the *contract curve*, which is the curve passing through all tangent points of the two indifference curves, given any initial endowment. Points along the contract curve are Pareto efficient, in that the allocation of resources cannot be changed to the advantage of any consumer without disadvantaging another consumer. In this model, we cannot suggest which point on the contract curve within the region of mutual advantage will be the final allocation, which will depend on factors like the bargaining skill of the two consumers.

The Edgeworth Box model does not allow us to describe the entire market system - for example, questions remain as to how one pairs up individual traders, or, if using this representation to describe the interaction of the two patches instead, how to model aggregate, communal utility. Furthermore, in our model the preferences of buyers from the two patches are identical. Trades still occur because we imagine buyers competing for the available animals according to their wealth, but would not under the Edgeworth box model as described above. One would either need to include wealth differences in the utility functions or explicitly model wealth as a commodity using a 3-commodity 2-agent “Edgeworth hyperbox” [353].

Integrating our work into a traditional micro-economics framework would no doubt simplify publication, and may increase the realism of the work. For the present, I emphasise two points. Firstly, our Dynamic Filter models are most aligned with an equilibrium view of market behaviour when supply equals demand, $2b = 2K\bar{m}$. Under these circumstances, both models demonstrate large changes in occupation equilibrium associated with small changes market supply/demand when the fitness advantage of the locally preferred breed is low. This result is important, in that the local availability of animals and the number of individuals seeking to buy animals clearly fluctuate. A relevant quality of the model, then, is the long-term average effect of these fluctuations, and investigating this behaviour would be an interesting extension to our work.

My second point is that, while our representation of the market might be modified or, by some measures, improved, we find that a wide variety of systems incorporating migration biases lead to system equilibria in which locally less-preferred, and less reproductively successful, animals reach locally high frequency. Given that market trading is expected to bias the flow of animals according to preferences and wealth, our modelling at the very least emphasises the need to consider this process when trying to understand observed animal distributions or alter the genetic composition of herds through breeding programmes.

3.5.3 The trading network

Our two-patch model is the simplest possible topology of mutually-interacting patches. This makes characterising the system easier, but actual animal movements occur over a complex trading network. A range of studies have characterised this network in the context of modelling livestock disease epidemics, making extensive use of animal movement data to build a realistic representation of movement patterns (e.g. both foot and mouth disease and bovine tuberculosis in the UK [285, 286, 284], epidemics affecting sheep in Scotland [287], and HPAI/H5N1 bird flu in Vietnam [281] and Cambodia [282]). Unfortunately, countries undertaking detailed tracking of animal movements tend to be those with advanced livestock industries in which breeding is tightly controlled. Nevertheless, these studies offer some guidance as to how available data might be analysed, and suggest an approach to capturing patterns of gene flow that avoids an explicit model of the trading process itself. There is data from the development literature on the movement of animals at a village-scale that might be able to guide future models - examples being the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) Village Dynamics in South Asia (VDSA) surveys and the Rural Economic and Demographic Survey (REDS) which were used by Anagol, for example, [310] in his analysis of the Indian cattle market.

3.6 Conceptual and experimental extensions of the market genetics framework

While we have chosen to investigate a single, simple market-genetics model, there are plenty of further questions that might be asked about the interaction between trading and genetic variation. I mention two, both related to breed formation, now.

3.6.1 The Market for Lemons and the emergence of breeds

The Market for Lemons is an economic model described by George Akerlof in 1970 [354] to describe the impact of information asymmetry between buyers and sellers on trading. A 'lemon' refers to a used car that is found to be of low quality only after the purchase has been made. The model dynamic can be described as follows. Consider a market consisting of many buyers and sellers of used cars. A seller knows more about his car than prospective buyers. As buyers have no prior information about the quality of cars, they will tend to offer the price they feel is appropriate for an average quality car. Sellers of high quality cars are likely to withdraw from the market as this price will be too low. This reduces the average quality of cars on the market, which reduces the price buyers should offer, and, by feedback, may lead to the complete collapse of the market [8].

A question arises as to what extent well-defined animal breeds allow traders to escape the Market for Lemons in livestock markets by reducing information asymmetry. Certainly animals from well-defined cattle breeds tend to obtain higher prices in India [318, 315, 336], though it is unclear to what extent this is due to reduced variance in productivity as opposed to improved productivity. An even more intriguing question is the extent to which the Market for Lemons promotes breed formation. There are two possibilities. If higher prices reduce the number of potential buyers, then wealthy regions will tend to accumulate animals with lower variance in productivity and linked visual traits, which we can call breeds. Alternatively, or concomitantly, the fact that clearly defined breeds reduce information asymmetry and thereby mitigate the Market for Lemons in leading to market collapse may mean that they flow through the market system more rapidly, disrupting breed formation.

3.6.2 Social stratification and the emergence of breeds

When conducting genetic sampling of cattle in India in 2009, I observed many different approaches to controlling reproduction in herds. Although artificial insemination using

[8] The literature building upon the Market for Lemons is vast - the paper had 22,497 citations as of September 2015. Under the model I have described, there is an interesting property whereby the average quality of a car actually is the true quality of the car when only one car is being sold. Once this car is sold, the next lowest-quality car may be sold, and so on. Investigation of details such as this, either in Akerlof's formal model or variations that better allow for it, probably exist among the many citations.

imported semen is becoming very widespread, farmers who did not crossbreed their cattle often did not control breeding in an active way. Most males, were, however, castrated for use as draught animals, and farmers with smaller herds especially were unlikely to own a reproductive bull. Thus, male animals might be shared within villages. In the past, reproductive males were often contributed for communal service by a more wealthy landowner.

The details of how reproduction is controlled in domestic animals, and the implications of this for breed formation, is very interesting. For example, it has been reported that 80% of Indian cattle are nondescript *desi* animals [339], and this will inevitably be related to the traditionally limited control of reproduction. Wealthier farmers are more able maintain self-sufficient herds or specifically chosen animals. The extent to which such herds are important in forming the nexus for breed formation, as opposed to simple geographic isolation, is an interesting subject that might be explored through modelling.

3.7 Simulations and modelling

The models I have explored in this chapter explore the implications of heritable animal movement biases, whereby the rate of migration depends on the animal type and the direction of travel. I have suggested that this allows for the representation of animals trading, and identified model behaviour that would be highly unexpected given isotropic or directionally symmetrical migration. In particular, under a range of different models (Dynamic Filter 1, Static Filters 1, 2 and 3), I find that it is possible for an animal type with globally low fitness to reach high frequency either locally or globally. Several models also suggest that market effects that might be expected to promote the assortment of animals according to their local fitness (Dynamic Filter 2 and Static Filters 3 and 4 when selection is symmetrically opposite in the two patches) can in fact lead to a distribution of animals that is highly non-optimal for the 'wealthier' patch.

These results are somewhat surprising, and are related to the interaction between migration and population regulation. They are robust to the implementation of slower population growth through a Beverton-Holt equation, though damped. The strong implication of our work is that breaking a common assumption in population genetics and ecological models - that animal movement is isotropic/directionally symmetric - can lead to radically different model behaviour.

My results derive almost entirely from numerical simulations, which have further helped to characterise and visualise system dynamics, and hence identify important narratives of model behaviour that might be sought out in real data or investigated in more detail mathematically. Furthermore, simulations were vital in testing the specific representation of the auction that we apply, and extending our model to probe important assumptions such as initial conditions and the representation of population growth. As noted above, some of our findings contrast with those of Bolnick and Otto [267], who did not identify the surprising behaviour that is generated when Static Filter 3 is applied to their system (Appendix 3, §3.3.8) because they focussed on a model variant more amenable to mathematical analysis.

How should this chapter be viewed, then, in the broader context of the thesis? In Chapter 2, I focused on a modelling assumption, that populations are infinitely divisible, taken in order to simplify the manipulation of a model of species invasions. The work primarily focused on improving model design by identifying unintended consequences of this assumption. Here, my main focus has been on extending a model to describe a system that is quite different from the one it was originally designed for - the impact of animal trading on livestock population dynamics. I follow Bolnick and Otto in breaking the assumption of symmetric, unbiased migration found in many population models, and adapt this, breaking the assumption of a constant migration rate as well. This allows me to extend the representation of the model. Thus, I argue that simulations are useful for flexibly assessing the implications of modelling assumptions, but that there may be multiple reasons for doing so - confirming the validity of an assumption taken to simplify manipulation of a model is one, while completely modifying the system targeted by the model is another.

In the next chapter, I extend this argument further, showing how simulations can be used to assess test statistics intended to detect natural selection in population genetic data. In particular, I consider how breaking the assumption of a constant recombination rate, known to be an incorrect, can help to us to improve the design of test statistics.

Chapter 4

Modelling natural selection: improving statistics to detect selection in population genetic data

4.1 Chapter introduction and summary

In this chapter, I use simulations to assess statistics designed to detect signatures of selection in population genetic data[9]. I am able to quantify the relative performance of different methods, and to suggest alternatives to currently used statistics. I then apply the best performing methods to previously-studied genetic data, confirming the existence of signals as known selection candidates. The work is a methodological contribution to answering the question of which genes or genomic regions have been subject to natural selection in the recent past. For example, humans show marked phenotypic differences but are genetically very similar. Some of these phenotypic differences are due to selection. When a trait offering a fitness advantage is associated with a genetic variant, this variant is more likely rise in frequency in a population, leaving signatures in the genome that can later be used to identify it.

The focus of this chapter marks a transition between theoretical and applied work in the thesis, but also a change in the way I use models. Both the first and second chapters involved designing plausible representations of population dynamics, focussing on migration and selection, to broadly describe the implications of different ecological processes on species dispersals and distributions. The models presented were inspired by

[9] The design and testing of selection statistics is a sizeable, and sometimes technical, subject area. To guide the unfamiliar reader through specialist genetic terminology, I include a glossary as Thesis Appendix A7. The first occurrence of a word that is defined in the glossary is underlined.

real world problems, but offered an exploration of possible worlds rather than a direct appraisal of the relevant one. Here, I take a more direct approach - using a well-explored and flexible model of molecular evolution, coalescent theory [355], to simulate realistic genetic data, which I then use to improve statistics intended to identify regions of the genome subject to natural selection through genome-wide selection scans (GWSSs). Using simulations to test the performance of selection statistics in evolutionary genetics is a standard approach (e.g. [356, 357, 358, 359, 360, 361, 362, 363, 364, 365]). I focus on the performance of test statistics based on pairwise linkage disequilibrium (LD) when an assumption that the recombination rate is constant over the genome is discarded. This assumption is commonly applied (e.g. [357, 360, 363, 364], but see e.g. [359, 361, 362] in the context of certain selection statistics that I do not investigate) and is known to be incorrect for a diverse range of species (e.g. [366, 367, 368, 369]). While test statistics are merely functions of some data, they are used and designed with a certain model of the data in mind. If the model makes incorrect assumptions then our beliefs about the performance (e.g. power and efficiency) of the test statistic may also be incorrect.

The ultimate aim of GWSSs is to make specific predictions about the past course of evolution. When applied to humans, this approach has the potential to both inform us about our evolutionary history, and to identify loci that may be medically important. However, questions have been raised about the validity of GWSSs findings, largely due to the possibility of false positive results and the difficulty in confirming selection candidates ([370, 371, 372]). Theoretical issues also arise. In the 1970s, the ‘adaptationist’ idea that a close correspondence between species characteristics and their habitats can be used as evidence of evolution was criticised [373]. In a similar vein, a signal of selection in population genetic data is not direct evidence of selection, and care is required when interpreting selection candidates [374, 371].

I begin this chapter by introducing linkage, an important phenomenon in genetics that helps us use GWSSs to infer which genes may have been subject to selection. With this property clarified, it is easier to visualise why different selection statistics work, and I continue by clarifying the concept of a signal of selection by describing several signatures that are used by statistics designed to detect selection. Having explained the basic principle of a selection statistic, I discuss the assumptions behind the search for genes and genomic regions subject to selection, before describing the design and statistical details of GWSSs. I then explain the standard approach used to assess statistics designed to detect selection through simulations, and present a manuscript in which I use both simulation and real genetic data to assess, refine and build upon a group of statistics designed to detect distortions in patterns LD caused by directional selection. Finally, I explore certain subtleties of the simulation method I apply, before discussing the relationship of work presented in this chapter to the broader use of simulations in population biology.

4.2 Signatures of natural selection

GWSS studies make use of a wide, and ever growing, variety of statistics to suggest selection candidate regions using population genetic data. These are based on several different genetic signals that arise due to natural selection, many of which are easily detected due to genetic linkage.

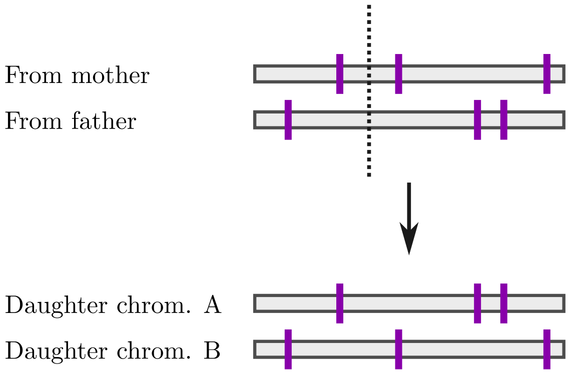
4.2.1 The concepts of genetic linkage and recombination

When a specific allele of a gene is associated with a high-fitness phenotype, the frequency of that allele will tend to increase in a population due to natural selection. Conversely, when the phenotype has low fitness, the frequency of the allele will fall. However, genes are not inherited in isolation. Rather, they are aranged in a specific sequence on a chromosome, with potentially large intergenic stretches of DNA separating them. Often, species have many different chromosomes, which are usually inherited independently in sexual species, such that a genetic variant inherited by an organism on one chromosome has no bearing on which variants are inherited on other chromosomes. Thus, genes that are on the same chromosome usually have a higher probability of being inherited together than genes on different chromosomes. The extent to which this is true depends on a process called recombination.

There are many different modes of reproduction among different species; in some cases, species switch between different modes depending on environmental conditions or the stage of their life cycle. In sexual, diploid species such as humans, each non-reproductive cell of the organism carries two copies of each autosomal chromosome, one from the father and one from the mother. Recombination occurs during the production of the haploid gametes (e.g. sperm and eggs) and involves these pairs of chromosomes swapping chunks of DNA, as illustrated in Fig. 4.1. This will tend to reduce the degree to which genes on the same chromosome are inherited together. The probability of recombination between two loci on the same chromosome is their genetic linkage, with high linkage indicating that the probability of recombination is low. At the other extreme, loci that are at opposite ends of a long chromosome with a high probability of recombination between them each generation will show near-independent inheritance, approximating the evolutionary behaviour of loci on different chromosomes.

The linkage between of two polymorphic loci has an important effect on the association between two alleles at those loci. When the allelic state of a locus is informative about the allelic state at another locus in a population, these loci are said to be in pairwise linkage disequilibrium. Conversely, when this is not the case, loci are in linkage equilibrium. In practice, unless a population is very large indeed most alleles will be in at least some LD in a given generation. In a small sample of a population, even alleles that are close to linkage equilibrium at the population level will often show some degree of LD.

FIGURE 4.1: Schematic diagram of the effect of meiotic recombination involving crossover. During the production of gametes in a sexual organism through meiosis, sections of the chromosome inherited from the mother and father may swap, leading to two recombinant daughter chromosomes. The grey rectangles represent homologous chromosomes (e.g. they all might be copies of Chromosome 1), and the purple lines indicate genetic variants with the derived state. The dotted line shows the location of the recombination event.



If the rate of recombination was constant along the genome, then the physical distance between two genes would be proportional to their linkage. However, the rate of recombination usually varies considerably over the genome. A description of the position of loci on a genome in terms of their linkage is called a genetic map, and the distance between two loci on such a map, which is intended to approximate their linkage, is called the genetic map distance. Genetic maps can be approximated using observed patterns of LD along the genome in a sample from a population (e.g. [375]) or by inferring recombination events in pedigrees (e.g. [376]). Sperm typing [377] can also be useful for obtaining patterns of recombination rate variation in individual males.

4.2.2 Measuring linkage disequilibrium

Pairwise LD can be quantified in several ways. The two most popular statistics are Lewontin's D' [378] and r^2 [379]. Both are normalised measures of the correlation between bi-allelic loci as described by D,

$$D = p_{a,b}^{1,2} - p_a^1 p_b^2, \tag{4.1}$$

where p_a^1 is the frequency of allelic state a at locus 1, p_b^2 is the frequency of an allele b at locus 2, and $p_{a,b}^{1,2}$ is the frequency of genotype $[a, b]$ at positions $[1, 2]$. D' is defined as

$$D' = \frac{D}{D_{\text{MAX}}} \tag{4.2}$$

where

$$D_{\text{MAX}} = \begin{cases} \min\bigl(p_a^1 p_b^1,\ (1 - p_a^1)(1 - p_b^2)\bigr) & \text{when } D < 0 \\ \min\bigl(p_a^1 (1 - p_b^2),\ (1 - p_a^1) p_b^2\bigr) & \text{when } D > 0 \end{cases}$$

while r^2 is

$$r^2 = \frac{D^2}{p_a^1 (1 - p_a^1) p_b^2 (1 - p_b^2)}. \tag{4.3}$$

The absolute value of D' equals one whenever all four possible genotypes described by two bi-allelic loci are not observed, such that rare alleles will frequently have high D' values. As I shall later explain, positive directional selection tends to increase the number of rare alleles in a genomic region, such that the average value of D' in a genomic region is likely to capture this signature of selection as well as any distortions in the association between alleles. Previous selection statistics based on LD have tended to focus on r^2 (e.g. Kelly's Z_{nS} [380], ω [357]), and given these reasons, I choose to use r^2 as the measure of pairwise LD throughout this chapter.

4.2.3 Genetic hitchhiking and the interaction between linkage and selection

To illustrate the importance of linkage for the localisation of genes subject to selection using GWSSs, it is useful to consider three possible scenarios arising given positive directional selection acting on a novel mutation in a sexual, diploid species - a high rate of recombination such that there is no linkage, no recombination such that the chromosome is in complete linkage, and the intermediate case, which is more biologically realistic for sexual organisms.

The signature of selection when there is no linkage If the recombination rate was exceedingly rapid, such that recombination events are probable each generation even between loci that are very close together on the chromosome, the inheritance of different genetic variants on a chromosome could be considered independent. The new, positively selected allele would rise in frequency without influencing the frequency of other alleles.

The signature of selection when there is complete linkage If there is no recombination, a novel genetic variant will almost always be found in conjunction with the variants that were on the chromosome on which it initially arose. The only process that can disrupt this is mutation. The entire chromosome upon which the selected variant first arose can be called its haplotype, and selection can be thought of as acting on this haplotype. Equivalently, we can view the entire haplotype as hitchhiking on the selection acting on the advantageous genetic variant. Assuming that the novel variant leads this haplotype to be the fittest version of the given chromosome in the population, the entire haplotype will rise in frequency over time.

The signature of selection when there is moderate linkage If the rate of recombination is moderate, loci on given chromosome that are nearby will be in greater LD than distant loci. A novel selected allele will, as in the case of complete linkage, initially exist on a single haplotype. However, recombination events will rapidly lead the selected variant to be associated with multiple alleles at long distances. Equivalently, it will appear on different haplotypes when a large genomic region is considered, but will tend to be on one or a few haplotypes when a short genomic region surrounding it is considered. Over time, the selected allele will rise in frequency, but alleles at loci in linkage with the locus subject to selection will also change in frequency.

The signal of selection is different in each case. When there is no linkage, selection can only be detected by observing the change in the frequency of the selected allele, either between populations or over time. If selection is detected, the exact position of the locus subject to selection is self-evident. The converse is true in the case of complete linkage. Here, evidence of selection can be detected by observing the frequency of different chromosome-length haplotypes. Frequency differences - again, over time or between populations - of any allele on the relevant chromosome can indicate selection on that chromosome, but it is difficult to identify the locus subject to selection.

The intermediate case is more biologically realistic for sexual, diploid organisms. Here, selection changing the frequency of an allele also impacts the frequency of alleles at other loci, especially if they are in tight linkage. This process is known as genetic hitchhiking [381]. The population genetic signature of a selection, then, will be focussed on the genomic location of the allele subject to selection, but will extend around it, such that it is theoretically relatively easy to detect and localise selection through GWSSs.

4.2.4 Some signals used to infer selection

The focus of this chapter is on only one signature of selection, distortions in patterns of linkage disequilibrium. However, an awareness of the expected effects of selection on genetic variation helps to clarify what a signal of selection is and the role of LD-based statistics in GWSSs. Focussing on signals of recent selection based on population genetic data from a single species, signals can be classified as follows [382, 370, 383]:

Population differentiation
Population genetic samples from two different populations will often have different frequencies of a given allele. This is due to several factors - how long ago the populations diverged, the extent of migration between the populations, the intensity of genetic drift in each population since divergence, and sampling effects are four examples. The average allele frequency difference across the entire genome is one measure of population differentiation. Selection can lead to distortions in allelle frequency differences between two populations in the genomic region subject to selection. This pattern of anomalous genomically local population differentiation can be used as a signature of selection. Unusually high population differentiation can indicate directional selection impacting the frequency of a variant in one of the populations but, potentially, not the other. Unusually low population differentiation may suggest that a genomic region is under strong

purifying selection. Statistics designed to assess population differentiation include F_{st} [384] and the population branch statistic [385].

Reduced genetic variation

When directional selection increases the frequency of a rare variant, it increases the frequency of alleles that are on the same haplotype. This corresponds to a reduction in the frequency of alleles that are not on the same haplotype, leading to a reduction in genetic variation as alleles become extinct and loci monomorphic in the population. Purifying selection has a qualitatively similar effect, in that novel variants with a negative fitness effect will be selected out, along with any rare alleles in LD with them. Thus, anomalously low levels of genetic variation in a genomic region can signal directional or purifying selection. Balancing selection can have leave a complex signature on neutral variation [386], but in some cases at least has lead to high nucleotide and haplotype diversity, the human leukocyte antigen (HLA) genes, which play a critical role in immune response, being an example. Here, anomalously high variation is suggestive of selection.

Site frequency spectrum

The site frequency spectrum (SFS) is distribution of allele frequencies in a sample at all polymorphic loci over a specified genomic region. It can therefore be used to assess the proportion of rare, intermediate frequency and common derived alleles. The genome-wide SFS is affected by many demographic processes, including changes in population size and admixture between partially isolated populations. When selection changes the frequency of a particular allele, the frequency of alleles at linked loci will also change, known as genetic hitchhiking [381]. This impacts the SFS. For example, positive selection tends to lead to a relative excess of low- and high-frequency derived alleles [387]. A wide range of statistics designed to detect selection are based on the SFS (e.g. Tajima's D [388], the Composite Likelihood Ratio (CLR) statistic suggested by Kim and Stephan [356], Fay and Wu's H [387]). These tend to have high power to detect recently completed selective sweeps that acted on a novel variant (rather than on standing genetic variation).

Linkage disequilibrium

Although I have presented the three signals above in the context of selection influencing the frequencies of the target allele and linked neutral variants, one can also view selection as changing the frequency of the haplotype surrounding a selected allele. Thus, each of the above signals can be considered in the context of haplotypes (e.g. increased haplotype differentiation between populations, or reduced haplotype diversity). As has already been indicated, in the case of a selective sweep acting on a novel variant, that variant is initially associated with a single haplotype (the chromosome on which it arose). Recombination rapidly causes selected alleles to be on different haplotypes at long genetic map distances, but the effect is much slower at a short genetic map distance. Thus, selection leads to a local distortion in the 'haplotype frequency spectrum'. The characteristic signature of a high-frequency, long-range haplotype has been used to infer

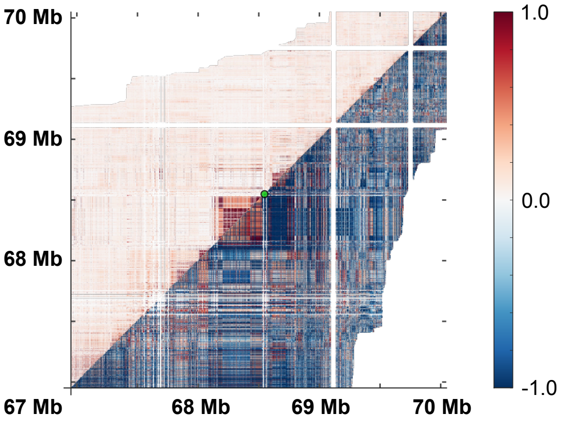


Figure 4.2: LD pattern around the selected single nucleotide polymorphism in *CPT1A*, c.1435C>T[11], the position of which is indicated by a green dot. The top-left of the matrix shows raw r^2 values, a measure of LD introduced formally below, while the bottom-right shows $\log \frac{r^2}{E[r^2]}$. $E[r^2]$ denotes the expected value of r^2 given the genetic map distance between loci, calculated using the observed r^2 values for all loci on Chromosome 11 in the northeast Siberian population and the HapMap Phase II genetic map [390]. The reduction in LD between single nucleotide polymorphisms on either side of c.1435C>T is indicative of a late-stage selective sweep [357], and is emphasised by taking into account expected r^2 values. Values >1.0 and <-1.0 were plotted as 1.0 and -1.0 respectively.

recent and ongoing positive selection especially, with a variety of selection statistics based on this signature (e.g. extended haplotype homozygosity [389], integrated haplotype score [359] and nS_L [365]).

The changing frequency of different haplotypes due to selection has an effect on pairwise LD, defined above. In the case of positive directional selection, an increase in LD is expected in the early stages of the selective sweep. More complex patterns arise as a selected allele nears fixation, with a characteristic reduction in LD between loci located on different sides of the selected variant (see manuscript below for details and references). This latter pattern is apparent in Figure 4.2, which shows the pattern of LD around a variant in the *CPT1A*[10] gene, identified as a likely selection candidate by Clemente *et al* [2]. This particular variant is at a high frequency in Arctic populations despite being associated with high infant mortality in Canadian and Greenland Inuits.

During the course of my PhD, I contributed a similar figure to that study (Supplementary Figure S5 in [2]), which included the pattern of LD in control populations, along with a discussion of the LD patterns in this region. My contribution helped to narrow down the variant likely to be associated with a phenotype that was positively selected in the past, possibly adaptation to a high-fat diet (see the study for evidence supporting this hypothesis).

[10] Following convention, genes name abbreviations are written in italic upper case

[11] This notation indicates a single base variation at position 1435 in the coding (c.) sequece of the gene, from an ancestral state 'C' (cytosine) to a derived state 'T' (thiamine)

The above signatures are those most commonly used to infer relatively recent selection acting on populations within a species. Other approaches exist, especially based on multiple-species comparisons (some are detailed in [383]), but are not directly relevant to the work in this chapter. Similarly, multiple signatures are often combined to detect selection (e.g. the composite of multiple scores (CMS) statistic [362] or the recent S/HIC statistic that uses machine learning to characterise the combined implications of a wide range of selection signatures [391]).

As I have already emphasised, the genomic region over which a signal of selection extends depends on linkage between the locus on which selection is focussed and nearby loci. This in turn depends the local recombination rate, which is highly variable. Briefly, if the recombination rate is unusually low then haplotypes associated with high frequency derived alleles may be unusually long, local LD will be unusually high, and the impact of selection on reducing local genetic variation and distorting the SFS will affect a larger genomic region. Unusually rapid recombination may mask signatures of selection, and lead to false positives for some statistics based on LD especially (again, see manuscript below).

A slightly subtler, but equally concerning, point is as follows. I have observed in simulations that low recombination rate tends to lead to more 'noise' in a wide range of population genetic summary statistics. This is because the genealogies[12] of loci in such regions are highly correlated - in the limit of no recombination, they all reflect a single genealogy, which describes the evolutionary history of the entire chromosome. When the local recombination rate is high, different loci evolve with near-independence and many possible genealogies are effectively sampled from those possible given a population's demographic history. The increased noise associated with slow recombination rate will lead to outlier signals in many statistics.

In this section, I have attempted to explain the concept of selection signals through important examples of signals that are frequently used in selection statistics, and the important role that recombination has in allowing us to localise detect signals of selection. I continue by describing some of the challenges that arise when trying to interpret these signals.

4.3 Challenges in identifying natural selection

Natural selection describes how, given heritable phenotypic variation that leads to differential reproductive success, phenotypic qualities that yield more offspring will tend

[12] See my discussion of the coalescent in section §4.5.1

to be emphasised - termed adaptation. This is a fundamental principle underlying the evolutionary process. Nevertheless, the role of adaptation in explaining the phenotypic (e.g. [373]) and genetic characteristics [392] of species has been extremely controversial.

4.3.1 Natural selection affecting allele frequencies in humans

A gene or genetic variant may be considered a promising selection candidate if there are multiple lines of evidence pointing toward adaptation, and especially if these link genotype, phenotype and fitness [371]. Gaining truly cohesive evidence is especially difficult in humans, where experimental manipulation of genes and environment is usually unsuitable in a scientific context. The list of genes that are thought to influence a selected phenotype based on GWSSs and other evidence continues to grow (e.g. examples in [393]), though at a far slower rate than those identified by GWSSs only [370]. To indicate the process by which evidence may be accumulated, I now describe a classic example - the effect of variation in the *MCM6* gene on lactase persistance.

Lactase persistence Lactase persistence (LP) is the expression of the lactase enzyme into adulthood, leading to the ability to digeset lactose after weaning. This allows adults to consume dairy products without discomfort, and hence access another source of nutrition. The LP phenotype is genetically determined and the distribution of LP follows the distribution of dairy farming (see citations in [394]). This has led to suggestions, from the 1970s onward, that the distribution of LP is a result of selection [395]. Population genetic investigation lead to the identification of two genetic single nucleotide polymorphisms (SNPs) associated with the LP phenotype in a Finnish cohort [396] located in *MCM6*, the gene neighbouring the lactase gene (*LCT*). The derived state of both SNPs is common in northern Europeans (which would suggest that the original mutation was old under neutrality) but is on a long haplotype (a signature of a recent mutation), as well as being in a genomic region of high population differentiation; combined, these signals suggest that the SNPs rose in frequency in some populations very quickly, indicative of natural selection [394].

Subsequent work has identified different SNPs associated with LP in Africa [397, 398] and the Middle East [399]. There is considerable evidence that several of these SNPs impact gene expression (e.g. in a human intestinal cell line for certain SNPs found in Africa [397] and in transgenic mice for the -13910T SNP (rs4988235[13] , one of the European SNPs) [400]). Finally, work on ancient DNA (aDNA) from European human remains has demonstrated a sharp, recent rise in the frequency of SNP rs4988235 over the last 5000 years [401]. For comparison, time-estimates based on modern genetic

[13] Specific SNPs are identifed by researchers using their unique rsID (Reference SNP cluster ID), which is assigned on submission of a novel SNP to the main database of SNPs, dbSNP

patterns[14] only suggested a date of 8-9000 years (95% confidence 2200-19200) [397]. Taken together, these studies identify SNPs showing evidence of a genotype-phenotype association, the potential to increase gene expression in both human cell lines and a mouse model, and, indirectly, fitness effects through observed frequency changes and signatures of selection in modern populations.

There is a similar breadth of support for selection acting on several other SNPs. For example, a variant in the *SLC24A5* gene is associated with skin pigmentation in multiple human populations [402, 403], and both directly impacts pigmentation in zebrafish and shows genetic signatures of selection in moden European populations [402]. The variant also shows a rapid frequency increase in Europe in the last 8000 years [401]. I search for selection signatures at both of these genes in the manuscript presented below.

I also consider a variant in the *EDAR* gene thought to have been subject to selection among Asian populations [404, 405]. This variant was found to be associated with hair thickness [406] and tooth shape [407] in modern Asians. A mouse line with the variant showed increased hair thickness, reduced mammary fat pad size but increased mammary gland branching, and increased eccrine sweat gland number [408]. After identifying these associations in mice, the association with eccrine gland density was replicated in humans [408]. This diversity of effects (pleiotropy) means that it is difficult to determine which phenotypic trait, if any, offered a selective advantage, and hence lead to allele frequency changes.

4.3.2 From differential reproduction to allele frequency changes

The situation with the *EDAR* variant is an interesting example of the challenges involved when interpreting signatures of selection in population genetic data. However, this particular case is an anomaly. Most of the time, a GWSS will highlight many regions of the genome, each of which may be several hundred kilobases long. Within each selection candidate there may be multiple (or no) genes and thousands of SNPs, as well as other genetic polymorphisms. Many genes have multiple biological roles. However, the large number of genome-wide association studies, with differing protocols and sample sizes, that have now been performed mean that genes often show putative association with multiple diseases or phenotypic traits. Even if a single SNP is considered a likely selection target information on its phenotypic impact in model organisms will usually have to be specially generated. These features together mean that it is hard to narrow down the loci contributing to signatures of selection, identify likely phenotypic associations, or test the impact of a variant in model organisms.

[14] I do not describe techniques to date the onset of selection or estimate the selection intensity in this thesis - the chapter is already long and I do not attempt to do either in the work presented below

Given these difficulties, inferring the narrative of adaptation from population genetic patterns is especially challenging [371]. But the narrative may be unclear even when there is supporting evidence for selection from multiple directions. For example, if a genetic variant known to have a specific phenotypic effect and also displays population genetic signatures of selection, it does not logically follow that the frequency of the allele has been influenced by selection on that phenotypic effect [374]. Furthermore, many complexities to the simple selective narrative, often the scenario selection statistics are designed to detect, are possible. An allele may have a phenotypic impact that is difficult to observe or not obvious to a scientist, or may impact many traits in ways that do not influence reproductive success. Alternatively, the phenotypic impact of an allele may depend strongly on the environment, or on the genetic state at other loci (epistasis), or on its own frequency (e.g. heterozygote advantage).

Despite these concerns, the use of population genetic patterns to infer selection has had notable successes. The examples of lactase persistence and pigmentation are joined by many others (examples include high-altitude adaptation [409], blue eye colour in Europeans [410] and height in African Pygmies [411]), with the implication that GWSSs can offer an important contribution to our understanding of evolutionary changes in genotype and phenotype. Nevertheless, given the challenges facing GWSSs, it is particularly important to have confidence in the statistical method employed. My work aims to further our understanding of the statistics used in GWSSs, and hence help to reduce the number of false positive selection signals identified in GWSS studies.

4.4 Statistical details of genome-wide scans for natural selection

To help focus on important statistical details of GWSSs, I begin by outlining a general GWSS protocol. Important terms are *italicised*, with a definition and discussion given below, in order, rather than in the glossary. Note that these include both terms from genetics and statistics, although my discussion focusses on the specific relevance of the term to GWSSs. Readers familiar with statistical concepts and the idea of a GWSS may wish to proceed to section §4.5, where I discuss the use of simulations in assessing the performance of selection statistics.

4.4.1 The standard model of a genome-wide selection scan

The two essential components of a selection scan are the *population genetic data* and a *test statistic* that allows a researcher to assess the evidence for selection in a given region. The population genetic data consists of genetic data from individuals which is partitioned into one or more *population groups*. Following suitable pre-processing of the data - for example, phasing may be necessary for non-haploid species - the test statistic is calculated for each of a number of *genomic regions*. In GWSSs, these genomic regions define subsets of the data to be used as input for the test statistic, and are distributed across the genome to identify the genomic location of unusual signals.

An effective test statistic will have a different distribution given the *null hypothesis* and *alternative hypothesis*. Calculating the *null distribution* of a test statistic is a critical step in identifying signatures of selection. There are two main approaches. Firstly, a demographic model of the population may be assumed, and the null distribution of the statistic then calculated, or, more often, sampled through simulations. Secondly, one can assume that the observed distribution of the statistic over the genome largely represents the null distribution. In the case of genome-wide selection scans, this assumes that the vast majority of genomic regions are selectively neutral. In both cases, a *critical value* of the test-statistic is proposed, and those regions yielding test statistic values more extreme than this critical value are considered *selection candidates*.

4.4.2 Important concepts in GWSSs

Population genetic data - the population genetic data consists of a set of samples for which genetic data is available. For a GWSS, the sequence data for an individual has tended to consist of a large number of known single nucleotide polymorphisms (SNPs), although increasingly whole genomes are available. Using whole genomes avoids problems associated with ascertainment bias. Information on the genomic location of sequenced loci is often required to define genomic regions over which to calculate the test statistic.

Test statistic - A test statistic is defined as a function of a series of observations ([412, pp.374]; [413, pp.246]),

$$W(X_1, \ldots, X_n) = W(X). \tag{4.4}$$

In the context of GWSSs, the observations are the population genetic data at each specific genomic region for which the statistic is calculated. The function can be quite a complicated algorithm, and may focus on one or more of a wide variety of genetic patterns that are thought to be created by selection.

Population groups - A population group is chosen so as to ask questions about the shared evolutionary history of the samples it contains. In human GWSSs, workers are often searching for signatures of selection that are specific to certain human groups, as defined geographically and culturally (e.g. ‘Europeans’, ‘East Asians’; see the HapMap project [414], HGDP-CEPH [415], 1000 Genomes Project [416]). Algorithms exist that cluster samples according to genetic similarity (e.g. STRUCTURE, [417]).

Genomic region - For any test statistic, it is necessary to define the genomic data used as the input. In GWSS, the test statistic is evaluated many times along the genome, often in regularly-spaced 'windows' covering a known physical or genetic distance, testing each for evidence of selection.

Null and alternative hypotheses - A hypothesis can be broadly defined as 'a statement about a population parameter' [412, pp.373]. Hypothesis testing involves making a decision regarding which of two possible hypotheses are accepted given the sample data - the null hypothesis (H_0) or the alternative hypothesis (H_1). The null hypothesis places some proposed constraint on the population parameter θ, and we can state H_0 as $H_0 : \theta \in \Theta_0$. Often, the alternative hypothesis is merely a statement that H_0 is false. In our notation, this is expressed as $H_1 : \theta \in \Theta_0^c$, where Θ_0^c is the complement of Θ_0.

In the context of GWSSs, the desirable null hypothesis is neutrality. The relationship between this and the hypothesis of selection is not trivial, and will be discussed shortly.

Null distribution - The null distribution is the distribution of the test statistic under the null hypothesis. When the test statistic deviates strongly from the null distribution this is taken as evidence that the null hypothesis is not true.

In GWSSs, determining the null distribution of a test statistic under neutrality is a critical step. The null distribution is usually impacted by features of population history such as population size and changes in population size and admixture between populations. Note that this demographic history impacts the entire genome, while selection will tend to affects the genomic region associated with relevant phenotypic traits [418]. There are two options for generating the null distribution of a statistic. The first is to explicitly simulate samples from the population according to some demographic model. This is often achieved using coalescent simulations. The second is to assume that the distribution of the test statistic over every genomic region in the genome approximates the null distribution.

Critical value - A critical value is a value of the test statistic that marks the boundary between values that are taken as insufficient evidence to reject H_0 and those that are taken as sufficient evidence to do so[412, pp.383; 397–398]. The critical value should be chosen before carrying out the hypothesis test, and usually corresponds to a certain probability of a false positive result given that H_0 is true, $P(\mathrm{FP}|H_0)$[15]. Traditionally, $P(\mathrm{FP}|H_0) = 0.05$.

An important consideration is determining a reasonable value for $P(\mathrm{FP}|H_0)$ in situations where multiple hypotheses are tested. This is especially relevant for GWSSs. Testing for non-neutrality in non-overlapping windows of 200kb over the human genome would involve over 10,000 hypothesis tests, with nearby windows often showing correlated

[15] This quantity is often indicated using the notation α, which I reserve for a certain family of statistics in the manuscript presented below

statistic values. $P(\text{FP}|H_0)$ is usually reduced to reduce the number of false positives, and many approaches have been recommended [419].

Selection candidate - A selection candidate is a genomic region for which the null hypothesis of neutrality has been rejected for a given test statistic. This makes no statement on whether that locus has actually been subject to selection, and indeed it is not strictly possible to make this statement using GWSSs alone - H_1, when defined as the complement of H_0, refers to an infinite number of hypotheses, with selection included in only a subset of these. Examples of possible confounding factors in GWSSs are local anomalies in recombination or mutation rate.

As the definition of a selection candidate depends on the test statistic, null distribution and the chosen critical value, we will have differing levels of confidence in selection candidates suggested by different analyses. Some caution is recommended by the fact that by 2009, combining the results of selection scan studies suggested that 23% of the human genome had been subject to detectable selection [370]. This is despite a small human ancestral population size, which leads most methods to have difficulty in identifying even moderately strong selection. Between these studies, the number of hypothesis tests conducted will be very large indeed, raising the possibility of many false positives in the literature.

4.4.3 H_0 and the concept of neutrality

The intention of GWSSs is to search for signatures of selection, such that neutrality is the logical suggestion for the null hypothesis. However, as indicated above there are multiple approaches to approximating the null distribution, and these have implications on what a region being identified as a selection candidate actually implies.

Neutrality in theory

In theoretical population genetics, there are four processes that affect allele frequencies over time - stochastic drift due to finite population size, migration, mutation and selection. Selection is possible when lifetime reproductive success depends on genotype, and changes in allele frequencies due to selection are expected when this is associated with variation at relevant loci (the locus of a given allele or a locus that is linked to it) in the genome.

In mathematical models of genetics, it is common to model two alleles at a locus. Each allele is assigned a fitness, describing its relative reproductive potential of a carrier. Perhaps the simplest model describes a system of two alleles at a haploid locus with frequencies p and q, and discrete, non-overlapping generations,

$$p_{t+1} = \frac{p_t(1+s)}{q_t + p_t(1+s)} \tag{4.5}$$

$$= \frac{p_t(1+s)}{1+sp_t} \tag{4.6}$$

where p_{t+1} is the frequency of allele p at time $t+1$, $q = 1-p$ and s is the selection coefficient, with relative fitnesses 1 and $1+s$ applying to alleles q and p respectively (Eq. II-19 in [420]). The absence of selection at a locus is termed neutrality; in such models, $s = 0$.

The description of selection changing an allele frequency changing over time given by Eq. (4.5) is a clear simplification. Although the mathematical model describes the idealised impact of selection quite accurately when the population is large, selection strong and the advantageous allele at an intermediate frequency, it does not include random genetic drift. Furthermore, features like fitness depending on environmental factors or frequency dependent selection are not included.

Selection and neutrality are easy to define in the context of models, but such definitions are only as useful as the models are good representations of real evolutionary processes. Given this, I approach the problem from a new angle by asking what null hypotheses is implied by the chosen null distribution in a GWSS.

Constraints on the null hypothesis through the null distribution

As mentioned above, the null distribution is usually determined in one of two ways - through simulation, or empirically using the observed distribution of the test statistic for the relevant population genetic sample.

When using simulations, a demographic model is selected, and genetic data is simulated according to this demographic model many times. The demographic model is usually retrieved by defining a plausible model structure, describing features like the number of populations modelled and how often these change in size, and fitting the parameters of the model to genetic data (examples in humans include [421] and [422]). The precise null hypothesis being tested when such a null distribution is used, then, is *local consistency with the stated demographic model.* The demographic model is a highly simplified representation of a population history that is most consistent, as compared to other parametrisations tested, with certain aspects of the population genetic data. While many true features of the population's history are not modelled, the effect of these on genetic diversity may be partially absorbed by parameters that are included. This leads to a deviation between the demographic model and the most accurate population history, but can be desirable for the purposes of detecting local selection signatures.

The alternative approach is to simply use the observed distribution of the test statistic over the whole genome as the null distribution. In this case, one inevitably captures many aspects of the data - the genome-wide influence of demography, variable recombination and mutation rate, and more complex patterns such as assortative mating driven by specific loci and natural selection. The null hypothesis, then, is *local consistency with the genome-wide statistic pattern*.

The relationship between the null hypothesis and neutrality is different in the two cases. In the former, the stated demographic model is chosen so as not to include selection (i.e. $s = 0$ for all loci in the model). However, the data used to fit the demographic model may include loci that have been subject to selection, with the impact of this partially absorbed into other demographic parameters. The null distribution reflects one of infinitely many neutral demographic models, such that any local deviation from this model due to variation in recombination rate or mutation, or due to complex non-selective processes such as an allele affecting mate choice or migration tendencies, may lead the null hypothesis to be rejected. Incorrect estimation of the neutral model can lead to huge numbers of false positives (e.g. [423]).

In the latter, outliers (based on a chosen threshold, such as the top 1% of signals) are taken to be selection candidates. The extent to which these are true positives - in the sense that the genetic variation in these regions has been affected by selection within the time frame probed by the selection statistic - is unclear. Ultimately, this will depend on the power of the test statistic given the true sample genealogy, weighted according to the frequency of different selection events. Although the demographic scenario no longer needs to be estimated, outliers may appear to offer convincing selection candidates if selection is rare and the null-distribution under neutrality fat-tailed.

The genomic signature of neutrality

However the null distribution is constructed, GWSSs rely on the idea that demographic history impacts the entire genome equally (albeit stochastically), while selection tends to affect local genomic regions only [418]. However, selection is not the only process that affects different regions of the genome in different ways. The recombination rate, for example, varies considerably over short genomic scales [367, 390], and increasing evidence suggests that the mutation rate may be quite variable also [424]. Different populations have different recombination hotspot locations [425], such that these must change over time.

Specific local patterns or changes in recombination rate and mutation rate over time could create many of the genomic patterns that are created by selection. The simplest examples include regions of low genetic variation or high LD, both signals of selection, as a result of low local mutation or recombination rate. More complex changes in local mutation rate could replicate distortions in the SFS expected to arise due to selection.

This by no means suggests that most selection candidates have been caused by quirks of local recombination or mutation, but emphasises the challenges in linking selection candidates to the process of selection itself. If the null-distribution of a test statistic is to truly represent neutrality, it would have to take into account these sort of phenomena. Unfortunately, processes such as changes in local mutation and recombination rates over time remain poorly characterised. While a null distribution approximated using the empirical distribution of a test statistic will reflect the impact of locally variable neutral features of molecular evolution, the relative frequency of such signals as compared to true selective sweeps is unclear.

The above discussion is not exhaustive, but broadly covers approaches in GWSSs and highlights critical challenges that these face. Many of these challenges have not been fully resolved. Simulations offer a useful tool for exploring the implications of the various assumptions in GWSSs or specific selection statistics.

4.5 Using simulations to improve test statistics for GWSSs

Simulations can be used to assess the accuracy of GWSSs when different approaches are taken to constructing the null distribution. They can also be used to assess the ability of a test statistic to correctly identify signatures of selection. In my work, I have considered the performance of test statistics rather than the details of constructing the null distribution, and I therefore focus on the latter.

The principle of using simulations to assess a test statistic in this way is as follows. Firstly, genetic data is simulated according to a neutral demographic scenario with a given model of molecular evolution, and the test statistic evaluated based on this data. Repeating this yields a null distribution of the test statistic. Secondly, genetic data is simulated according to the same model of demography and molecular evolution, but with some form of selection included also. Now, many replicates gives the distribution of the test statistic under this specific violation of the neutral null hypothesis. The two distributions can be compared to assess the ability of the test statistic to distinguish between the two scenarios.

There are two main approaches to simulating genetic data. Forward-time simulations usually represent the entire population as it evolves according to a specified scenario. If I want a sample of $n = 10$ chromosomes from a Wright-Fisher [16] population of size

[16] A Wright-Fisher population is a standard idealisation of a finite and constant-sized population of identical, randomly mating individuals which evolves by replacing a parent population with a daughter population each generation [13]. Another commonly used idealisation in forward-time simulation and mathematical analysis is the Moran model [426].

$N = 1000$ at equilibrium diversity, the simplest approach is to initialise the N-member population with identical chromosomes, and then evolve this population forward in time, using a specified model of mutation and recombination, until diversity no longer changes according to my preferred metric. I would then randomly sample n chromosomes from this population. A more common approach is backward-time simulation using coalescent models. Here, the genealogy of the n sampled chromosomes is simulated only. This approach is much more computationally efficient, but is less flexible, especially when trying to represent complex mating systems (i.e. highly non-random mating) or, depending on the algorithm used, highly skewed offspring distributions. A more comprehensive discussion of the difference between forward and backward-time simulations can be found in [427].

The approach to testing selection statistics faces two challenges. Firstly, the simulation protocol is assumed to accurately capture the specified neutral and selection scenarios. Secondly, the scenarios chosen should be relevant to the evolutionary history real populations. I proceed by introducing coalescent simulations in more detail, before presenting the draft manuscript and, finally, returning to these two challenges in the context of the manuscript.

4.5.1 The coalescent simulation method

Coalescent simulations make use of coalescent theory to efficiently obtain a sample of simulated genetic sequences. The theory underlying the concept dates to the early 1980s [355, 428, 429], but relies on and extends previous results [430]. The simplest form of the model is also the first [355], and represents neutral evolution given a single panmictic population of constant size, without recombination. In brief, a genealogy is first constructed for the samples using the distribution of time between coalescent events, which is known, see Eq. (4.7) and Fig. 4.3. A coalescent event occurs when two copies of a chromosome share common ancestry, and can, prior to this, be viewed as a single evolving lineage. I use the term chromosome here to mean a contiguous genetic region. The simulation algorithm is as follows:

1. Start with a sample of n chromosomes partitioned into $k = n$ sets, each containing a single chromosome. Each set contains a group of chromosomes that have coalesced at time t pastward.

2. Repeat, until $k = 1$:

   (a) Draw a random number, T_k, distributed according to the waiting time for the next coalescent event (see below).

   (b) Where t_{k-1} is the time, pastward, at which there are first $k - 1$ sets, $t_{k-1} = t_k + T_k$

(c) Randomly pick two sets of chromosomes and combine them (a coalescent event), setting $k = k - 1$

Membership of the k sets over time can be drawn as a bifurcating tree (see Fig. 4.3). The waiting time for a coalescent event is approximated by an exponential distribution. Formally, with n finite and population size $N \to \infty$, a limiting process is described such that the mean of this exponential is

$$E[T_k] = \frac{1}{\binom{k}{2}}, \tag{4.7}$$

with time re-scaled in proportion to N [355, 431, 432] and $E[T_k]$ the expected waiting time for the next coalescent event given k sets of chromosomes. In the Canning's model [433], which generalises the Wright-Fisher model to allowing for offspring distributions that are not multinomial, time is scaled as N/θ^2, where θ^2 is the variance in offspring number as $N \to \infty$ and is a positive constant. For a haploid Wright-Fisher model [13], $\theta^2 = 1$, while re-scaling occurs as $\frac{N^2}{2}$ [355] in the Moran model [426]. The accuracy of Eq (4.7) in approximating these classic population genetic models relies on $N \gg n$. As k increases, Eq. (4.7) gets smaller, reflecting a reduction in the waiting time associated with an increasing number of different possible coalescent events in a generation. Eq. (4.7) can be used to calculate properties of the tree, such as the time to the most recent common ancestor of all n samples (e.g. [432]),

$$\begin{aligned}
E[T_{\text{MRCA}}] &= \sum_{k=2}^{n} \frac{N}{\binom{k}{2}} \\
&= 2N \sum_{k=2}^{n} \frac{1}{k(k-1)} \\
&= 2N \sum_{k=2}^{n} \frac{1}{k-1} - \frac{1}{k} \\
&= 2N(1 - \frac{1}{2} + \frac{1}{2} - \frac{1}{3} + \frac{1}{3} - \cdots - \frac{1}{n-1} + \frac{1}{n-1} - \frac{1}{n}) \qquad (4.8)\\
&= 2N\frac{n-1}{n}, \qquad (4.9)
\end{aligned}$$

now assuming a haploid Wright-Fisher population and time in generations.

Having sampled the coalescent to obtain a genealogy, it is trivial to convert this into sequence data by randomly scattering mutations on the branches, see Fig. 4.4. Although changes in mutation rate over time are not commonly implemented in computer programs simulating the coalescent (e.g. they are absent in the classic program ms of Hudson), this is easy to incorporate by conditioning the probability of a mutation on time. Similarly, mutations are typically assigned random numbers representing location based on a uniform distribution between 0 and 1, corresponding to an infinite sites model. Local mutation rate variation could be included by using a different distribution.

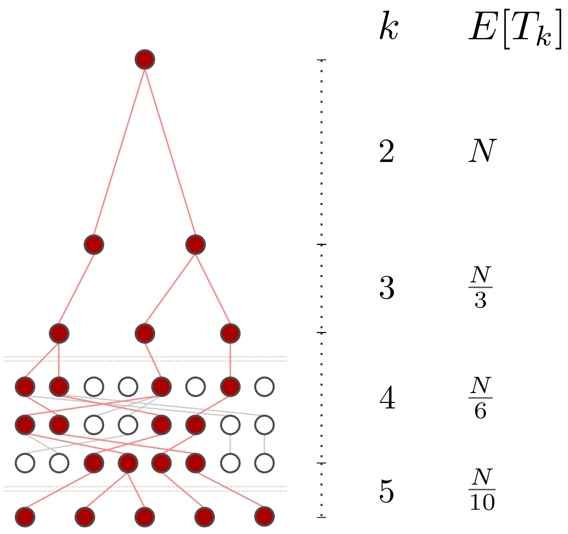


Figure 4.3: The coalescent process given a haploid Wright-Fisher population of constant size $N = 8$ and sample size $n = 5$, with no recombination. In the coalescent model, time runs backward from the present, $t = 0$, to the final coalescent event, corresponding to T_{MRCA}; in the figure, the present is at the bottom of the tree, where $k = n = 5$. Each circle corresponds to an individual in a specific generation, with dark red indicating individuals belonging to lineages that are ultimately sampled in the present and white indicating lineages that are not sampled. The detailed process of reproduction for the entire population over two generations is shown between the two horizontal dashed lines, with the lighter lines indicating ancestor-descendant relationships. The number of independently evolving lineages, k, and the expected time to the next coalescent event in generations under this model, $E[T_k]$, are indicated.

The simulations I perform rely on three developments of this method - complex demography, recombination and selection.

Complex demography in coalescent simulations

A single population of constant size represents the simplest possible demographic history of a population, described by a single parameter N. In reality, processes such as population size changes are common, while non-random mating due to population structure, either geographic or otherwise, can often be described as migration between subpopulations. Coalescent models that include these properties have been described. In the case of population size changes, an alteration to the waiting time to coalescence is involved (e.g. see [431] and references therein). The case of multiple-subpopulations is marginally more complicated, as in practice it is necessary to keep track of lineages switching between the sub-populations and coalescing within them (e.g. see [434] and references therein). Programs implementing coalescent simulations, such as Hudson's ms [435], as well as MSMS [3] and Cosi2 [4] which are both used in the manuscript below, allow the specification of complex demography.

Recombination

The evolution of unlinked genes is described by different coalescent trees, as they segregate independently each generation. Conversely, when genes are fully linked, their history is described by a single tree. The situation is more complex for partially linked loci. Now, the genealogy of genes will be correlated according to the recombination distance between them. When recombination occurs within a simulated region, an appropriate coalescent model will no longer create a bifurcating tree, but an ancestral recombination graph (ARG, [429]).

Backwards in time, a recombination event will lead to the splitting of a chromosome into two lineages, each following one portion of the chromosome only. The location of the divisor point is chosen according to the relative recombination rate along the chromosome. Knowing the total probability of recombination over the entire chromosome, we can determine a waiting time distribution for such splitting events. An example of a possible ARG is shown in Fig. 4.4b.

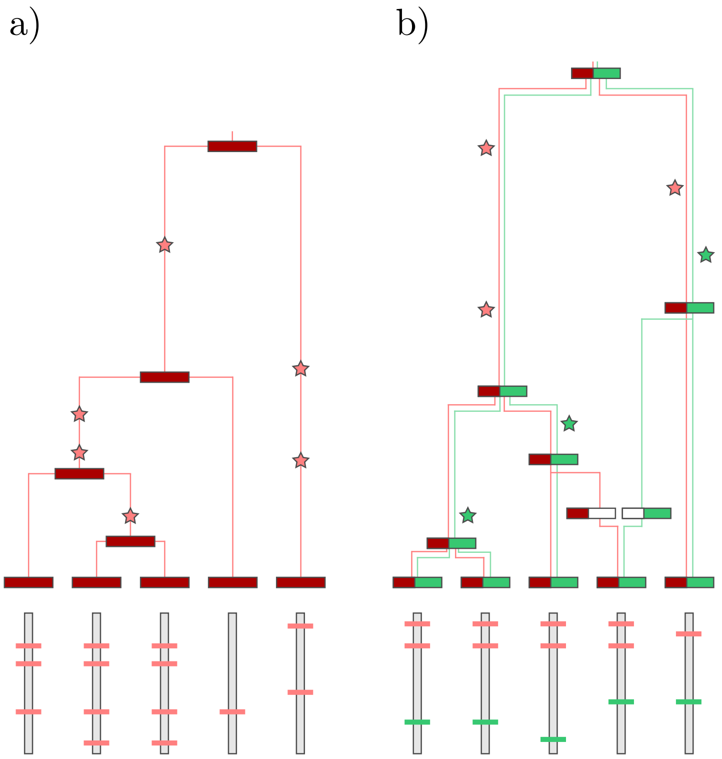


Figure 4.4: Once a coalescent tree has been simulated, mutations (stars) are scattered at random - assuming a constant mutation rate over time - along the branches. With a constant mutation rate over the chromosome, these are also places at random positions along the chromosome, and are indicated on the chromosomes (vertical grey rectangles) beneath each sample as red or green bars. The two plots correspond to a) a tree topology and genetic samples generated without recombination; and b) an ancestral recombination graph (ARG) and samples generated with recombination. The ARG in b) includes a single recombination event, whereby two sections of the chromosome in a sample evolve independently for a period of time - the two sections are coloured red and green for clarity throughout the ARG, as are mutations in the ARG and on the sampled chromosomes.

Selection

When a selective sweep is included, we can consider a population as two subpopulations defined by the presence or absence of the selected allele. Coalescence occurs within these subpopulations only. The size of each subpopulation is updated over time according to the simulated frequency trace of the selected allele. These two subpopulations are initialised by splitting the ancestral population into two, according to an allele chosen to have frequency equal to the initial frequency of the selected allele, at a time that is often a parameter of the simulation.

In the coalescent program I use, MSMS, the frequency trace is created using a forward-time structured diploid Wright-Fisher model. Briefly, with very slight notation differences from the Internal Manual of MSMS [3], we consider two alleles at a locus, a and A, with positive selection acting on an allele A. Genotypes have fitness values in the ith deme at time t of $1+s_i^{\mathrm{aa}}(t)$, $1+s_i^{\mathrm{Aa}}(t)$ and $1+s_i^{\mathrm{AA}}(t)$ for the homozygous a, heterozygous and homozygous A forms. The finite number of demes are linked by migration rates $m_{ij}(t)$ between from deme j to deme i at time t, with the proportion of non-migrants $m_{ii}(t) = 1 - \sum_j m_{ij}(t)$.

The life-cycle consists of four stages - selection, migration, mutation and then random mating - with the census taken after the random mating stage. Hardy-Weinberg equilibrium is assumed to hold.

The proportion of allele A in deme i at time t is denoted $x_i^{\mathrm{A}}(t) = \frac{n_i^{\mathrm{A}}(t)}{2N_i(t)}$, where $n_i^{\mathrm{A}}(t)$ is the number of A alleles in deme i and $N_i(t)$ the population size of deme i at time t. Selection implemented first,

$$\eta_i^{\mathrm{A}}(t_s) = x_i^{\mathrm{A}}(t)\Big(1 + \Big(1 - x_i^{\mathrm{A}}(t)\Big)s_i^{\mathrm{aA}}(t) + x_i^{\mathrm{A}}(t)s_i^{\mathrm{AA}}(t)\Big) \tag{4.10}$$

$$\eta_i^{\mathrm{a}}(t_s) = \Big(1 - x_i^{\mathrm{A}}(t)\Big)\Big(1 + x_i^{\mathrm{A}}(t)s_i^{\mathrm{aA}}(t) + \Big(1 - x_i(t)\Big)^{\mathrm{A}}(t)s_i^{\mathrm{AA}}(t)\Big), \tag{4.11}$$

where $\eta_i^{\kappa}(t_s)$ is the 'amount' of κ alleles in deme i after selection, with amount indicating a not-yet-normalised updated proportion of an allele. This is followed by migration,

$$\eta_i^{A}(t_m) = \sum_j m_{ij}(t)\,\eta_j^{\mathrm{A}}(t_s) \tag{4.12}$$

$$\eta_i^{a}(t_m) = \sum_j m_{ij}(t)\,\eta_j^{\mathrm{a}}(t_s), \tag{4.13}$$

where $\eta_i^{\kappa}(t_m)$ is the number of κ alleles in deme i after migration. Mutation then occurs, with the new relative frequency in an *infinite* population of allele A, $x_i^{\mathrm{A}}(t_M)$, given by

$$x_i^{\mathrm{A}}(t_M) = \frac{(1-v)\eta_i^{A}(t_m) + u\eta_i^{a}(t_m)}{\eta_i^{A}(t_m) + \eta_i^{a}(t_m)}, \tag{4.14}$$

where the mutation rate from allele A to a is v and the mutation rate from a to A is u. Demographic stochasticity due to a finite population size is now included through sampling of an infinite population with proportion $x_i^{\mathrm{A}}(t_M)$ of allele A using a binomial random variable. The distribution sampled is

$$\Pr\Big(n_i^{\mathrm{A}}(t{+}1)|n_i^{\mathrm{A}}(t)\Big) = \binom{2N_i(t)}{n_i^{\mathrm{A}}(t+1)}\Big(x_i^{\mathrm{A}}(t_M)\Big)^{n_i^{\mathrm{A}}(t+1)}\Big(1{-}x_i^{\mathrm{A}}(t_M)\Big)^{2N_i(t)-n_i^{\mathrm{A}}(t+1)}, \quad (4.15)$$

which is then used to calculate $x_i^A(t{+}1)$. This model is sufficient to generate a frequency trace of allele A in any demography that can be specified by MSMS.

Arbitrary frequency traces can also be given as input to either MSMS or Cosi2.

I now present a draft manuscript in which I use coalescent simulations to assess the performance and suggest improvements to selection statistics based on signatures of LD. The work is complete in the sense that I do not currently intend to conduct further simulations, but the final submitted manuscript will likely be re-formatted and cut.

4.6 Refining the use of linkage disequilibrium as a robust signature of selection *(Unpublished manuscript)*

G. S. Jacobs, T. J. Sluckin and T. Kivisild

4.6.1 Abstract

Natural selection causes distortions to patterns of linkage disequilibrium (LD) in the genomic region surrounding a selected locus. These patterns have been used to infer past selective sweeps. However, recombination rate is known to vary substantially along the genome for many species. We here investigate the effectiveness of current (Kelly's Z_{nS} and $\omega_{\max}$) and novel statistics designed to identify LD distortions due to selection in various scenarios, including a human-realistic demographic model and recombination rate variation. When recombination rate is constant, Kelly's Z_{nS} offers high power, but is out-performed by a novel statistic we test, which we call α. When recombination rate fluctuations are included, there is a considerable reduction in power for all LD-based statistics. However, this can largely be reversed by appropriately controlling for expected LD using a genetic map. To further test these different methods, we perform selection scans on well-characterised HapMap data, finding that all three statistics - $\omega_{\max}$, Kelly's Z_{nS} and α - were able to replicate signals at many selection candidates previously identified based on population differentiation or distortions to the site frequency spectrum. Modifying Kelly's Z_{nS} and α to control for expected LD increases the number of selection candidates they replicate, while the $\omega_{\max}$ statistic was the most successful method that did not control for recombination rate variation.

Keywords: Natural selection, linkage disequilibrium, human evolution, recombination rate

4.6.2 Introduction

Natural selection can create characteristic population genetic signatures focussed on genomic regions that have influenced past reproductive success. Reading this evolutionary fingerprint allows us to gain insight into the lives of long-dead individuals, while also identifying a subset of phenotypically and evolutionarily important genes. In that both evolutionary success and medical science are interested in survival, such genes may have important implications for human health. Among domestic and wild species, veterinary applications are supplemented by the potential to improve productivity or further conservation respectively.

The idea of identifying signatures of selection using linkage disequilibrium (LD) data is not a new one (e.g. [436]). However, increasing availability of genetic data and computational power offer us new tools. In the case of human genetics, these resources have fascilitated the discovery of thousands of putative selection signals [370]. However, for the vast majority of detected signals no phenotypic association has been suggested, let alone functionally tested using model organisms, and a relatively small number of signals are replicated across studies. Given the large number of hypotheses tested in genome scans for selection, many signals may be false positives [437, 370, 372]. This concern may be compounded by systematic biases related to ascertainment [438] or data quality [439]. Distinguishing true signatures of selection requires a multi-faceted approach, but improvements to statistics used to infer selection is an important step. In this work, we use simulations to explore the ability of different statistics to correctly infer selection from local patterns of LD. In particular, we assess the impact of variable recombination rate and complex demography on the power of this approach. We then assess the extent to which selection statistics based on pairwise LD can replicate previously detected selection candidates.

Distortion to LD in the genomic region surrounding a selected variant is one of many genetic effects of natural selection related to genetic hitch-hiking [381]. In the case of positive selection, a selective sweep may lead to a local increase in LD. As the selected variant reaches high frequency, this local increase in LD is followed by a reduction in LD between SNPs located on opposite sides of the variant (both patterns shown in Figure 4.5). Statistics designed to detect both the first (Kelly's Z_{nS}, [380]) and second (ω, [357]) patterns have been suggested, and the theoretical dynamics of LD under selection have been explored [440, 441, 442]. It has been argued that one of these signals, the ω statistic, is relatively robust to non-equilibrium demographic history of the target population, which can vastly reduce the performance of other statistics [364, 360].

In humans [366] and other species (e.g. [367, 368, 369]) the local recombination rate is known to be highly variable. Genetic maps, which provide estimates of the recombination rate along a genome, are created to describe these fluctuations, either through patterns of linkage disequilibrium in a random sample drawn from a population (e.g.

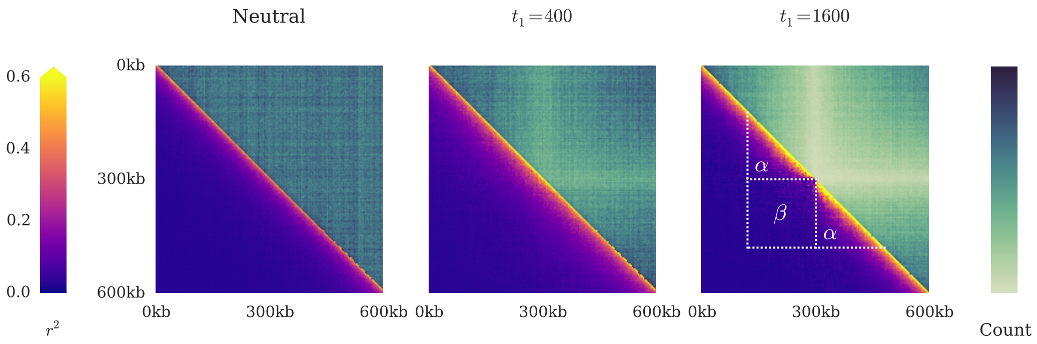


FIGURE 4.5: The average pattern of pairwise linkage disequilibrium (r^2, lower triangle of each matrix) and SNP diversity (as represented by the total number of LD measurements over all simulations at the given pairwise distance, upper triangle) created by a selective sweep, based on 2000 simulations. The human demographic model of Gravel *et al* [422] was used in the simulations (see Extended Methods, §4.6.8, for details), with 40 chromosomes sampled from the European population. For models including selection, the sweep begins at time t_1 generations pastward, using an initial selected allele frequency of 0.0005 and an additive selection model with the homozygous state corresponding to $s = 0.04$. The frequency of the selected allele in the present is, on average, approximately 0.7 and > 0.99 when selection started 400 and 1600 generations ago respectively. In the right-most plot, the Along (α) and Over (β) regions used to calculate test statistics are indicated.

[375]) or by inferring recombination events in pedigrees (e.g. [376]). As an indication of the extent of recombination rate variation, human genetic data suggests that 60% of recombination events happen in 6% of the genome [390]. The portion of the genome with high-recombination rate manifests as highly local extreme peaks in the recombination rate known as recombination hotspots. The implications of recombination rate variation on methods to detect selection based on pairwise LD has not yet been thoroughly explored.

The structure of this paper is as follows. We will first describe two well-known statistics (Kelly's Z_{nS} and ω) used to detect positive selection through patterns of LD, and discuss their advantages and potential complications in their use. We then use simulations to identify, from a barrage of possible formulations designed to address some challenges these statistics face, the most effective selection statistics. Finally, we test all methods on well–studied HapMap Phase II data [390], assessing their ability to replicate selection candidates identified based on the site frequency spectrum and/or population differentiation. We find that controlling for expected LD based on a genetic map tends to improve statistic performance in simulations invovling variable recombination rate and increases the replication of selection candidate signals.

4.6.3 Selection statistics based on LD

For the purposes of this paper, we focus on statistics that use pairwise LD to infer positive selection, as opposed to other statistics driven by hitch-hiking such as haplotype homozygosity [389]. Several methods that use pairwise LD patterns alone have been proposed - Kelly's Z_{nS} [380], Rozas' Z_A and Z_Z [443], and the ω statistic [357]. The last of these has been adapted for use in genome scans, as implemented in the program OmegaPlus [444]. A separate approach, which we discuss but do not investigate, leverages a procedure for detecting the recombination rate from population genetic data [375] through LD distortions with recombination rate estimates based on pedigree data to detect selection [445]. The focus of this paper is on Kelly's Z_{nS} and ω, and we describe these statistics now.

Kelly's Z_{nS} is simply the average pairwise LD between all SNPs over a fixed region of the genome,

$$Z_{nS} = \frac{2}{S(S-1)} \sum_{i=1}^{S-1} \sum_{j=i+1}^{S} r_{i,j}^2 \tag{4.16}$$

where S corresponds a list of polymorphic sites in the genetic region numbered $[1, ..., S_{\max}]$, i and j are indicators refering to loci in the list S, and $r_{i,j}^2$ is a standardised measure of LD corresponding to the squared correlation of allelic identity between loci i and j [379]. Visually, an example calculation of this statistic would be the average r^2 among all SNPs contained in the region x in Fig. 4.6. Given the dynamics of LD driven by a selective sweep, this statistic is expected to be most effective when a selected variant has reached a moderate to high frequency. The approach also has relatively high power to detect soft sweeps [446] in which a locus experiences recurrent beneficial mutations.

The ω statistic tries to identify a characteristic LD pattern that emerges toward the end of a selective sweep, represented by an increase in LD between SNPs downstream or upstream of a selected variant (*Along* the chromosome), but a reduction in LD between those on either side of it (*Over* the selected locus). It achieves this by taking the list of S polymorphic sites in a region and dividing this into two contiguous groups, L and R, which are located on either side of the l^{th} locus. L and R contain l and $S-l$ SNPs respectively. Given these definitions, ω is defined as

$$\omega = \frac{(\binom{l}{2} + \binom{S-l}{2})^{-1} (\sum_{i \in L, j \in L} r_{i,j}^2 + \sum_{i \in R, j \in R} r_{i,j}^2)}{(l(S-l))^{-1} \sum_{i \in L, j \in R} r_{i,j}^2}. \tag{4.17}$$

The position of l is iteratively moved along the chromosome to obtain the maximum value of ω. Referring again to Fig. 4.6, an example calculation of this statistic would be the average value of r^2 measurements indicated by blue circles divided by the average value of r^2 measurements indicated by pink circles. Large values of $\omega_{\max}$ may indicate non-neutral evolution. The pattern detected by ω is apparent in the right-most plot of Figure 4.5.

The OmegaPlus program implements a genome scan adaptation of ω. Instead of locating the divisor on SNPs, a regularly sampled grid is defined over the region to be scanned. At each point on the grid, the value of $\omega_{\max}$ is determined by varying the window size evaluated and hence the SNPs in sets L and R. In the original ω formula, the windows size was limited by the length of sequence data available. As this would imply whole-chromosome windows in a genome-scan situation, the OmegaPlus implementation of ω differs in defining constraints on window size for each ω calculation (Supplementary Information, [444]; first applied in [363]). Normalisation occurs as above, and depends on the number of LD evaluations made in each region. The implementation of the statistic is highly optimised [447], such that large amounts of genetic data can be rapidly scanned, even on consumer-grade computers. $\omega_{\max}$ has shown promise in identifying selection in simulation studies [364, 360]. This work explores the potential of LD as a selection signal in genomic scans, and we use therefore used OmegaPlus program to calculate $\omega_{\max}$ values rather than using the standard ω statistic construction. Henceforth we use the notation $\omega_{\max}^{\text{minwin,maxwin}}$ to denote an individual run of OmegaPlus with window size flags '-minwin' and '-maxwin' set as the specified number of kilobases.

4.6.3.1 Challenges faced by Kelly's $\mathrm{Z_{nS}}$ and the ω statistc

Both Kelly's Z_{nS} and ω have been used to infer natural selection (egs. [448, 449, 450]). However, each also faces challenges, related to the nature and impact of recombination. We here describe three complications - variable recombination rate, which affects both Kelly's Z_{nS} and ω, variable window size, which is only a concern for ω, and fluctuations in diversity.

Variable recombination rate

As previously noted, recombination rate shows high levels of genome-wide variation within a number of species. This clearly has significant implications for statistics searching for unusual patterns of LD to infer non-neutral evolution [445]. Recombination hotspots cause local LD to plummet, which mimics the pattern of reduced LD expected at the end of the selective sweep and can lead to large values of the ω statistic. Conversely, certain regions of the genome have low recombination rates [367, 375], and these coldspots, with correspondingly high LD, will raise Kelly's Z_{nS}. If the null distribution of ω of Kelly's Z_{nS} is based on simulations that do not accurately represent recombination rate variation, recombination hot and cold spots may lead to false positives. When the null distribution is based on empirical whole-genome data the signal of variable recombination rate will be captured, but we expect a reduction in statistical power as compared to the constant-recombination rate case due to outliers associated with recombination hotspots and coldspots.

Variable window size

Most selection scans calculate statistics based on a sliding window of fixed size. The OmegaPlus algorithm takes a different approach, calculating the statistic at static positions l on a regular grid while changing the size of the regions used to define R and L on each side of the target locus to find $\omega_{\max}$. This approach has several implications. Firstly, we note that the spatial extent of LD distortions in the genome caused by a selective sweep will depend on recombination rate, selection strength and the age of the sweep [381]. Given this, it is possible that a single scan with OmegaPlus using variable window size can identify a broader range of selective sweeps leading to different sized LD distortions than a fixed-window approach. Secondly, different sizes of R and L have different expected ω values under neutrality. A simple illustrative example is provided in Appendix 2, §4.7. The impact of R and L on expected ω value is not necessarily a problem, but does make understanding the statistics less intuitive. It may also lead the value of $\omega_{\max}$ to reflect local patterns of SNP diversity as well as LD. Thirdly, when L and R are smaller, the variance in the average value of both the numerator and denominator in Eq. (4.17) will be higher due to the reduced number of LD measurements. If the denominator randomly becomes small, Eq (4.17) diverges. It is unclear whether this effect will favour the detection of selection (local reduction in SNP diversity due to selection is important here) or just lead to greater noise in the neutral signal.

In practice, greater power is often observed for statistics that follow the same principle as ω but apply a constant window size, as shown in Appendix 1 (§4.6.9).

Fluctuations in diversity

The spatial distribution of SNPs along a region of the genome can also impact both Kelly's Z_{nS} and ω. Variation in the mutation rate [424] could potentially lead to regions with far fewer SNPs, and hence higher variance in the statistics under neutrality. This is a particular concern for the variable window sizes used in OmegaPlus (although the flag '-minsnp' allows the specification of a minimum number of SNPs in a window). Local patterns of mutation rate variation in humans have not yet been accurately mapped (though information on certain patterns of *de novo* mutation exists, e.g. [451]). However, in addition to mutation rate variation, selection can also reduce diversity [381]. Both Kelly's Z_{nS} and ω are based on average LD in a genomic region, with each LD measurement equally weighted. With greater genetic map distance from a positively selected locus the density of SNPs will increase and the degree of LD distortion will decrease. This suggests that the signal of LD distortion targeted by these statistics may disappear if the window size is too large.

The value of statistics designed to capture local patterns of LD may depend, then, on features such as the window size used and neutral variation in SNP diversity. Furthermore, variable recombination rate substantially impacts observed LD values. The implication is that some control for these phenomena may improve the ability of these statistics to detect positive selection. In this work, we focus on developing and testing LD statistics calculated over a fixed-size genomic window that control for variable recombination rate.

4.6.4 Methods

Our approach to improving LD-based statistics is a pragmatic one. In essence, we use simulations to explore a large number of possible LD-based statistics and assess their potential to detect selection by comparing their power to Kelly's Z_{nS} and $\omega_{\max}$. The details of this process are described in the Extended Methods[17] , §4.6.8.

4.6.4.1 Designing selection statistics

Briefly, we define a window of fixed length x base pairs, centred on a target locus. As in ω, we divide SNPs within this window into two sets - those that are to the left of the target locus, L, and those on the right, R, see Fig. 4.6. The ω statistic, Eq. 4.17, averages LD measurements between SNPs in the same set (the Along region) and divides this by the average LD between SNPs in different sets (the Over region). Kelly's Z_{ns} takes the average LD over all SNPs in the window. A measure of the average value of LD in the Along region is

$$\alpha = \frac{\binom{l}{2}^{-1} \sum_{i \in L, j \in L} r_{i,j}^2 + \binom{S-l}{2}^{-1} \sum_{i \in R, j \in R} r_{i,j}^2}{2} \tag{4.18}$$

and in the Over region

$$\beta = \frac{\sum_{i \in L, j \in R} r_{i,j}^2}{l(S-l)}, \tag{4.19}$$

see Fig. 4.5. Assuming the number of SNPs in L and R are similar, Kelly's Z_{nS} will be approximately the average of α and β, while ω is approximately $\frac{\alpha}{\beta}$. When designing test statistics, we take a similar approach in calculating a measure of average LD in the Over and Along regions, with the possibility of some simple operation (such as addition and

[17] Providing a brief summary of methods in the main text and then giving a more detailed description in an Extended Methods section, which appears at the end of the main text rather than as an Appendix, is relatively standard in the evolutionary genetics literature.

division, as above) then performed on these. Unlike ω and Kelly's Z_{ns}, we include the option of subtracting or otherwise controlling for the expected value of the test statistic, based on expected LD between SNPs under neutrality. In total, we test 39 different statistics, 29 of which control for expected LD.

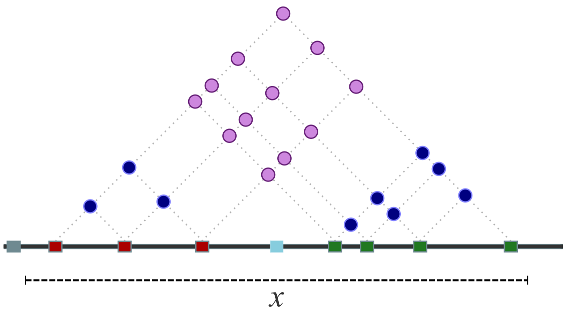


FIGURE 4.6: Schematic illustration of a test statistic calculation. The dark grey line indicates a section of the chromosome, the rectangles intermittently spaced along it SNPs, and the dots above it LD measurements between SNPs. The SNP targeted by selection is coloured light blue, while SNPs within the target region, x, that are in set L are red and those in set R green. LD measurements are coloured to indicate whether they are in the Along set α (dark blue) or Over set β (pink).

4.6.4.2 Assessing the power of statistics through coalescent simulation

We use coalescent simulations conducted with the programs MSMS [3] and Cosi2 [4] to assess the performance of different statistics. We first approximate the distribution of each statistic under neutrality by simulating 1000 samples of $n = [20; 40; 80]$ chromosomes for each demographic and recombination model. For each sample, we evaluate the statistic at each SNP within a 200kb window, with the maximum or minimum value in the window used as the test statistic value. The 1000 values thus generated describe the null distributions for a test statistic. To approximate the distribution of the test statistics given a positive selective sweep, we simulate at least 300 samples with positive selection ($s = [0.01; 0.02; 0.04; 0.08]$) acting on a single SNP located in the middle of the chromosome, using various selection scenarios (e.g. final selected allele frequency). The distribution of each test statistic under neutrality and given positive selection is then used to calculate power and receiver operating characteristic (ROC) curves [452] over several demographic, recombination and selection scenarios.

We use two demographic scenarios, one with constant population size of $N_e = 10^4$ and the other following an Out of Africa (OOA) model suggested by Gravel *et al* [422], with samples taken from the European population. Details can be found in the Extended Methods, §4.6.8. The recombination rate is either constant (both models) or variable (constant population size model only). For the variable recombination rate model, the rate is sampled from the HapMap human Chromosome 2 (b36) genetic map, as estimated based on HapMap Phase II populations ([390]), excluding regions close to the centromere.

When a test statistic requires the calculation of expected LD given genetic map distance, we generate an 'LD profile' describing the expected properties of LD at a given genetic map distance, such as the average or standard deviation of r^2. We approximate the LD profile by simulating approximately $3 * 10^9$bp of neutral genetic data according to the appropriate demographic scenario and sample size, and then binning LD measurements according to their genetic map distance. For all simulations involving a variable recombination rate, we consider three situations regarding the availability of genetic map data. We either assume that no genetic map is available, in which case physical distance serves as a proxy for genetic map distance (PhysMap), or that the true genetic map is available (TrueMap), or that a lower resolution genetic map is available (LowResMap). An appropriate LD profile is used in each case - for example, if we are calculating statistic that attempt to control for expected LD using the physical map as a proxy for the genetic map (PhysMap), then the LD profile used to calculate expected LD would also be generated using the PhysMap model.

The internal algorithm of MSMS involves conditioning a coalescent model on a stochastically generated selected allele frequency trajectory. The allele frequency trajectory is created using the selection coefficient s, the demographic model, and several of four possible parameters describing features of the selection scenario - the time at which the selection phase begins (pastward), t_0 generations, the time at which selection stops, t_1, and the frequency of the selected allele at these times, q_{t_0} and q_{t_1}. To compare summary statistics, we define three selection scenario categories - Low Frequency, where $q_{t_0=0} = [0.3; 0.5; 0.7]$, High Frequency, with $q_{t_0=0} = [0.9; 0.99]$ or $q_{t_0=800} = [0.99]$, and selOOA, denoting the Out of Africa demographic scenarios. Performance is assessed using the power of test statistics (e.g. Tables 4.1 and 4.2, and Fig. 4.8) or the partial area under the ROC curve (pAUC, [453]) between a false positive rate (FPR) of 0 and 0.05, Fig. 4.7. Using the pAUC gives greater emphasis on performance when the FPR < 0.05, which is relevant for genome-wide selection scans. For each selection category, we determine the best pAUC for each included selection scenario (i.e. selection strength and final allele frequency), and average these, giving an indication of the maximum potential performance of each statistic. Again, further details on selection and the assessment of statistic performance are provided in the Extended Methods, §4.6.8.

4.6.4.3 Replication of previously suggested selection candidates

We compare the performance of our test statistics with Kelly's Z_{nS} and $\omega_{\max}$. Several statistics, particularly the α statistic, Eq. 4.18, and Kelly's Z_{nS} when these control for expected LD, have high power given a variable recombination rate. To help further assess the utility of different statistics, we present selection scan results using HapMap Phase II (NCBI b36, [390]) data for human Chromosome 2 (CEU and CHB+JPT) and Chromosome 15 (CEU). Recombination rate is controlled for using either the HapMap genetic map (derived from LD patterns, [390]) and the deCODE genetic map (derived from inferred recombination events in a large Icelandic cohort, [376]). Now performance is assessed based on the ability of the various statistics to replicate selection signals previously identified based on the site frequency spectrum and/or population differentiation [358, 454, 455, 361, 456, 457, 458, 6], genetic features that should be relatively independent of recombination rate variation under neutrality (though see e.g. [365]). High performance is indicated by enrichment of signals identified by our LD-based test statistics in these candidate regions.

4.6.5 Results

Our simulations identify certain statistics (Kelly's Z_{nS} and α) as particularly powerful. They also suggest that controlling for expected LD increases the power of these statistics when recombination rate is variable, but marginally reduces power when recombination rate is constant. We also found that controlling for expected LD increased the number of previously suggested selection candidates that these statistics replicated in HapMap Phase II SNP data. Although the interpretation of signal replication is not trivial (signals may be false positives or false negatives), the overall impression is that controlling for expected LD increases the performance of certain LD-based selection statistics.

4.6.5.1 Controlling for expected LD increases simulated power of statistics when recombination rate is variable

As expected, our simulations exploring the impact of human-realistic recombination rate variation supported the use of a genetic map in the design of LD-based selection statistics. In total, we tested 29 methods that incorporated some form of control for expected linkage and 10 methods that did not. Comparing these as groups of similar statistics - for example, all those that divide average LD in the Along region by that in the Over region (like ω), or all those that add average LD in these two regions (like Kelly's Z_{nS}) - the average improvement over methods that did not consider genetic map data was 79% (in absolute terms, 0.22) by our pAUC metric, assuming a constant population size demographic model and variable recombination rate (see Figure 4.7). Focussing on the case of $n = 40$, window size $x = 200$kb and $s = 0.01$, this reflects an average increase in power at 1% FPR of 120% (0.07) for the Low Frequency scenarios and 90% (0.08) for the High Freqency scenarios.

When controlling for variable recombination rate, we found that both the full genetic map (TrueMap) and the lower resolution genetic map (LowResMap) yielded similar results, with a performance reduction according to the pAUC metric of just 10% and 4% for the lower resolution map given Low Frequency and High Frequency scenarios respectively. However, trying to control for expected LD based on physical distance

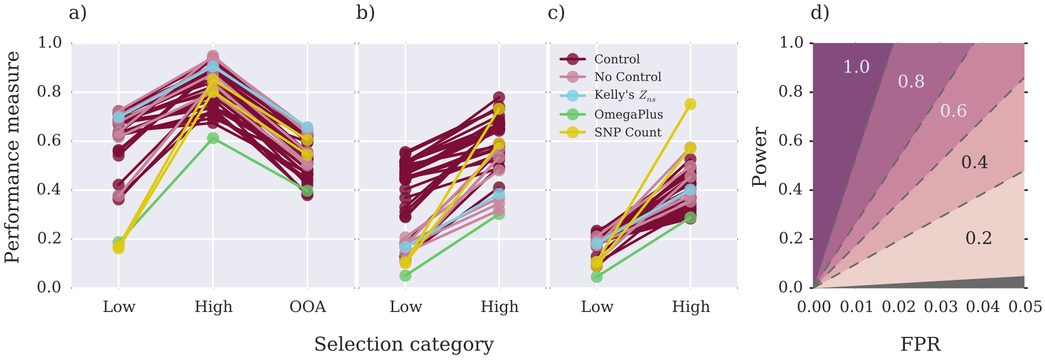
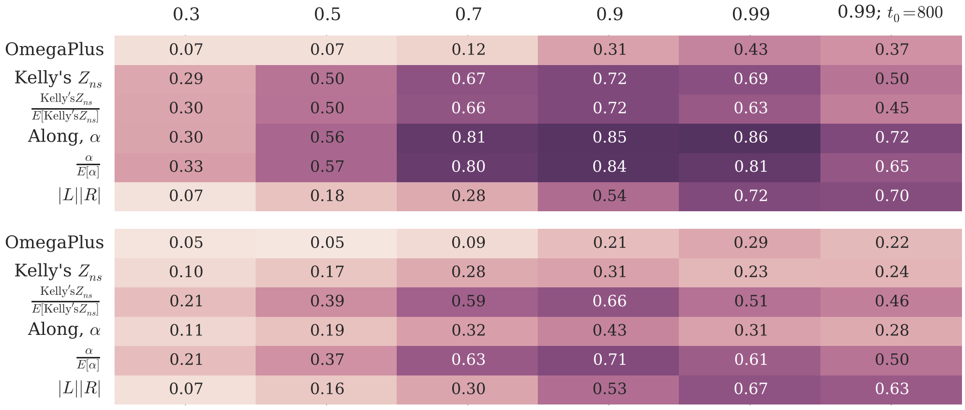


FIGURE 4.7: Performance of different categories of selection statistic tested according to our pAUC metric, under a) a constant recombination rate, b) a variable recombination rate and the true genetic map and c) a variable recombination rate using the physical map as an approximation of the genetic map. The statistic categories are indicated in the key, corresponding, in order, to those that control for expected LD, those that do not, Kelly's Z_{nS} statistic, the $\omega_{\max}$ statistic with various window sizes, and methods based on SNP diversity. The relationship between our pAUC metric and power is shown in d).

yielded poor results, and those statistics that did not take variable recombination rate into account were of equal or better performance (see Figure 4.7c).



	0.3	0.5	0.7	0.9	0.99	0.99; $t_0=800$
OmegaPlus	0.07	0.07	0.12	0.31	0.43	0.37
Kelly's Z_{ns}	0.29	0.50	0.67	0.72	0.69	0.50
$\frac{\text{Kelly's}Z_{ns}}{E[\text{Kelly's}Z_{ns}]}$	0.30	0.50	0.66	0.72	0.63	0.45
Along, α	0.30	0.56	0.81	0.85	0.86	0.72
$\frac{\alpha}{E[\alpha]}$	0.33	0.57	0.80	0.84	0.81	0.65
$\lvert L\rvert\lvert R\rvert$	0.07	0.18	0.28	0.54	0.72	0.70
OmegaPlus	0.05	0.05	0.09	0.21	0.29	0.22
Kelly's Z_{ns}	0.10	0.17	0.28	0.31	0.23	0.24
$\frac{\text{Kelly's}Z_{ns}}{E[\text{Kelly's}Z_{ns}]}$	0.21	0.39	0.59	0.66	0.51	0.46
Along, α	0.11	0.19	0.32	0.43	0.31	0.28
$\frac{\alpha}{E[\alpha]}$	0.21	0.37	0.63	0.71	0.61	0.50
$\lvert L\rvert\lvert R\rvert$	0.07	0.16	0.30	0.53	0.67	0.63

TABLE 4.1: The power (FPR = 0.05) of several test statistics given a constant population size ($N_e = 10^4$) demographic model and selection ($s = 0.01$) to various selected allele frequencies q_{t_0}. Unless otherwise stated, $t_0 = 0$, corresponding to the time of sampling in generations. The upper table indicates power given a constant recombination rate and the lower table power given recombination rate variation, with recombination rate sampled from a HapMap genetic map, see Extended Methods (§4.6.8).

The power to detect selection at different stages of a selective sweep for some representative statistics is shown in Tables 4.1 and 4.2. Again, an increase in power is associated with controlling for expected LD given genetic map distance when recombination rate is variable (Table 4.1). We also note the generally high performance of α as a test statistic, as compared to Kelly's Z_{nS} and especially $\omega_{\max}$ when recombination rate is not variable (Table 4.1) and the slower decay of reduced SNP diversity as a signal of selection than LD distortions (the $|L||R|$ statistic, describing the number of LD measurements in the

Over region). Note that our estimate of the power of $|L||R|$ is conservative as we did not calculate test statistics on simulated data with few SNPs. This avoids high variance in LD-based statistics. Nevertheless, our underestimation of power for $|L||R|$ was usually low (less than a 0.01 reduction in power) and always under 0.07 for the data presented in Tables 4.1 and 4.2.

	400	800	1600	2400	3200	4000
OmegaPlus	0.05	0.07	0.35	0.35	0.18	0.14
Kelly's Z_{ns}	0.06	0.17	0.55	0.52	0.29	0.13
$\frac{\text{Kelly's}Z_{ns}}{E[\text{Kelly's}Z_{ns}]}$	0.06	0.18	0.47	0.57	0.32	0.19
Along, α	0.06	0.19	0.63	0.58	0.32	0.17
$\frac{\alpha}{E[\alpha]}$	0.06	0.20	0.58	0.52	0.29	0.18
$\lvert L\rvert\lvert R\rvert$	0.05	0.16	0.62	0.63	0.51	0.45

TABLE 4.2: The power (FPR = 0.05) of several test statistics given an Out of Africa demographic model and selection $s = 0.01$ beginning on an allele at starting frequency 0.0005 at the time indicated, t_1 in generations. The average frequency of the selected allele at sampling time was approximately 3%, 40%, 93% for $t_1 = 400$, 800 and 1600, with $q_0 > 0.99$ when $t_1 \geq 2400$.

Identifying the 'best' statistic from our set was not simple; full results included in Appendix 1, §4.6.9 show that many approaches to controlling for expected LD had similar performance. Statistics based on the average LD in the Along region, like α, tended to be particularly successful, and we therefore focus on these. In Tables 4.1 and 4.2 we controlled for the expected statistic value given genetic map distance by dividing the observed α by that expected given the SNP distribution under neutrality. While this is relatively intuitive, there are other possible approaches to controlling for expected LD, several of which are shown in Figure 4.8.

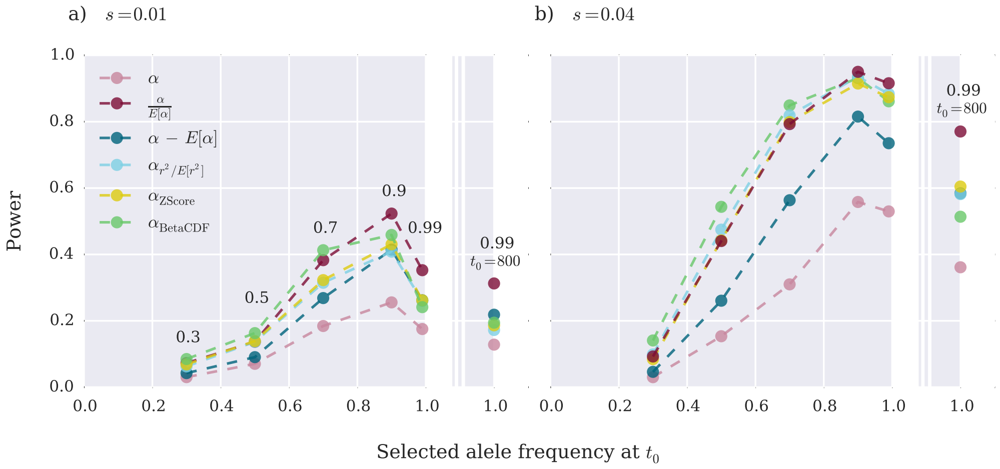


FIGURE 4.8: Power (FPR = 0.01) of the test statistic α and several methods that also focus on LD in the Along region but control for expected LD, see Extended Methods (§4.6.8). A constant population size demographic model ($N_e = 10^4$) was used with variable recombination rate and a) weaker, $s = 0.01$ or b) stronger, $s = 0.04$ selection. Unless otherwise indicated $t_0 = 0$.

Overall, the $\frac{\alpha}{E[\alpha]}$ statistic is simple to calculate and generally performed as well or better than other LD-based statistics explored given variable recombination rate.

4.6.5.2 Controlling for expected LD increases selection candidate signal replication in HapMap data

To further assess this and other selection statistics, we performed selection scans using publicly available HapMap genotype data [390]. We chose to focus on Chromosomes 2 (European descent and Asian populations, CEU and CHB+JPT) and 15 (CEU) as these chromosomes/populations include well-characterised selection signals (*MCM6/LCT*, *SLC24A5*, *HERC2* and *EDAR*).

Selection scan results using OmegaPlus ($\omega_{\text{max}}^{50,400}$), Kelly's Z_{nS}, α and $|L||R|$, as well as LD-controlled variants of Kelly's Z_{nS} and α, are shown in Figure 4.9. For each chromosome, candidate selection signals suggested by other studies which did not use statistics based on LD or haplotype patterns (see Extended Methods, §4.6.8) are indicated. Table 4.3 shows the average value of each assessed selection statistic in the 200kb genomic windows containing these selection signals, as well as the average rank of selection candidate windows among all windows and an indication of how unusual the observed findings are. Results for scans using several other LD-based statistics that we investigated are included in Appendix 4(§4.7.2), along with scans using the deCode [376] rather than HapMap genetic map to control for expected LD. While the main purpose of these scans was to assess the various statistics, we did find several novel peaks. As we are not aware of selection statistics based on pairwise LD being applied to this data (though see [445]), we tabulate the strongest signals in Appendix 5, §4.7.3.

Interpreting the selection scans shown in Figure 4.9 in not simple. Our metrics show no signal at many selection candidates, and in some cases controlling for expected LD actually leads to a reduction in signal strength. This in itself is not concerning, in that many suggested selection targets may be false positives. For example, of the 65 200kb selection candidate windows we identified from a range of studies on European Chromosome 2, some using similar statistical methods, only 5 of these were found in multiple studies. Nevertheless, it is clear that controlling for expected LD does not always improve the signal even for well-accepted candidates. While the signal at *SLC24A5* (CEU) increases, that of *MCM6/LCT* (CEU) appears to fall. This was apparent for both Kelly's Z_{nS} and α, which, unsurprisingly, are highly correlated. Interestingly, the ω_{max} statistic actually displayed an unusually low value for some selected regions (such as *LCT* in CEU) rather than the high value that is generally expected, using various window sizes (Appendix 4, §4.7.2). This is presumably because the selective sweep is recent and incomplete [459, 401], and the reduction in LD in the Over region that drives high ω_{max} values only appears in later stages of the sweep [440]. A similar pattern

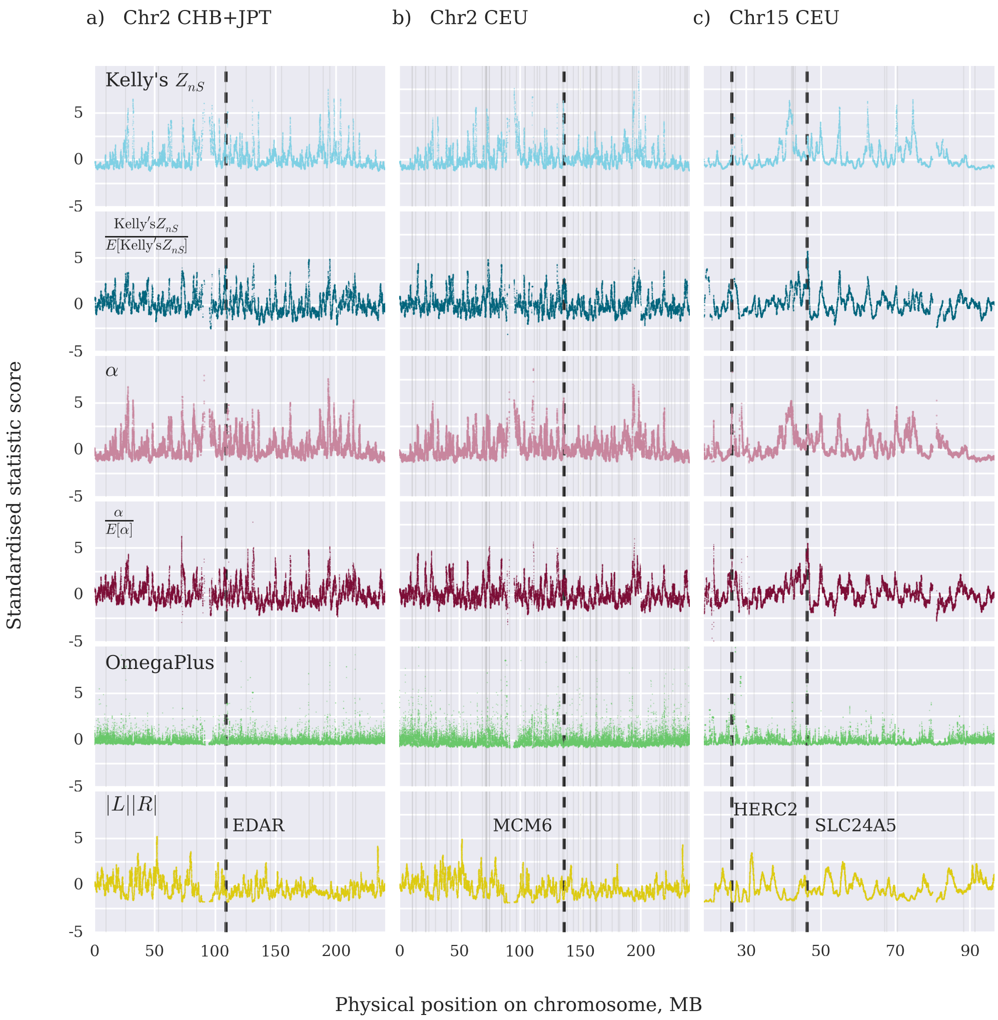
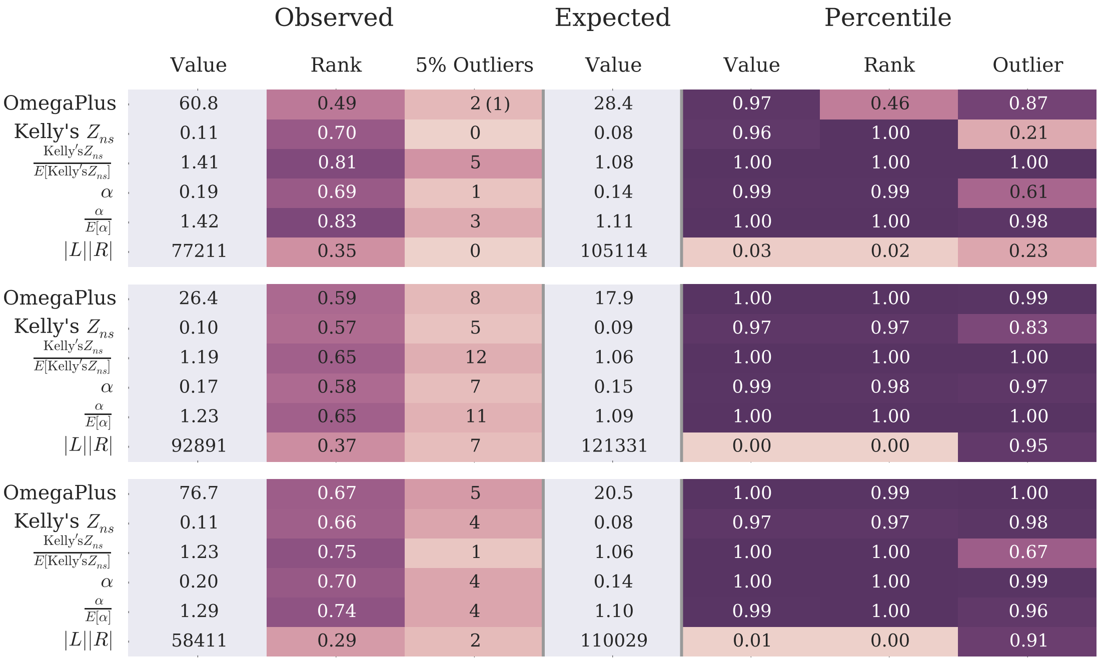


Figure 4.9: Selection scans (standardised to aid comparison) on HapMap Phase II data, using a range of LD-based statistics and a diversity measure. From top to bottom, the scans represent Kelly's Z_{nS}, Kelly's Z_{nS} controlled for expected LD, α, α controlled for expected LD, the OmegaPlus program calculating $\omega_{\max}^{50,400}$, and the diversity measure $|L||R|$. Excluding OmegaPlus, all statistics were calculated with statistic window size 400kb. The HapMap combined genetic map was used to estimate expected LD. Thick dashed lines indicate the genomic targets of four relatively well-established signatures of selection, while the light grey lines indicate signals found in a range of studies based on population differentiation and the site frequency spectrum, see Extended Methods, §4.6.8, and Appendix 3, §4.7.1

was sometimes observed in our simulations, and was found to be strongly influenced by choice of window-size parameters.

Greater detail on the performance of the various statistics is given in Table 4.3. Before discussing these results, it is useful to note that, in addition to the problem of false positives, many statistics used to scan for positive selection have greatest power when selection is stronger and the selected locus has swept to high frequency. This regime overlaps the optimal power for LD-based statistics. The implication is that many true signals which our statistics might find will have already been detected. We can, however,



	Observed			Expected	Percentile		
	Value	Rank	5% Outliers	Value	Value	Rank	Outlier
OmegaPlus	60.8	0.49	2 (1)	28.4	0.97	0.46	0.87
Kelly's Z_{ns}	0.11	0.70	0	0.08	0.96	1.00	0.21
$\frac{\text{Kelly's} Z_{ns}}{E[\text{Kelly's} Z_{ns}]}$	1.41	0.81	5	1.08	1.00	1.00	1.00
α	0.19	0.69	1	0.14	0.99	0.99	0.61
$\frac{\alpha}{E[\alpha]}$	1.42	0.83	3	1.11	1.00	1.00	0.98
$\|L\|\|R\|$	77211	0.35	0	105114	0.03	0.02	0.23
OmegaPlus	26.4	0.59	8	17.9	1.00	1.00	0.99
Kelly's Z_{ns}	0.10	0.57	5	0.09	0.97	0.97	0.83
$\frac{\text{Kelly's} Z_{ns}}{E[\text{Kelly's} Z_{ns}]}$	1.19	0.65	12	1.06	1.00	1.00	1.00
α	0.17	0.58	7	0.15	0.99	0.98	0.97
$\frac{\alpha}{E[\alpha]}$	1.23	0.65	11	1.09	1.00	1.00	1.00
$\|L\|\|R\|$	92891	0.37	7	121331	0.00	0.00	0.95
OmegaPlus	76.7	0.67	5	20.5	1.00	0.99	1.00
Kelly's Z_{ns}	0.11	0.66	4	0.08	0.97	0.97	0.98
$\frac{\text{Kelly's} Z_{ns}}{E[\text{Kelly's} Z_{ns}]}$	1.23	0.75	1	1.06	1.00	1.00	0.67
α	0.20	0.70	4	0.14	1.00	1.00	0.99
$\frac{\alpha}{E[\alpha]}$	1.29	0.74	4	1.10	0.99	1.00	0.96
$\|L\|\|R\|$	58411	0.29	2	110029	0.01	0.00	0.91

Table 4.3: The performance of test statistics in replicating a signal in candidate selected windows. From top to bottom, the tables correspond to CHB+JPT (Chr2), CEU (Chr2) and CEU (Chr15). For these population/chromosome combinations, we identified (and could test using the HapMap data) 17(15), 65(64) and 17(17) previously reported selection candidates respectively. OmegaPlus calculated $\omega_{\text{max}}^{50,400}$ across a grid rather than at SNPs, and was able to test all selection candidates (17, 65, 17); for the CHB+JPT (Chr2) data, one of the replicated candidates was not tested by the other statistics. The test statistic for a 200kb window corresponds to the maximum value of the indicated statistic in that window, or the minimum value for the diversity metric $|L||R|$. The average window test statistic across all selection candidate windows is reported (observed value), and can be compared to the average across all 200kb windows (expected value). The average rank of selection candidate windows is also reported (observed rank), as well as the number of selection candidate windows in that are top-5% outliers (or bottom 5% for $|L||R|$). A re-sampling approach yielded percentiles of the three assessment metrics as compared to their chromosome-wide distribution.

gain some insight into the problem by assuming that the density of such sweeps is broadly consistent across chromosomes, and the demographic history of Europeans and East Asians to be sufficiently alike to yield a similar efficiency of selective processes. Based on our candidate selection set, the proportion of Chromosome 2 within 200kb candidate windows is 5.4% and 1.4% for Europeans and Asians respectively, while for Chromosome 15 the figure is 4.4% (Europeans). Thus, our general heuristics for a successful statistic are

- A relatively low number of novel candidates, such that proportionately more top signals replicate previously suggested candidate windows.

Population	Chr	MB	Signals	Source	Genes	Target
CHB+JPT	2	72-73	2,3,4,5,6	[456, 458]	CYP26B1,EXOC6B,SPR,EMX1	
CEU	2	74.2-75	2,4	[358, 455, 456]	30 genes	
CEU	2	84.2-85	2,6	[358, 455, 6]	FUNDC2P2,SUCLG1,DNAH6, TRABD2A,TMSB10	
CEU	2	87.2-87.8	5	[458]	5 non-coding	
CHB+JPT	2	108.4-108.8	2	[6]	GCC2,LIMS1,RANBP2,CCDC138	EDAR
CEU/ CHB+JPT	2	121.2-121.4/ 121.4-121.6	6/ 6	[456]/[456]	GLI2	
CEU	2	135.8-136.2	1,3	[358]	ZRANB3,R3HDM1,MIR128-1	MCM6
CHB+JPT	2	177-177.8	2,4,6	[454, 361, 456]	HNRNPA3 and 5 non-coding	
CEU	2	182.2-182.4	6	[6]	CERKL,NEUROD1	
CEU/ CHB+JPT	2	194.2-195/ 194.4-195	0,1,2,3/ 1,2,3,5	[6]/[454, 361, 456]	LOC101927406	
CEU	15	26.2-27	3,4,5,6	*	12 genes, including HERC2	HERC2
CEU	15	41.4-42.2	1,3	[457, 6]	23 genes	
CEU	15	46-46.6	2,4,6	[456, 458]	SLC24A5,MYEF2,CTXN2, SLC12A1,DUT,FBN1	SLC24A5
CEU	15	88.2-88.4	6	[458]	AP3S2,C15orf38-AP3S2,ARPIN, ZNF710,MIR3174	

Table 4.4: Table showing replication of candidate selection regions at the more stringent $p < 0.01$ level. Selection statistic key: $1 =$ Kelly's Z_{nS}, $2 = \frac{\text{Kelly's}Z_{nS}}{E[\text{Kelly's}Z_{nS}]}$, $3 = \alpha$, $4 = \frac{\alpha}{E[\alpha]}$, $5 = |L||R|$, $6 = \omega_{\max}^{50,400}$. Genes are listed in full with resampled p-values in the SOM (Appendices 3 and 4). *Several studies identified selection candidates marginally downstream or upsteram of this signal [457, 455, 456]

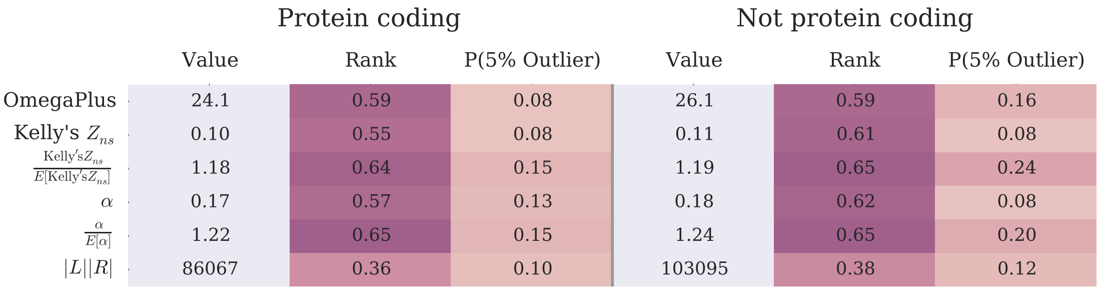
- Greatest ability to replicate signals in Chromosome 2 (CHB+JPT), then Chromosome 15 (CEU), then Chromosome 2 (CEU).

Over all selection candidate windows, Kelly's Z_{nS} replicated 9/96 signals as 5% outliers, the same as $|L||R|$ but less than α (12) and $\omega_{\max}^{50,400}$ (15/101). The estimate for $|L||R|$ is slightly pessimistic, as our pipeline removed windows with very few LD measurements, including two selection candidate windows (both in Chr 2, one Asian and the other European) that would otherwise have yielded positive results. Controlling for expected LD considerably improved the performance of both Kelly's Z_{nS} and α, with both replicating 18/96 signals. Interestingly, the window values and ranks for Kelly's Z_{nS} and α exactly follow our predictions, being highest in CHB+JPT (Chr2), then CEU (Chr15), then CEU (Chr2). This pattern is clearer when LD was controlled for, but is not apparent for $\omega_{\max}^{50,400}$ or $|L||R|$. In general, all statistics replicated considerably more signals than expected by chance, as strongly suggested by the re-sampling results (Table 4.3).

The purpose of this study has been to assess certain LD-based selection statistics rather than to search for selection signatures as such. We therefore tabulate novel hits in the Supplementary Material only. Often, several outlier windows occurred in succession, such that it is difficult to identify specific genes, variants or features that might be driving a signal. Different approaches using pairwise LD [2], other population genetic patterns (e.g. DIND, [460, 6], DDAF [461, 458]), or, especially, biological information on the impact of variants can simplify this task. Certain novel 200kb windows contained a single or few genes, such as signals overlapping *MKRN3* and *ARHGAP11B* (Europeans), and *ABCA12* (Asians). *ARHGAP11B* is a human-lineage duplication that has recently been found to influence neocortex size [462], and its low variation supports a recent origin and selection. Mutations in *MKRN3* can impact the onset of puberty [463], and variants in this region have been associated with age at menarche [464]. Finally, *ABCA12* has been previously identified as a possible selection candidate based on population differentiation, possibly related to adaptation of the skin in response to ultraviolet light exposure [458]. We did not formally class this as a replication as our 200kb window does not include the previously highlighted variant.

4.6.5.3 Selection candidate signal replication is not increased for candidate windows containing protein-coding genes

A sufficient number of selection candidates have been suggested for Chromosome 2 that we can divide windows into those that contain protein coding genes (39) and those that do not (26). The naïve expectation is that relatively more real selection events will have occurred in windows containing protein-coding genes. Results for this analysis are shown in Table 4.5.



	Protein coding			Not protein coding		
	Value	Rank	P(5% Outlier)	Value	Rank	P(5% Outlier)
OmegaPlus	24.1	0.59	0.08	26.1	0.59	0.16
Kelly's Z_{ns}	0.10	0.55	0.08	0.11	0.61	0.08
$\frac{\mathrm{Kelly's}Z_{ns}}{E[\mathrm{Kelly's}Z_{ns}]}$	1.18	0.64	0.15	1.19	0.65	0.24
α	0.17	0.57	0.13	0.18	0.62	0.08
$\frac{\alpha}{E[\alpha]}$	1.22	0.65	0.15	1.24	0.65	0.20
$\|L\|\|R\|$	86067	0.36	0.10	103095	0.38	0.12

Table 4.5: Average statistic value and rank for selection candidate windows that either do or do not contain protein-coding genes. As there were different number of windows in each set, the probability of a selection candidate being a 5% outlier rather than the number of 5% outliers is shown.

The range of known biological mechanisms through which genetic variation in a window might impact the phenotype is greater for those windows that contain protein coding genes as compared to those that do not. We therefore expected the test statistics to show values more indicative of selection in candidate windows containing protein coding genes. Testing this hypothesis on CEU Chromosome 2 data did not generally support the idea (Table 4.5), and indeed for some of the statistics the converse pattern was observed. An exception was the diversity measure $|L||R|$, with candidate windows containing protein coding genes containing fewer SNPs, potentially reflecting the longer-term signal of repeated selective sweeps or of purifying selection. We note that recent work has highlighted selection on regulation as especially important in recent human evolution [465], such that it is not clear that non-genic candidate windows are more likely to be false positives or have been subjected to weaker selection.

4.6.6 Discussion

Our results suggest that incorporating information from the genetic map into calculations of selection statistics based on pairwise LD can substantially improve their performance, both in simulations and in replicating previously identified selection candidates. We have also suggested a modification of Kelly's Z_{nS}, α, that has higher power to detect the LD distortions caused by positive selection in simulations. Although the ω statistic implemented in OmegaPlus performed relatively poorly in simulations, it was effective in replicating selection candidate signals. Indeed, of the LD-based statistics that we tested which did not incorporate information from the genetic map, this statistic identified most previously suggested selection candidates at the 5% significance level, and most overall at the 1% significance level (Table 4.4).

The partial contradiction between power as suggested by simulations and replication of signals in real data deserves consideration, as simulations are often used to justify the use of specific selection statistics. There are several possible confounding factors, and we begin by discussing our simulation modelling before turning to detailed implications of selection candidate replication. We finally discuss the difference between results obtained using a genetic map estimated using observed LD in HapMap data [390] and one derived from inferred recombination events in a large Icelandic dataset [376].

The use of coalescent simulations to represent complex demographic scenarios (e.g. [466]) and selection [467] is well established. We used two coalescent programs, MSMS [3], which has been widely used and cited, and Cosi2 [4], a development of the well-known Cosi program [466]. In both cases, selection is incorporated by dividing the population according to allelic state at the selected site [468] and conditioning the coalescent process on an independently generated allele frequency trajectory describing the size of the two subpopulations over time. Simulations according to this approach yield qualitatively expected patterns of LD [357, 440] and reduced diversity (see Fig. 4.5).

Although the detailed implementation of selection in coalescent simulations can lead to complications that are not immediately obvious [469], aspects of our simulations - an additive model of dominance, for example, and the use of stochastic frequency trajectories [470] - bypass some of these. The programs we use are based on Kingman's coalescent [355], such that coalescent events correspond to bifurcations in the genealogical tree (or ancestral recombination graph). When selection is strong or the sample size large, the probability of multiple mergers in a single coalescent event can no longer safely be ignored and other coalescent models may be more appropriate [469]. The maximum selection strength we apply is $s = 0.08$ with a sample size of $n = 80$, and while multiple mergers are certainly possible we do not consider a qualitative distortion to our power analysis likely. This problem can also arise under neutrality, but standard models of genetic change are well approximated by the Kingman coalescent even when n is quite large [471].

We therefore assume that genealogies under both neutrality and positive selection are approximated with reasonable accuracy by the coalescent simulation programs we employed. The question to ask of our simulation results, then, is whether the correct selection, demographic and recombination scenarios were modelled. In the case of recombination, the variable rate was sampled from a genetic map [390] estimated using the HapMap data we investigate, which in turn appears to be broadly consistent with other information on recombination rate variation (e.g. [376, 472]). The demographic model we apply was estimated based on the joint site frequency spectrum using low-coverage data from the 1000 Genomes project [422]. This method involves fitting the site frequency spectrum calculated using a diffusion approximation of a Wright-Fisher population that incorporates drift, selection and migration to observed data [421]. Kingman's coalescent can accurately approximate genealogies generated by a Wright-Fisher model [355, 471]. As such, even though any demographic model is a vastly idealised version of a population's history, the parametrisation will capture qualities of human genetic data that in turn should be recapitulated in simulated genealogies. The model of human demography used is broadly consistent with understanding of the Out of Africa dispersals, and is similar, in terms of divergence times and population size estimates, to a model estimated based on haplotype sharing [473].

Our choice of selection scenario is more constrained, in that we only modelled hard selective sweeps. 'Soft' sweeps, in which the ultimate fixation of an allele caused by selection involves multiple copies of that allele [474], often due to selection on standing variation rather than *de novo* mutation, can lead to different patterns of genetic variation [475]. Furthermore, a range of other selective phenomena (purifying selection, balancing selection and selection on epistatically interacting loci, for example) also lead to characteristic distortions to genetic diversity. Some of these can create signatures that resemble positive selective sweeps (e.g. see [476] and references therein). Balancing selection in particular can lead to a signal in Kelly's Z_{nS} [380], as can soft sweeps [446], while selection on loci with epistatic interactions can also impact LD patterns [477]. Finally, the frequency at which selective sweeps actually occurred in recent human evolution remains a subject of debate [478, 465]. The overall impression is that our simulations of selection may only correspond well to the evolutionary history of subset of selection signals, or, equivalently, that some selection candidates do not correspond to hard selective sweeps. Nevertheless, characterising the performance of selection statistics is an iterative process, with the range of possible selection and demographic scenarios infinite and our understanding of which are important in real data imperfect. An indication of the utility of genetic maps in controlling for variable recombination rate in simple selection scenarios provides a useful baseline for further work.

While simulation of selection offers an idealised representation of the evolutionary process, real data is complex. Replication of a selection candidate may be random, or be a true positive for both statistics involved, or reflect similar causes of false positives in both signals. Assessing the proportion of shared true and false positives depends on the power of different statistics, their correlation, and our underlying model of frequent and strong selective sweeps are. Sorting individual replicated signals into true and false positives is even harder. The critical property here is the extent to which selection statistics are correlated under neutrality versus their correlation under selection. For replication to provide evidence of true positives, the former should be minimal and the latter high. There is a danger that substantial correlation between statistics under neutrality can create the impression of a robust selection signature even when no selection occurred. We tried to avoid this problem by ignoring selection candidates suggested by approaches based on other measures of allelic associations such as haplotype homozygosity statistics. Nevertheless, characterising the detailed correlation between different statistics under neutrality, and the impact of complex demography, variable recombination and variable mutation rate on this, is an important step toward correctly interpreting replication.

This general point is highly relevant to the signal overlap between the LD-based statistics we tested, for example, which correlate strongly (Fig. 4.9), but are not especially informative concerning the unexpectedly high number of selection candidates replicated by OmegaPlus. Part of the pattern may be related to the $\omega_{\max}$ statistic searching for a signal of positive selection most apparent at the end of a selective sweep. This coincides with the stage at which population differentiation and distortions to site frequency spectrum due to selection are strongest. The power of OmegaPlus peaked at fixation in our simulations (e.g. Table 4.1), but was nevertheless lower than that of the other LD-based statistics. Ultimately, further work will be needed to precisely clarify the relationship between signal replication and power as assessed through simulations.

When controlling for expected LD given genetic map distance, we used two different genetic maps. We focus on results using the combined HapMap genetic map [390] in the main text, which is based on LD patterns in Europeans, Asians and Africans. As such, there is a danger of underestimating the recombination rate in regions with high LD due to natural selection, and of incorrectly inferring hotspots when LD is low over a high-frequency selected site. Where recombination rate is underestimated, expected LD is correspondingly over-estimated, and controlling for recombination rate is thus likely to degrade the signal of selection. Despite this, we found that results were not substantially different when using the deCODE genetic map [376]. This is interesting, in that differences between the genetic map estimated from LD patterns and that based on observed recombination events have been suggested as indicative of selection [445]. We speculate that the method of combining recombination rate estimates from multiple populations used in the HapMap genetic map substantially mitigates this effect. This is because LD-based methods tend to identify signatures of recent selection, which post-date population divergence and will affect different genomic regions in the different populations.

We finally note that our method of controlling for expected LD is, in many ways, the simplest approach, and that more complex alternatives may improve the power of these statistics further. For example, widely separated alleles that are always found on the same haplotype show an r^2 value of 1, but such distant co-segregating alleles are far less expected under neutrality if the haplotype is at high-frequency (the signal exploited by methods derived from EHH [389]). Incorporating information about the derived allele frequency of each allele in a pairwise LD measurement in addition to genetic map distance may give a better indication of how unusual the observed LD pattern is, and hence increase statistical power. The question, ultimately, would be whether this closer approximation of haplotype-based methods has advantages over the range of well-developed haplotype-methods currently used.

4.6.7 Conclusion

Our work has demonstrated that the power of selection statistics based on LD can often be improved by controlling for variable recombination rate. Doing so is likely to reduce the number of false positive selection candidates and give a clearer indication of the relative strength of selection signals. Of the methods we tested, the α statistic and $\frac{\alpha}{E[\alpha]}$ showed highest power in simulations. In the absence of information on the genetic map, OmegaPlus was most successful at replicating selection candidates identified by other selection scan studies, despite often demonstrating low power in simulations. Simulations are often used to test the performance of selection statistics, and this pattern creates an intriguing contradiction. Focussing on this problem, we conclude pragmatically - without a greater understanding of the correlation between signals, under neutrality especially, it is difficult to interpret precisely what signal replication implies. Thus, in our study the OmegaPlus statistic was effective at finding signals that have already been identified, but we are unable to suggest the precise evolutionary meaning of these signals or whether this reflects shared true or false positive results. Based on both simulation and replication, incorporating information on expected LD using a genetic map can substantially improve the performance of selection statistics.

4.6.8 Extended Methods

Designing simple test statistics

As indicated in the main text, we defined a series of possible test statistics based on some average measurement of LD in the Along region and the Over region. When the average value of r^2 was used, we retrieve α, Eq. (4.18), and β, Eq. (4.18), which are two of the statistics we explored. However, it was often useful to use an alternative to the observed r^2 value in order to control for expected LD. A simple example would be the statistic $\alpha - \alpha_{E[r^2]}$, corresponding to the average r^2 in the Along region minus the average expected r^2 in the Along region,

$$\alpha_{E[r^2]} = \frac{\binom{l}{2}^{-1} \sum_{i \in L, j \in L} E[r^2_{i,j}] + \binom{S-l}{2}^{-1} \sum_{i \in R, j \in R} E[r^2_{i,j}]}{2} \tag{4.20a}$$

where $E[r^2_{i,j}]$ is estimated based on the generated 'LD profile', described below, and the genetic map distance between loci i and j. As a rule, we use subscripts to α and β to indicate cases where we are calculating a measure of LD in the Along or Over region respectively in a manner analogous to α and β but not based solely on the observed r^2 between loci. We used four other approaches to controlling for expected LD within the Along and Over calculations,

$$\alpha_{r^2/E[r^2]} = \frac{\binom{l}{2}^{-1} \sum_{i \in L, j \in L} r^2_{i,j}/E[r^2_{i,j}] + \binom{S-l}{2}^{-1} \sum_{i \in R, j \in R} r^2_{i,j}/E[r^2_{i,j}]}{2} \tag{4.20b}$$

$$\alpha_{\log(r^2/E[r^2])} = \frac{\binom{l}{2}^{-1} \sum_{i \in L, j \in L} \log(r^2_{i,j}/E[r^2_{i,j}]) + \binom{S-l}{2}^{-1} \sum_{i \in R, j \in R} \log(r^2_{i,j}/E[r^2_{i,j}])}{2} \tag{4.20c}$$

$$\alpha_{\text{ZScore}} = \frac{\binom{l}{2}^{-1} \sum_{i \in L, j \in L} \frac{r^2_{i,j} - E[r^2_{i,j}]}{\sigma[r^2_{i,j}]} + \binom{S-l}{2}^{-1} \sum_{i \in R, j \in R} \frac{r^2_{i,j} - E[r^2_{i,j}]}{\sigma[r^2_{i,j}]}}{2} \tag{4.20d}$$

$$\alpha_{\text{BetaCDF}} = \frac{\binom{l}{2}^{-1} \sum_{i \in L, j \in L} F(r^2_{i,j}; a, b) + \binom{S-l}{2}^{-1} \sum_{i \in R, j \in R} F(r^2_{i,j}; a, b)}{2} \tag{4.20e}$$

where, in Eq. (4.20e), $F(r^2_{i,j}; a, b)$ denotes the value, at $r^2_{i,j}$, of the cumulative distribution function of a Beta distribution with parameters a and b, fitted by maximum likelihood to r^2 measurements in the appropriate genetic map distance bin generated when creating the LD profile. This final approach is essentially attempting to estimate a p-values for each observed r^2 measurement, and averages these (which is far more conservative than using Fisher's method to combine p-values, and likely more representative given the strong correlation between LD at nearby pairs of loci).

Analogous quantities are defined when calculating variants of β, while Kelly's $Z_{nS,E[r^2]}$ is the average expected r^2 between all pairs of loci in a region. The test statistics we assessed involved simple operations on estimates of LD in the Along and Over regions, and consisted of:

1. LD and deviation from expected LD in the Along region
   - $\alpha, \quad \alpha_{r^2/E[r^2]}, \quad \alpha_{\log(r^2/E[r^2])}, \quad \alpha_{\text{ZScore}}, \quad \alpha_{\text{BetaCDF}}, \quad \alpha - \alpha_{E[r^2]}, \quad \dfrac{\alpha}{\alpha_{E[r^2]}}$
2. LD and deviation from expected LD in the Over region
   - $\beta, \quad \beta_{r^2/E[r^2]}, \quad \beta_{\log(r^2/E[r^2])}, \quad \beta_{\text{ZScore}}, \quad \beta_{\text{BetaCDF}}, \quad \beta - \beta_{E[r^2]}, \quad \dfrac{\beta}{\beta_{E[r^2]}}$
3. Kelly's Z_{nS} and deviation from expected Kelly's Z_{nS}
   - $Z_{nS}, \quad \dfrac{Z_{nS}}{Z_{nS,E[r^2]}}, \quad Z_{nS} - Z_{nS,E[r^2]}$
4. The $\omega_{\max}$ statistic and similar constant window-size alternatives that can control for their expected value
   - $\omega_{\max}, \quad \dfrac{\alpha}{\beta}, \quad \dfrac{\alpha}{\beta} - \dfrac{\alpha_{E[r^2]}}{\beta_{E[r^2]}}, \quad \dfrac{\alpha_{\text{BetaCDF}}}{\beta_{\text{BetaCDF}}}$
5. Methods similar to Kelly's Z_{nS} with more diverse approaches to controlling for expected LD
   - $\alpha+\beta, \quad \alpha_{r^2/E[r^2]}+\beta_{r^2/E[r^2]}, \quad \alpha_{\log(r^2/E[r^2])}+\beta_{\log(r^2/E[r^2])}, \quad \alpha_{\text{ZScore}}+\beta_{\text{ZScore}}, \quad \alpha_{\text{BetaCDF}}+\beta_{\text{BetaCDF}}, \quad (\alpha - \alpha_{E[r^2]}) + (\beta - \beta_{E[r^2]}), \quad \dfrac{\alpha}{\alpha_{E[r^2]}} + \dfrac{\beta}{\beta_{E[r^2]}}$
6. Alternatives to $\omega_{\max}$ that instead use the difference between LD in the Over and Along regions
   - $\alpha-\beta, \quad \alpha_{r^2/E[r^2]}-\beta_{r^2/E[r^2]}, \quad \alpha_{\log(r^2/E[r^2])}-\beta_{\log(r^2/E[r^2])}, \quad \alpha_{\text{ZScore}}-\beta_{\text{ZScore}}, \quad \alpha_{\text{BetaCDF}}-\beta_{\text{BetaCDF}}, \quad (\alpha - \alpha_{E[r^2]}) - (\beta - \beta_{E[r^2]}), \quad \dfrac{\alpha}{\alpha_{E[r^2]}} - \dfrac{\beta}{\beta_{E[r^2]}}$
7. The product of average LD in Over and Along
   - $\alpha\beta, \quad \alpha_{\text{BetaCDF}}\beta_{\text{BetaCDF}}$
8. The number of SNPs in the Along and Over region, where $|R|$ indicates the cardinality of set R
   - $\binom{|L|}{2} + \binom{|R|}{2}, \quad |L||R|$

Many of these statistics rely on the creation of an 'LD profile', which we now describe.

The LD profile

The LD profile consists of descriptive statistics of LD measurements between loci separated by a given genetic map distance. Generating the profile using simulated data involved repeatedly generating a samples of n 3MB chromosomes under neutrality and according to the relevant model of recombination and demography. As in our power simulations, loci with MAF < 0.05 were removed, followed by the removal of random loci until the average spacing between polymorphic sites was 2500bp. r^2 values were calculated for all pairs of loci up to a distance of 2cM, with the genetic map distance between loci either based on the true genetic map, low resolution map or physical distance. These LD measurements were assigned to 20000 bins according to the genetic map distance between the two loci, such that each bin represents 0.00001cM. The process was repeated 1000 times. Finally, the average LD, $E[r^2]$, for each bin and the standard deviation, $\sigma[r^2]$, were calculated, and a maximum likelihood fitting using the Scipy module [479] (scipy.stats.beta.fit) performed to obatin values of a and b.

A different LD profile was generated for each combination of sample size, demographic model, recombination model and assumed known genetic map. LD profiles were constructed for the HapMap data individually for each chromosome and population, using the two different genetic maps [390, 376].

Genetic maps

In our simulations with variable recombination rate, we considered three possible scenarios concerning the genetic map. Two of these simply involved using the physical map as a proxy (PhysMap) or providing the real section of the HapMap genetic map according to which the data was simulated (TrueMap). The third used a lower resolution version of the true genetic map (LowResMap). This was generated by downsampling the HapMap map by a factor of 15, reducing the average distance between reported map positions from approximately 817bp to 12260bp. Note that the genetic map is still accurate in the sense that it was generated using all loci, but that hotspots will be considerably smoothed out.

Simulations

Simulations were performed using MSMS [3] and Cosi2 [4]. Each simulation replicate involved simulating a sample size of $n = [20; 40; 80]$ chromosomes of length 1.5MB, which may or may not have been subject to selection at a site located in the middle of the chromosome. The mutation rate was $\mu = 1.7*10^{-8}$ in the constant size demographic model, and followed Gravel *et al* [422], $\mu = 2.38*10^{-8}$, in the Out of Africa demographic model. When recombination rate was constant this was set to $\rho = 1.1 * 10^{-8}$; the variable recombination rate was retrieved as described in Methods section, §4.6.4. The generation time was taken to be 25 years, as in [422]. To approximate SNP panel data and avoid LD measurements involving singletons, loci with MAF < 0.05 in the sample were removed before randomly removing loci until the average spacing between SNPs was 2500bp. A 200kb window was then defined from positions 650-850kb. We calculated the value of all statistics other than $\omega_{\max}$ using in-house scripts at each SNP, using three statistic window sizes ($x = [100\text{kb}; 200\text{kb}; 400\text{kb}]$). Statistics were not calculated if $|R|$ or $|L|$ were under 4, or if $|R||L| < 25$. $\omega_{\max}$ was calculated on a grid using OmegaPlus with a resolution of 2500bp using statistic window-size flags '-minwin' = [1000; 10000], '-maxwin' = [100000; 400000] and '-minsnps' = 5. The maximum and minimum value of each statistic in the 200kb window was recorded, unless SNP diversity was too low to obtain any statistic calculations in which case the replicate was not used in power calculation. Note that the removal of very SNP-sparse replicates will make our power estimate for diversity-based statistics conservative - most windows removed would have been true positives. We estimate this distortion to generally be of the order of 0.01 in the tables presented in the main text, rising to a maximum of 0.07 for the OOA scenario with selection starting at $t_1 = 1600$ generations (Table 4.2).

Two demographic scenarios were used, one with constant population size and one following an Out of Africa model [422] with samples taken from the European population. In order to obtain selective sweep trajectories under both demographic models in MSMS, we used two types of selection scenario. In the first, selection of strength s begins (pastward in time) at t_0 generations with an allele frequency of q_{t_0}. The time at which the selection phase of the model ends, t_1 generations, corresponds to the time at which the *de novo* selected allele first appears, and is determined stochastically when MSMS generates the selected allele frequency trajectory on which later coalescent simulations are conditioned. This approach is used when the population size is constant. The second method involves specifying s and times t_0 and t_1, as well as the frequency of the selected allele when the selection phase ends, q_{t_1}. Although this allows for selection on standing variation, we only consider the situation where $q_{t_1} = 0.0005$. This time, q_{t_0} is determined by the generated selected allele frequency trajectory. We use this approach when applying the OOA demographic model.

The selection scenarios investigated used an additive selection model and a selective advantage of $s = [0.01; 0.02; 0.04; 0.08]$ for the homozygote. For the constant population size demographic model, we conditioned the selection simulations on final allele frequency, $q_{t_0} = [0.3; 0.5; 0.7; 0.9; 0.99]$ with $t_0 = 0$ or $q_{t_0} = [0.99]$ with $t_0 = 800$. For the Out of Africa model, we conditioned selection simulations on the starting allele frequency $q_{t_1} = 0.0005$ and with $t_1 = [200; 400; 800; 1600]$ and $t_0 = 0$, and further removed simulations in which the selected allele became extinct such that $q_0 \neq 0$. The ROC curves from these scenarios were used to calculate the selOOA pAUC performance measure (Fig. 4.7), with otherwise similar supplementary runs using $t_1 = [2400; 3200; 4000]$ performed to assess the decay of the LD signal (Table 4.2).

The program MSMS was used for all simulations in which the recombination rate didn't vary, while Cosi2 was used for simulations with a variable recombination rate. The allele frequency trajectories used to simulate selection in Cosi2 were generated using MSMS. Example MSMS scripts are shown in Table 4.6, while Fig. 4.10 offers a schematic of

Demography	n	N_e	Mutation rate, μ	Selection, s	t_0	q_{t_0}	t_1	q_{t_1}	MSMS command line
Constant Size	20	10000	$1.7 * 10^{-8}$	Neutral	NA	NA	NA	NA	ms 20 1 -t 1020 -r 660 100000
Constant Size	40	10000	$1.7 * 10^{-8}$	0.01	800	0.99	NA	NA	ms 40 1 -t 1020 -r 660 100000 -N 10000 -SAA 200 -SAa 100 -SF 0 0.99 -Sp 0.5
OOA	80	14474	$2.36 * 10^{-8}$	Neutral	NA	NA	NA	NA	ms 80 1 -I 3 0 80 0 -t 2049.5184 -r 955.284 100000 -N 14474 -n 1 1.0 -n 2 2.33618 -n 3 3.13457 -m 1 2 1.4474 -m 2 1 1.4474 -m 2 3 1.8006 -m 3 2 1.8006 -m 3 1 0.4516 -m 1 3 0.4516 -g 2 219.59 -g 3 277.24 -em 0.0158901 1 2 8.6844 -em 0.0158901 2 1 8.6844 -en 0.0158901 2 0.129 -eg 0.015889 2 0 -eg 0.015889 3 0 -ej 0.01589 3 2 -ej 0.03524 2 1 -en 0.10225 1 0.50504
OOA	20	14474	$2.36 * 10^{-8}$	0.02	0	NA	400	0.0005	ms 20 1 -I 3 0 20 0 -t 2049.5184 -r 955.284 100000 -N 14474 -n 1 1.0 -n 2 2.33618 -n 3 3.13457 -m 1 2 1.4474 -m 2 1 1.4474 -m 2 3 1.8006 -m 3 2 1.8006 -m 3 1 0.4516 -m 1 3 0.4516 -g 2 219.59 -g 3 277.24 -em 0.0158901 1 2 8.6844 -em 0.0158901 2 1 8.6844 -en 0.0158901 2 0.1286 -eg 0.015889 2 0 -eg 0.015889 3 0 -ej 0.01589 3 2 -ej 0.03524 2 1 -en 0.10225 1 0.50504 -SI 0.006909 3 0 0.0005 0 -Sc 0 2 578.96 289.48 0 -Sp 0.5 -Smark -oOC -oTrace
OOA	40	14474	$2.36 * 10^{-8}$	0.04	0	NA	1600	0.0005	ms 40 1 -I 3 0 40 0 -t 2049.5184 -r 955.284 100000 -N 14474 -n 1 1.0 -n 2 2.33618 -n 3 3.13457 -m 1 2 1.4474 -m 2 1 1.4474 -m 2 3 1.8006 -m 3 2 1.8006 -m 3 1 0.4516 -m 1 3 0.4516 -g 2 219.59 -g 3 277.24 -em 0.0158901 1 2 8.6844 -em 0.0158901 2 1 8.6844 -en 0.0158901 2 0.1286 -eg 0.015889 2 0 -eg 0.015889 3 0 -ej 0.01589 3 2 -ej 0.03524 2 1 -en 0.10225 1 0.50504 -SI 0.027636 3 0 0.0005 0 -Sc 0.015891 2 1157.92 578.96 0 -Sc 0 2 1157.92 578.96 0 -Sp 0.5 -Smark -oOC -oTrace
OOA	80	14474	$2.36 * 10^{-8}$	0.08	0	NA	3200	0.0005	ms 80 1 -I 3 0 80 0 -t 2049.5184 -r 955.284 100000 -N 14474 -n 1 1.0 -n 2 2.33618 -n 3 3.13457 -m 1 2 1.4474 -m 2 1 1.4474 -m 2 3 1.8006 -m 3 2 1.8006 -m 3 1 0.4516 -m 1 3 0.4516 -g 2 219.59 -g 3 277.24 -em 0.0158901 1 2 8.6844 -em 0.0158901 2 1 8.6844 -en 0.0158901 2 0.1286 -eg 0.015889 2 0 -eg 0.015889 3 0 -ej 0.01589 3 2 -en 0.0352399 2 1.0 -ej 0.03524 1 2 -en 0.10225 2 0.50504 -SI 0.055272 3 0 0.0005 0 -Sc 0.035241 2 2315.84 1157.92 0 -Sc 0.015891 2 2315.84 1157.92 0 -Sc 0 2 2315.84 1157.92 0 -Sp 0.5 -Smark -oOC -oTrace

Table 4.6: Example MSMS command lines used to simulate different demographic and selection scenarios with a constant recombination rate of $\rho = 1.1 * 10^{-8}$. Note that certain slight timing offsets and commands, such as setting growth rates to zero before merging populations, may be technically redundant, but were included in the simulations to ensure robust behaviour. For the OOA scenarios, population indices 1, 2 and 3 usually correspond to African, Eurasian/European and East Asian. When selection began before the OOA bottleneck at 51,000 years ago, however, we re-code the simulation such the populations 1, 2 and 3 correspond to post-OOA African, pre-OOA African/Eurasian/European and East Asian respectively, with the OOA bottleneck implemented by transferring all lineages from the post-OOA population (1) to the pre-OOA population (2) and converting the Eurasian population (2) to the pre-African population (2) by increasing its effective population size. The selection phases for the OOA model given different values of t_1 are shown in Fig. 4.10

the OOA demograhpic model, indicating the populations that may be under selection (depending on t_1).

ROC curves were calculated for each statistic by comparing 1000 neutral replicates with at least 300 replicates involving selection. As the statistics we used employed different signatures to detect selection, four ROC curves were calculated, based on top or bottom

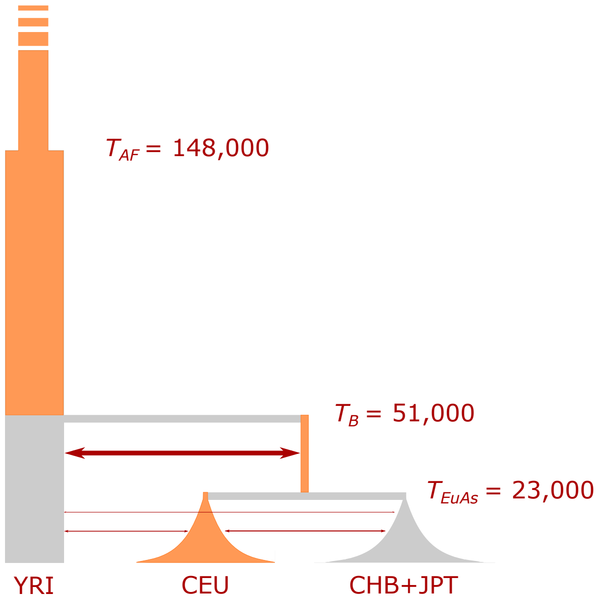


FIGURE 4.10: Schematic of the Out of Africa model of Gravel *et al* [422]. The vertical axis represents time, running from bottom (the present) to top, and the horizontal axis population size. For population size and migration parameters, see Table 2 in [422]. Three time events are indicated - the time of an ancient African bottleneck, T_{AF}, the time of the Out of Africa bottleneck, T_{B}, involving the Eurasian population splitting from the African (YRI) population, and the time of the split of the Eurasian population into Europeans (CEU) and East Asians (CHB+JPT), T_{EuAs}. Selection ends, pastward in time, at t_1 generations, such that, depending on t_1, selection will act in different populations. The populations that would, if $t_1 > 5920$ (148,000 years), experience selection are coloured orange. Note that the exact implementation of the model when $t_1 > 2040$ (51,000 years) is modified slightly, see Table 4.6.

outliers of the maximum or minimum statistic value in the 200kb window indicating selection. These were used to determine the pAUC between an FPR of 0 and 0.05. When assessing the performance of statistics, we did not want to make assumptions about the direction of deviation indicating selection or the window size used in the selection scan. We therefore chose the maximum pAUC for each statistic at a sample size of $n = 40$ (usually three window sizes and 4 pAUC each, so the maximum of 12 pAUC values) as its measure of performance under a given selection, recombination rate and demographic scenario. To summarise the performance of statistics under different selection models, we separated the selection scenarios into three groups, Low Frequency, High Frequency and selOOA as detailed in the Methods section. We averaged the maximum pAUC across scenarios in these groups to give an overall indication of average statistic performance. Note that we implicitly give equal weight to each selection coefficient. In the case of selOOA, we did not include those scenarios for which the average frequency of the selected allele at sampling was low, $q_0 < 0.05$, corresponding to $s = 0.01$ with $t_1 = [200; 400]$, and $s = [0.02; 0.04]$ with $t_1 = 200$.

Defining selection candidates

To compile a list of previously suggested selection candidates, we searched for a range of studies performing selection scans based on the site frequency spectrum (SFS) or population differentiation. We identified 8 appropriate selection scans, summarised in Table 4.7.

We converted the suggested signal locations into single 200kb windows by identifying

Reference	Data	Relevant population	Method	Class	Outliers reported	European Chr2	Asian Chr2	European Chr15
[358]	HapMap SNP data	CEU	CLR	SFS	23 (Chr 2)	23		
[454]	Perlegen SNP data	European and Asian	Tajima's D	SFS	23 (European) 29 (Asian)	2	5	
[455]	Low-coverage (184k SNPs)	European	S2Fst	Differentiation	162	13		5
[361]	CEPH SNP data	European and East Asian	CLR	SFS	10 (European) 10 (East Asian)	1	1	1
[456]	HapMap Phase II SNP data	European	XP-CLR	Differentiation, SFS	40 (CEU vs YRI) 40 (CHBJPT vs YRI)	3	3	4
[457]	1000 genomes project	European	XP-Sfselect	Differentiation, SFS	40 (Europeans vs Africans)	5		2
[458]	1000 genomes project	European and Asian	DDAF	Differentiation	110 SNPs (Europe) 110 SNPs (Asian) 73 SNPs (CEU)	19	12	5
[6]	Full genome data	European	Tajima's D	SFS	65 (Top 0.5%, SWEuropeans)	6		4
Total						72	21	21

Table 4.7: Selection scan studies used to define the candidate selection signal set.

which 200kb window overlapped the central point of the signal or the SNP reported. Very few windows were identified in multiple studies (6 overlaps in European Chr2, 4 in European Chr15 and 3 in Asian Chr2), although there were obvious clusters of signals. A complete table of signals is included in Appendix 3, §4.7.1.

To assess whether 200kb selection candidate windows containing protein-coding genes were preferentially replicated, we focussed on windows suggested based on European Chromosome 2 data. These were split into two groups, those that contained protein coding genes and those that did not, based on hg18 RefSeq Genes track refGene table accessed through the UCSC table browser (genome.ucsc.edu/cgi-bin/hgTables), before being assessed for signal overlap with our LD-based methods as usual.

Resampling to assess signal overlap

To give an indication of how unexpected observed signal replication was, we resampled the appropriate number of random windows 10,000 times and compared the observed statistic value, rank of candidate windows and number of outliers to this set. We noticed that windows were often clustered and that peaks in several of the LD-based statistics often included several consecutive windows. We therefore approximated this clustering in our resampling regime, copying distribution of consecutive windows seen in the candidate signal data. For example, in the European Chromosome 15 data there were 9 solitary windows, 2 runs of 2 consecutive windows and a single run of four consecutive windows, and when resampling we followed this pattern.

Supplementary Information

Refining the use of linkage disequilibrium as a robust signature of selection

G. S. Jacobs, T. J. Sluckin and T. Kivisild

4.6.9 Appendix 1: Detailed simulation performance results for tested statistics

The performance metric data used to create Fig 4.7 is presented in Table 4.8. The similar performance of statistics that control for expected LD in the LowResMap and TrueMap variable recombination rate scenarios is evident. The consistency of statistics that do not control for recombination rate (e.g. α, β, Kelly's Z_{nS}, OmegaPlus etc.) over the three recombination rate scenarios give an indication of the consistency of the performance metric, as these power assessments were (unecessarily) repeated. While only two demograhpic scenarios were assessed, we considered many variations on simple LD-based statistics. As such, full power results are not presented here but are available from the author on request.

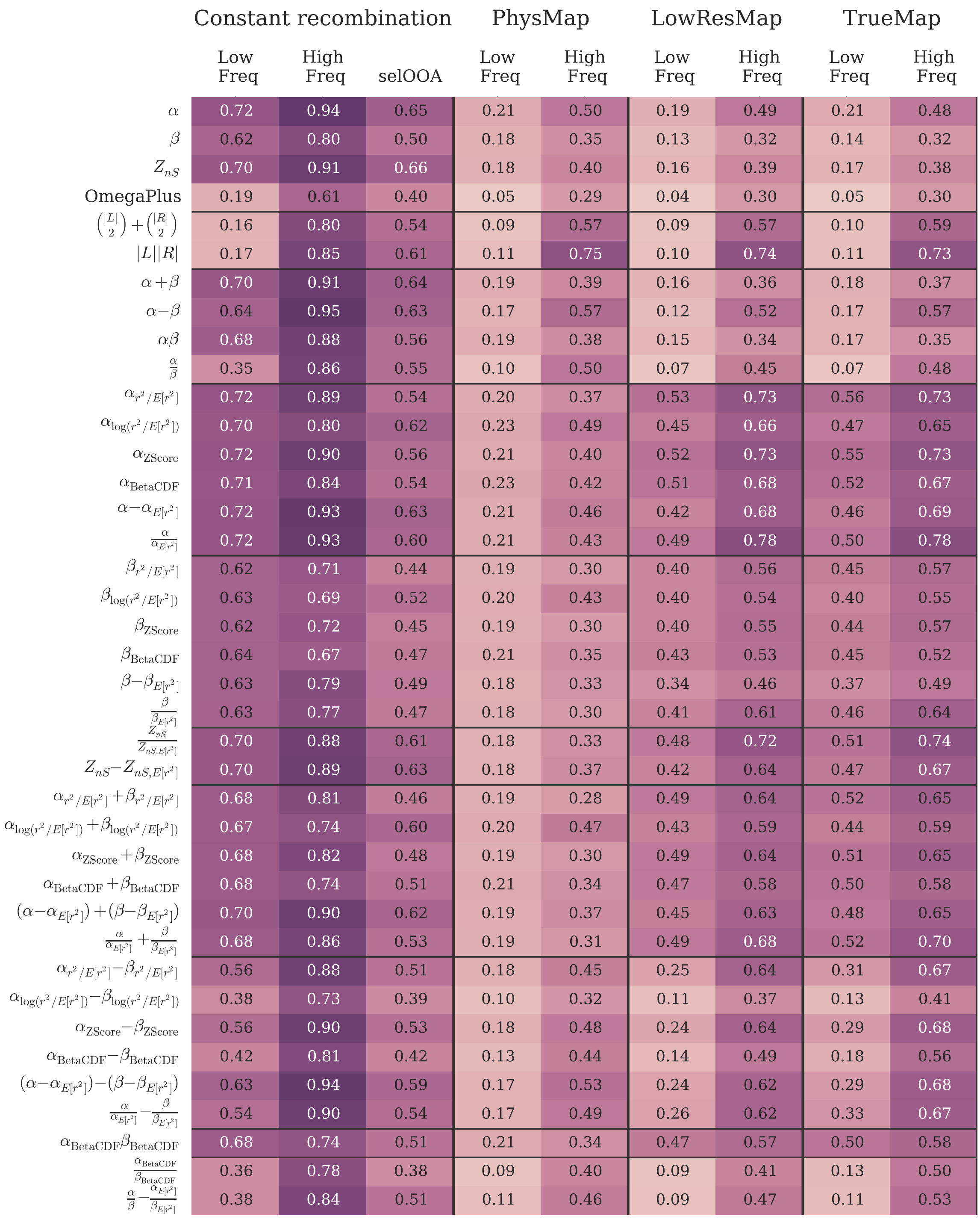
As indicated in the Extended Methods, these performance measures equally weight the four selection scenarios (additive positive selection on the homozygote $s = [0.01;\ 0.02;\ 0.04;\ 0.08]$). As might be expected, most statistics displayed very high power for scenarios involving very strong selection that is often unrealistic, such that the performance estimate is most useful for comparing statistics rather than as a strict indication of their power when applied to real data. Note also that selection is scaled by population size in coalescent simulations, such that the parameter $2N_e s$ describes selection strength and not s itself. Thus, if N_e is larger than that of humans in a natural population (*Drosophila melanogaster* for example) then sweeps caused by weaker positive selection are more likely to be identified.

4.7 The role of window size in determining expected average LD

Both the original formulation of the ω statistic [357],

$$\omega = \frac{(\binom{l}{2} + \binom{S-l}{2})^{-1}(\sum_{i,j \in L} r_{i,j}^2 + \sum_{i,j \in R} r_{i,j}^2)}{(l(S-l))^{-1} \sum_{i \in L, j \in R} r_{i,j}^2}, \qquad \text{(4.17 revisited)}$$

and the OmegaPlus genome-scan version [444] attempt to maximise the value of ω by varying the genomic region used to define the sets of SNPs L and R. OmegaPlus also allows for variable overall statistic window sizes, the total genomic region considered by the statistic that contains the S SNPs that are partitioned into L and R, using



	Constant recombination			PhysMap		LowResMap		TrueMap	
	Low Freq	High Freq	selOOA	Low Freq	High Freq	Low Freq	High Freq	Low Freq	High Freq
α	0.72	0.94	0.65	0.21	0.50	0.19	0.49	0.21	0.48
β	0.62	0.80	0.50	0.18	0.35	0.13	0.32	0.14	0.32
Z_{nS}	0.70	0.91	0.66	0.18	0.40	0.16	0.39	0.17	0.38
OmegaPlus	0.19	0.61	0.40	0.05	0.29	0.04	0.30	0.05	0.30
$\binom{\lvert L\rvert}{2}+\binom{\lvert R\rvert}{2}$	0.16	0.80	0.54	0.09	0.57	0.09	0.57	0.10	0.59
$\lvert L\rvert\lvert R\rvert$	0.17	0.85	0.61	0.11	0.75	0.10	0.74	0.11	0.73
$\alpha+\beta$	0.70	0.91	0.64	0.19	0.39	0.16	0.36	0.18	0.37
$\alpha-\beta$	0.64	0.95	0.63	0.17	0.57	0.12	0.52	0.17	0.57
$\alpha\beta$	0.68	0.88	0.56	0.19	0.38	0.15	0.34	0.17	0.35
$\frac{\alpha}{\beta}$	0.35	0.86	0.55	0.10	0.50	0.07	0.45	0.07	0.48
$\alpha_{r^2/E[r^2]}$	0.72	0.89	0.54	0.20	0.37	0.53	0.73	0.56	0.73
$\alpha_{\log(r^2/E[r^2])}$	0.70	0.80	0.62	0.23	0.49	0.45	0.66	0.47	0.65
α_{ZScore}	0.72	0.90	0.56	0.21	0.40	0.52	0.73	0.55	0.73
$\alpha_{\mathrm{BetaCDF}}$	0.71	0.84	0.54	0.23	0.42	0.51	0.68	0.52	0.67
$\alpha-\alpha_{E[r^2]}$	0.72	0.93	0.63	0.21	0.46	0.42	0.68	0.46	0.69
$\frac{\alpha}{\alpha_{E[r^2]}}$	0.72	0.93	0.60	0.21	0.43	0.49	0.78	0.50	0.78
$\beta_{r^2/E[r^2]}$	0.62	0.71	0.44	0.19	0.30	0.40	0.56	0.45	0.57
$\beta_{\log(r^2/E[r^2])}$	0.63	0.69	0.52	0.20	0.43	0.40	0.54	0.40	0.55
β_{ZScore}	0.62	0.72	0.45	0.19	0.30	0.40	0.55	0.44	0.57
β_{BetaCDF}	0.64	0.67	0.47	0.21	0.35	0.43	0.53	0.45	0.52
$\beta-\beta_{E[r^2]}$	0.63	0.79	0.49	0.18	0.33	0.34	0.46	0.37	0.49
$\frac{\beta}{\beta_{E[r^2]}}$	0.63	0.77	0.47	0.18	0.30	0.41	0.61	0.46	0.64
$\frac{Z_{nS}}{Z_{nS,E[r^2]}}$	0.70	0.88	0.61	0.18	0.33	0.48	0.72	0.51	0.74
$Z_{nS}-Z_{nS,E[r^2]}$	0.70	0.89	0.63	0.18	0.37	0.42	0.64	0.47	0.67
$\alpha_{r^2/E[r^2]}+\beta_{r^2/E[r^2]}$	0.68	0.81	0.46	0.19	0.28	0.49	0.64	0.52	0.65
$\alpha_{\log(r^2/E[r^2])}+\beta_{\log(r^2/E[r^2])}$	0.67	0.74	0.60	0.20	0.47	0.43	0.59	0.44	0.59
$\alpha_{\mathrm{ZScore}}+\beta_{\mathrm{ZScore}}$	0.68	0.82	0.48	0.19	0.30	0.49	0.64	0.51	0.65
$\alpha_{\mathrm{BetaCDF}}+\beta_{\mathrm{BetaCDF}}$	0.68	0.74	0.51	0.21	0.34	0.47	0.58	0.50	0.58
$(\alpha-\alpha_{E[r^2]})+(\beta-\beta_{E[r^2]})$	0.70	0.90	0.62	0.19	0.37	0.45	0.63	0.48	0.65
$\frac{\alpha}{\alpha_{E[r^2]}}+\frac{\beta}{\beta_{E[r^2]}}$	0.68	0.86	0.53	0.19	0.31	0.49	0.68	0.52	0.70
$\alpha_{r^2/E[r^2]}-\beta_{r^2/E[r^2]}$	0.56	0.88	0.51	0.18	0.45	0.25	0.64	0.31	0.67
$\alpha_{\log(r^2/E[r^2])}-\beta_{\log(r^2/E[r^2])}$	0.38	0.73	0.39	0.10	0.32	0.11	0.37	0.13	0.41
$\alpha_{\mathrm{ZScore}}-\beta_{\mathrm{ZScore}}$	0.56	0.90	0.53	0.18	0.48	0.24	0.64	0.29	0.68
$\alpha_{\mathrm{BetaCDF}}-\beta_{\mathrm{BetaCDF}}$	0.42	0.81	0.42	0.13	0.44	0.14	0.49	0.18	0.56
$(\alpha-\alpha_{E[r^2]})-(\beta-\beta_{E[r^2]})$	0.63	0.94	0.59	0.17	0.53	0.24	0.62	0.29	0.68
$\frac{\alpha}{\alpha_{E[r^2]}}-\frac{\beta}{\beta_{E[r^2]}}$	0.54	0.90	0.54	0.17	0.49	0.26	0.62	0.33	0.67
$\alpha_{\mathrm{BetaCDF}}\beta_{\mathrm{BetaCDF}}$	0.68	0.74	0.51	0.21	0.34	0.47	0.57	0.50	0.58
$\frac{\alpha_{\mathrm{BetaCDF}}}{\beta_{\mathrm{BetaCDF}}}$	0.36	0.78	0.38	0.09	0.40	0.09	0.41	0.13	0.50
$\frac{\alpha}{\beta}-\frac{\alpha_{E[r^2]}}{\beta_{E[r^2]}}$	0.38	0.84	0.51	0.11	0.46	0.09	0.47	0.11	0.53

Table 4.8: pAUC-based performance metric, which assesses relative performance for all tested statistics, over all demographic and recombination rate scenarios. The three columns 'PhysMap', 'LowResMap' and 'TrueMap' indicate performance when a variable recombination rate was applied and the specified genetic map was assumed to be available for calculating statistic. Groups of similar statistics (for example, those that measure diversity or those that make different attempts to control for expected LD in α) are indicated; the first 10 statistics are those that do not control for variable recombination rate. For more details on statistic definitions and simulation scenarios see Methods and Extended Methods.

the command flags '-minwin' and '-maxwin'. Each approach involves comparing several evaluations of Eq. 4.17 to find $\omega_{\max}$. We here consider the expected value of these evaluations under a simplified representation of neutral LD patterns, and suggest that the expected value of ω will often vary depending on window size. This makes the value of $\omega_{\max}$ more difficult to interpret and precludes comparison of $\omega_{\max}$ values obtained using different window size parameters.

An intuitive understanding can be built by considering simple representations of the decay of LD with distance between loci. The most basic case is diagrammatically illustrated in Figure 4.11. We consider a model by which r^2 is a constant value $0 \ll r_c^2 < 1$ between all loci separated by a distance y_c or less, and 0 for all loci separated by a distance greater than y_c. When the statistic window size x is small, $x \leq y_c$, all LD measurements will equal r_c and Eq. 4.17 will evaluate to 1. When the window size is moderately large, $y_c < x \leq 2y_c$, the numerator of Eq. 4.17 will evaluate to 1 while the denominator is less than 1; ω is correspondingly greater than 1. When $x = 2y_c$, shown in Fig. 4.11, ω will be 2. Finally, as the window size is allowed to become very large, $x \to \infty$, both the numerator and denominator of Eq. 4.17 will tend to 0, leading to erratic statistic behaviour.

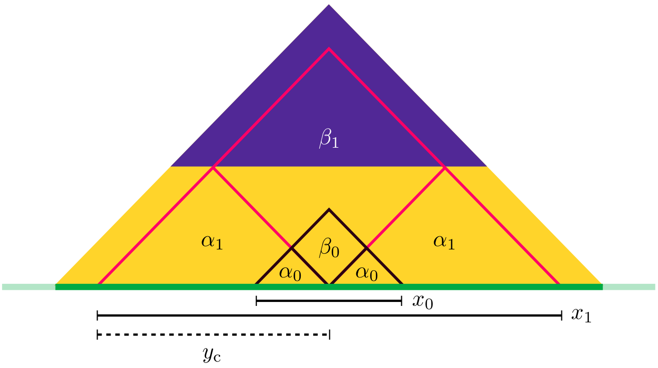


FIGURE 4.11: Schematic of the step-decay in LD described in the text, and two example statistic window size constructions leading to different expected values of ω. The horizontal bar (here green) indicates the chromosome, with the yellow and purple triangle representing the matrix of expected pairwise LD values between loci along the chromosome. Yellow shows positive LD between SNPs separated by a distance y_c or less, while purple indicates no LD. When the window size is x_0, the value of ω is expected to be 1 in this simple illustrative scenario; when the window size is $x_1 = y_c$, ω is expected to be 2. The 'Along' and 'Over' regions are indicated as α and β respectively.

Although the step-change decay in LD qualitatively discussed above is not a particularly good approximation of the decay of LD with distance, the essential point - that window size construction is likely to strongly impact the value of Eq. (4.17) - is clearly illustrated.

4.7.1 Appendix 3: Selection candidate windows

This appendix is included as Thesis Appendix A4, at the end of the thesis.

4.7.2 Appendix 4: Further selection scans and replication data

This appendix is included as Thesis Appendix A5, at the end of the thesis.

4.7.3 Appendix 5: Top 1% hits based on selected statistics

This appendix is included as Thesis Appendix A6, at the end of the thesis.

The draft manuscript presented above uses coalescent simulations to assess the power of a range of selection statistics, with suggestive support for my findings arising in the replication of previously identified selection candidates. I now extend my discussion of coalescent simulations in power analyses.

4.8 Assessing test statistic performance through coalescent simulations

4.8.1 Robustness and flexibility of the coalescent

The most complex scenario we investigate involves a series of subpopulations linked by migration, with population size changes, a variable recombination rate, and a selective sweep. To identify aspects of this model for which the coalescent approach may offer a poor approximation, it is necessary to consider the two assumptions underlying Eq (4.7). These are

- That coalescence of lineages occurs in pairs, and only one coalescence occurs at a time.
- That chromosomes are exchangeable - that is, assigning offspring to parents occurs according to identically distributed, though not independent, draws from a distribution

The first assumption has implications for population size (N should be large compared to n) and for the variance in reproduction, as very high reproductive variance can involve many lineages descending from a single one in a generation. The second assumption means that there is no population structure of any sort, and no selection. The non-independence of offspring numbers arises from the deterministic population size of the model.

Although Eq (4.7) does not represent a structured population with varying size, these features are easily incorporated into the coalescent framework [434, 431], and are implemented in the standard coalescent programs I apply (MSMS [3] and Cosi2 [4]). I focus, then, on two issues - the accuracy of the coalescent in modelling small population sizes and selection, and the generality of the coalescent approach with respect to the breadth of models that it captures.

Small population sizes and the coalescent

In our models, the population size of subpopulations is sometimes $N < 1000$. To avoid coalescent events including more than two lineages, n should be small; the appropriate limit is $n \leq \sqrt{2N}$ for Kingman's n-coalescent [471]. Our largest sample size is $n = 80$, which exceeds that limit. Fortunately, recent work suggests that the standard coalescent is remarkably robust to small N. For example, a wide range of statistics summarising tree structure and topology approximate the true values to within 10% when $n < \frac{N}{2}$, and often much more effectively [471]. Given this, and noting that small population sizes are not in the very recent past (such that the number of remaining lineages will be significantly below 80) I do not consider it likely that these parameters will strongly bias our neutral simulations.

A greater issue arises in our modelling of selection. The approach implemented in MSMS involves using a forward time diffusion approximation to obtain a stochastic allele frequency trajectory, and then conditioning the coalescent process on this [3]. A similar method is used by Cosi2, except that the stochastic frequency trajectory must be generated externally [4]. The important point is that the size of the subpopulation with the selected allele can become very small very quickly if the selection coefficient s is large. If the selective sweep occurs quickly enough, it is likely that many lineages will survive, pastward, to the early stages of selection, with multiple mergers likely [469]. Strong selection is often modelled as a star-like phylogeny (e.g. in [441]), with all lineages coalescing at the same time, a structure that is impossible given the standard coalescent.

The strongest selection coefficient we investigate is $s = 0.08$. As individuals are diploid and fitness is additive, we can approximate the time taken for fixation in our scenarios by iterating Eq. II-49 of [420],

$$p_{t+1} = \frac{p_t(1 + s + sp_t)}{1 + 2sp_t}, \tag{4.21}$$

which describes the frequency of an allele p at time $t+1$ given its frequency at time t and selection coefficient s. With initial frequency $p_0 = 0.0005$, it takes roughly 205 generations for the allele frequency to reach $p = 0.9995$. Given that our maximum sample size is $n = 80$, coalescent events involving more than two lineages in a generation seem probable. This issue arises, but is not commented on, in many simulation studies involving selection, such that it may deserve detailed study. For the present, I note that patterns of linkage disequilibrium under selection have been investigated theoretically using a range of approaches, including a star-like genealogical approximation of strong selection [441], forward time recurrence equations for selection and recombination [442], and a two-locus genealogical approach [440]. The patterns of linkage disequilibrium due to selection are broadly consistent in these models, and correspond well with those found in our simulations, Fig. 4.5. The implication is that the patterns of linkage disequilibrium that we hope to investigate are captured by the coalescent algorithms implemented in MSMS and Cosi2.

Coalescent process exist that incorporate simultaneous coalescence events [480], coalescence events involving more than two lineages do exist [481], or both. These are not implemented in standard simulation programs, but could be useful in investigating strong selection as an intermediate representation between the standard bifurcating tress and a star-like genealogy.

Comments on the robustness of the coalescent

The coalescent described by Kingman is remarkably robust [355, 482, 483] in that it approximates many models of populations evolving under neutrality for large $N >> n$. These include the classic Wright-Fisher model and Moran process [355]. Interestingly, a wide range of violations to the Wright-Fisher model manifest as a linear re-scaling of time for Eq (4.7), or, equivalently, of N [482]. The important factor is the time-scale of different demographic events as compared to coalescence events. For example, the genealogy created when subpopulations are linked by fast migration is well characterised by a panmictic population with a re-scaled population size. The same is true when population size fluctuates at a rate that is fast or slow versus the rate of coalescence [482].

One implication of this is that the specified coalescent models capture a wider range of evolutionary behaviours than our parameter settings imply, suggesting that selection might be detected given many unknown demographic models that we do not explicitly investigate. The converse - that our selective scenarios may map to population models that lack selection - is not overly problematic, as selection will tend to affect a local genomic region only. Methods that allow a worker to easily summarise the breadth of scenarios explored by a model with given parameter settings would be of immense practical use, but, as far as I am aware, do not exist.

The above suggests that the coalescent method can represent the demographic and selection models we specify with reasonable, though not perfect, accuracy.

4.8.2 Coalescent models and human evolution

I have already noted in the manuscript that the coalescent method should offer a good approximation of the human demographic model suggested by Gravel *et al*, which is based on a Wright-Fisher model of population evolution [421, 422]. Indeed, coalescent simulations are utilised in that modelling work to estimate parameter confidence intervals [421]. Although their model is inevitably a simplification of real human demography, it is broadly consistent with our prior beliefs about human evolutionary history, albeit without incorporating admixture from Neanderthal and other archaic populations. The hope is that their various parameters (which include subpopulation sizes that change over time, with migration between them) will absorb subtleties of neutral human evolution, such as archaic admixture or patterns of inbreeding, that will tend to have a genome-wide impact on the pattern of human genetic variation.

A more complex problem is raised by selection. In my simulation work, I compare scenarios incorporating a single selective sweep with demographically matched neutral scenarios. The assumption here is that selection is rare enough that the vast majority of the genome will appear, in terms of genealogy, neutral. In real evolving populations, this is not necessarily true. Since the 1960s, the neutral [392] or nearly-neutral [484] theories of molecular evolution have been the mainstream models of molecular evolution [372]. These suggest that most novel mutations are either neutral or weakly selected, with some portion also being strongly deleterious and rapidly purged from the population. Current data suggests that this model applies to some species more than others [467]. For example, the genome of the model organism *Drosophila melanogaster* shows consistent deviations from neutrality, with strong evidence of purifying selection as well as adaptive selective sweeps [485]. The role of purifying selection as a significant evolutionary force is also supported across the mammal phylogeny [486], and within humans [461, 478]. Positive selective sweeps in humans are thought to be quite rare [478, 487], such that many detected selection candidates may be false positives [370, 372].

Despite the relatively low N_e of humans, which reduces the efficiency of selection, there is evidence for both purifying selection and occasional selective sweeps. A test statistic designed to detect positive directional selection, then, would ideally also be assessed for its ability to differentiate neutrality and selection given a genetic background subject to purifying selection. In the coalescent framework, purifying selection leads to an excess of terminal branches [488, 489]. This effect can be approximated using a neutral coalescent model with a shrinking population size, pastward in time, from the genome-wide N in the present to some constant $< N$ [490]. The outcome is a reduction in local diversity coupled with a relative enrichment of rare alleles, which resembles the effect of positive selection and would lead to false-positives for directional selection when some statistics (e.g. Tajima's D [388]) are used [488]. Approaches for simulating purifying selection exist [491, 492], although currently only forward-time simulation programs offer the flexibility to model directional selection on a genetic background subject to purifying selection.

In the manuscript above, I do not investigate the behaviour of LD-based test statistics given purifying selection, for several reasons. Fistly, and most critically, I have chosen to focus on the robusticity of certain selection statistics to variable recombination rate. Extending the modelling to include a background of purifying selection warrants detailed study in its own right. Secondly, I apply the test statistics to humans. The relatively low effective population size of humans mean that purifying selection is likely to be less prevalent in general. Finally, some of my demographic models include recent exponential population growth, which is sometimes coupled with positive directional selection. It may be that including these models at least gives an indication of statistic performance under a background of genetic variation subject to purifying selection. Nevertheless, investigating the impact of purifying selection on LD-based test statistics would be an interesting extension to my work.

4.9 Simulation as a tool to improve complex test statistics

Simulations have been a critical tool in the study I present above. Investigating another measure of selection statistics performance - their ability to detect known selection signals - complements the computational work, but emphasises the difficulties in interpreting successful replication. The results of the two approaches are broadly in agreement. Other useful approaches could include confirming statistic performance using data from selection experiments that are sufficiently artificial to record details of phenotype, genotype and reproductive success, but sufficiently faithful to demography and selection scenario relevant to natural populations.

Simulations, nevertheless, have substantial advantages over alternative methods. First among these is practicality. Manipulating a very large experimental population over many generations would be expensive and time consuming. Simulation using the coalescent method is rapid, such that many replicates of selection given a range of demographic scenarios can be practically explored. The second advantage is their flexibility. Developments in the mathematical framework of coalescent theory have allowed workers to design simulation algorithms that accurately represent a wide variety of demographic and selection scenarios and models of molecular evolution. Although it is not feasible to explore the parameter space of possible evolutionary models in detail, focussed simulations can simply assess the robustness of a statistic's performance to specific assumptions - in my case, a constant recombination rate. The work presented above can be considered a qualitative indication of the importance of variable recombination rate in reducing the power of selection statistics based on LD, and an exploration of methods to mitigate this effect.

In this chapter, I have tried to emphasised both the challenges facing GWSSs and their promise. Similarly, I have explored the assumptions involved in using simulations to investigate the performance of statistics used in GWSSs. While it is important to have an awareness of the limitations of this approach, the flexibility and computational efficiency of simulations mean that they are an invaluable tool for characterising the performance of test statistics in population genetics under different assumptions about the evolutionary process.

Chapter 5

Discussion and conclusions

The work presented in this thesis has involved using simple simulation models to represent populations changing over time. Through this approach, I have been able to test the modelling assumptions of a mathematical representation of species invasions, extend and explore a model of population subdivision to investigate migration biases arising from animal trading, and clarify the impact of an incorrect simplification on the power of previously suggested and novel selection scan statistics. The specific advantage of simulation over other approaches is different in each case, but can be broadly characterised as the ability to use flexible representation to quantitatively explore the various models. Flexibility is best understood in comparison to verbal description or mathematical analysis of models. Computational simulation maintains the valuable logical exactness of mathematical analysis, while allowing the exploration of systems that would be challenging to mathematically analyse. Although it may be easier to represent some models verbally, it is much more challenging to compare plausible but contradictory verbal models and identify any logical inconsistencies.

When designing a model of a target system, simplifying assumptions may be introduced for various reasons. We can describe the assumption of an infinitely divisible population, as taken in integro-difference models of species invasions, as a manipulation assumption. The assumption is not meant to accurately represent the target system, but allows the velocity behaviour of the wave of advance to be derived given long-range dispersal kernels that are difficult to accurately represent using a reaction-diffusion equation (though see work on fractional reaction-diffusion equations, eg. [93]). In the case of my work on using linkage disequilibrium as a signal of natural selection, the assumption of constant recombination rate can be considered a constraint assumption. Again, the assumption is known to be incorrect, but in the context of power analysis the ease of model manipulation is not effected. Rather, the parameter space is constrained, reducing the amount of computational processing and data analysis required. An assessment of statistical power using a constant recombination rate is likely to be incorrect because the causal processes leading to the relevant signal of selection are poorly represented.

The assumption of constant, directionally symmetric, genetically invariable migration rate taken when modelling gene flow can be considered an assumption of system representation. This has been actively chosen to provide a best-guess representation of the target system. In the absence of specific evidence for biased migration rates - and, indeed, given certain results suggesting that they are evolutionarily unlikely under some circumstances (eg. [269]) - the simplest representation might be considered as the most appropriate. In such circumstances, assumptions should be revised as a more comprehensive description of the specific target system becomes available, as is the case when we model gene flow occurring through markets as opposed to other mechanisms.

The three models I have explored, then, provide a range of examples of how simulations can be used to investigate different modelling assumptions in population biology. Through this work, I have further presented a range of novel results, showing that dispersal stochasticity is an important feature when modelling long-range dispersal, that surprising patterns of genetic variation can arise due to migration biases such as those caused by market-mediated gene flow, and how selection scan statistics can be improved by controlling for variable recombination rate. The utility of simple simulations in testing and extending models in population biology is clear.

5.1 Assessing the robustness of my modelling results

In the introduction of this thesis, I discussed the challenges involved in assessing the robustness of modelling results. Much of the work I have conducted can be characterised in this light, investigating the importance of specific modelling assumptions on determining model behaviour. I have attempted to confirm the robustness of my own findings in various ways. In Chapter 2, I supplemented my main modelling work with investigation of alternative species invasion models, including one closely based on the integro-difference equation of Kot *et al* [46]. I also confirmed that my simulation results corresponded to analytically retrieved velocities in certain limits, as retrieved from marginal stability analysis of the linearised wave front for example, and that certain observed velocities were within the bound derived from the work of Clark *et al* [106].

In Chapter 3, I confirmed the robustness of the essential result - that migration biases, such as those potentially caused by animal trading, can lead to a high equilibrium frequency of a locally or globally negatively selected animal type - to various representations of the market process and population regulation. I also attempted to parameterise the model according to properties of the Indian cattle market, and discussed evidence for some of the modelling narratives I observed among agricultural communities in Africa and India especially. I have not, as yet, directly incorporated quantitative data on animal trading dynamics into the model, and this remains an important future step.

In Chapter 4, I was able to use real genetic data to assess the selection statistics proposed, finding that controlling for expected recombination rate increased the number of previously identified selection signals that statistics could replicate. This generally supports the results of my power analysis, although it is difficult to precisely interpret the meaning of selection signal replication. While I do not make extensive use of real data in this thesis, when simulation is intended to represent the world it is important to incorporate observations of the world when designing and refining models.

Although I have attempted to be thorough in assessing the robustness of my various results and models, a clear finding from my work is that relatively small changes to the assumptions of a model can radically alter its behaviour. This can be observed in Chapters 2 and 3 especially. The implication is that some reservation is recommended when interpreting all modelling results, including my own.

5.2 The utility of models when robustness is uncertain

The finding that modelling results can depend strongly on seemingly innocuous modelling assumptions is widely echoed in other modelling work. A trivially simple example arises in the modelling of selection, where deterministic models guarantee the survival and approach to fixation of a novel advantageous allele at long times while stochastic models allow for the extinction of the allele, or its ultimate fixation. Correctly representing population growth using a stochastic, as opposed to a deterministic, frequency trajectory of a selected allele has been found to be important in coalescent simulations (eg. [470]).

Remaining on the topic of selection, I have been involved in work showing that the algorithmic details of the representation of fitness in population models impacts model behaviour [5]. Fitness differences are often represented as different reproduction probabilities, as in the classic Moran model [426], termed the *birth-death process*. However, it is also possible to represent fitness differences as different probabilities of mortality, which is known as the *death-birth process*. A simple example arises in the Prisoners Dilemma game described by Kaiping *et al* [5], in which the time taken for a previously monomorphic population to fix with a single strategy after a population is invaded either by a co-operator or defector is approximately 105.3 generations given the birth-death process, as compared to 11.4 generations given a death-birth process. The difference in times is due to the efficiency of the death-birth process in quickly eliminating low-fitness variants, while the birth-death process merely guarantees that they are unlikely to reproduce. When mutation rate is rapid such that three or more strategies exist in the population at once, we found several models to show different equilibria in addition to different dynamics depending on the update rule.

Work on the evolution of the migration rate is also relevant. Again, models show contradictory results depending on their representation of the system. Various models [263, 265, 262] have found that an initially positive migration rate should tend toward zero when there is a spatially heterogeneous environment, as organisms will quickly migrate to and rise to high frequency in their selectively favoured environment - after which point migration will tend to be disadvantageous. A contrasting picture appears when kinship is taken into account [264], or when the environment is both spatially and temporally heterogeneous [261]. In the former case, individuals should migrate away from their natal site in order to avoid kin competition, while in the latter migration allows individuals to escape a local environment that temporarily becomes poor. Ultimately, each effect will be important, the essential point being that the results of models can depend strongly on the features of a system that a modeller subjectively chooses to incorporate. An example following a similar principle is the inclusion, or otherwise, of an Allee effect [493] in models of population spreading, which can radically alter the velocity of the wave of advance (eg. [46]). In Chapter 2, I focus specifically on the role of dispersal stochasticity given long-range dispersal in species invasion models, intentionally ignoring Allee effects, environmental heterogeneity, heritable individual variation in dispersal distance, and many other factors likely to influence wave velocity behaviour.

Topical examples that are not directly related to population biology come from economic modelling. Economic theories have historically used the representative agent approximation, in which a typical decision maker - usually a rational utility maximiser - is used to represent the diversity of agents and their strategies. Our model of cattle markets, though not a conventional economic model, can be characterised in this way, in that we only implicitly incorporate a wealth distribution in each patch, and other aspects of agent variation - such as preferences, or the probability of being a seller or buyer, or assets - are ignored. This approach often leads to economically untenable results [494], such as the absence of trade or periodic (and hence predictable) boom and bust cycles, with agent-based simulation suggested as a useful alternative [495].

Incorrect modelling assumptions may also actively contribute to sudden and severe economic crashes in several ways, as strongly argued by Colander *et al* [496]. The Black-Scholes formula [497, 498] for pricing derivatives, introduced in the early 1970s, as well as other models from mathematical finance, offered apparently rigorous guidance on appropriate strategies to hedge risk, and encouraged banks to use more leverage [499]. Unfortunately, these models often make simplifying assumptions that are known to be incorrect, and consequently make unreliable predictions. An example is the normal distribution of returns in the Black-Scholes model, which does not capture the relatively high frequency of extreme market fluctuations actually observed [496, 500]. Interestingly, the details of when modelling assumptions are violated can depend on the penetrance of a specific model among traders. When many agents follow similar strategies, their behaviour becomes correlated, which may promote wilder swings in the market. It has been suggested that the October 1987 US stock market crash, in which US stock fell in value by 20% in a day, offers an example, with automated hedging strategies, built on models like the Black-Scholes equation, leading to a feedback of sell orders [496].

Modelling, including simulation, is also used to justify the introduction of new financial products [496]. Given the substantial simplifications used in models, the relationship between their results and the expected behaviour of a financial product is unclear. An example might be the optimistic credit ratings assigned to tranches of subprime mortgages arising, in part, due to models taking insufficient notice of correlations between mortgage defaults [501, 502]. While the detailed application of models in population biology does have real world relevance - integro-difference models, which poorly represent finite populations, have recently been recommended as an approach to modelling species invasions [105] - the suggested role of inappropriate models in contributing to the stock market crashes offers a particularly dramatic illustration of the dangers of incorrect assumptions.

Drawing on this, I emphasise another important quality of mathematical and simulation models. While verbal models lend themselves to qualitative conclusions, the output of simulation models is usually quantitative. It is often difficult to assess the range of error of this quantitative output - especially given that the assumptions of the model are likely to deviate from reality in unknown ways, with unknown impacts on results. In the context of mathematical analysis of models, it is possible to have a comprehensive understanding of a model with a limited and uncertain resemblance relationship to the real world. In the context of financial models, it has been suggested that this can lead to a *control illusion* [496], whereby workers are liable to overestimate their ability to influence a target system in a predictable way based on apparently detailed modelling results. The same is true for our understanding of evolutionary and ecological systems through models.

This principle is especially relevant to our interpretation of the cattle market modelling work presented in Chapter 3. Here, the model is useful in illustrating the potential for migration biases, possibly through livestock markets, to distort the course of evolution in unexpected ways. It also suggests certain narratives of population change and patterns of trading that may be of especial interest if identified in the real world. However, I have been careful to emphasise that this representation is just one of many possible representations. My results should be taken only as a guide to the possible outcomes of biased migration. The system I have chosen to investigate - market-mediated gene flow - has not been subject to rigorous investigation, and as such this early modelling work remains useful in promoting and directing future work.

5.3 Equifinality of models

In my discussion of movement models, I indicated that an asymptotically constant velocity wave of advance arises from a range of models, including the Fisher-Kolmogorov reaction-diffusion equation, Eq. (2.1), and integro-difference models when the dispersal kernel is not fat-tailed [46]. The generality of this asymptotically constant velocity result, and its value, has been explored mathematically (eg. [77]), and I have previously related this finding to the concept of equifinality (when a single end state is reached through multiple processes, or system trajectories, or initial conditions). When a modelling result is not disrupted by changes to the modelling assumptions, it might be considered robust. Equifinality may be important either in representations of a target system, in a target system itself, or in both. The details of which is true depend on the detailed relationship between a target system and its representation. For example, models that display equifinality may all neglect to represent a critical property of the target system which leads to far greater diversity in behaviour.

The concept of equifinality is important at several points in this thesis. The coalescent framework that I use to simulate realistic genetic samples from a population is considered robust to a wide range of model details (see §4.8.1), such that multiple different demographic models will produce sample sequences with tree topologies following near-identical distributions of branching order and branch length. Again, this can be seen as model equifinality - the question of which demographic histories in actual evolving populations yield genetically similar samples will be more complex, as various assumptions of the coalescent may be broken. The wide range of processes that can generate Lévy flight patterns of movement also display equifinality, although, as discussed in §2.4.3, the specific manner in which a Lévy flight dispersal kernel is generated is likely to be important in determining the results of species invasion models.

As already mentioned, robustness of results to different modelling assumptions is an example of equifinality. However, when a specific result displays such robustness, it is possible that the model generating it has little resemblance to a target system. An example might be the Fisher-Kolmogorov model of species invasions. The constant wave velocity can be determined from real species prevalence data. Using Eq. (2.4), a suggested value of diffusivity or the maximum growth rate can be used to derive D or α. However, there is no certainty that these parameters provide widely useful descriptions of species behaviour, and the value of D, for example, derived from one species invasion might be incorrect in another context. When introducing his results on the generality of mathematical results concerning the asymptotic behaviour of the Fisher-Kolmogorov equation, Weinberger makes a related point, noting that “the other side of robustness, of course, is that a model which is found to predict [the asymptotic behaviour of an ecological system] correctly may be far from being a good model for predicting other phenomena” [77].

It is interesting that both the sensitivity of modelling results to assumptions, as explored in all three chapters of this thesis, and their robustness can lead to challenges in interpreting model findings. The more robust a finding is, the less likely that the specific design of a model is informative regarding the causal relations in the target system, while the more sensitive a finding is to variations in modelling assumptions the less likely that it will be relevant to the target system.

5.4 Closing remarks and conclusion

In this closing chapter of my thesis, I have argued that the application of modelling faces several challenges. In particular, both my work and examples from theoretical population biology and economics suggest that modelling results can be highly dependent on modelling assumptions, sometimes in relatively subtle ways. This leads to a potentially dangerous situation in which comprehensive mathematical characterisation of a specific model, or detailed quantitative simulations, can lead to a potentially fragile illusion of understanding a target system. When a model finding is known to be robust, such that a range of models with different assumptions display equifinality in their results, a separate issue can arise whereby the details of model design are relatively uninformative about the details of the target system. This can lead to difficulties in interpreting parameters biologically and extrapolating model findings.

These points do not invalidate the use of mathematical and computational models. The logical exactness of such models make them useful conceptual tools, helping to formalise the assumptions of verbal arguments and facilitating the comparison of competing interpretations of the world. Indeed, I suggest that simple simulations especially offer one route through which challenges of assessing robustness and investigating the detailed relationship between a model and the target system can be explored. In this thesis, I have presented three examples of how this might be achieved in population biology, using a range of techniques to investigate the implications of different assumptions used in different contexts for different reasons. As a result of this work, I have identified novel results of relevance to models of species invasions, the impact of migration biases on genetic variation, and the refinement of statistics designed to detect natural selection in population genetic data. That these findings were previously unknown emphasises the scientific value of using simulations to clarify the impact of modelling assumptions in population biology.

Bibliography

[1] G. Jacobs and T. Sluckin. Long-range dispersal, stochasticity and the broken accelerating wave of advance. *Theoretical Population Biology*, 100:39–55, 2015.

[2] F. J. Clemente, A. Cardona, C. E. Inchley, B. M. Peter, G. Jacobs, et al. A selective sweep on a deleterious mutation in *CPT1A* in Arctic populations. *The American Journal of Human Genetics*, 95(5):584–589, 2014.

[3] G. Ewing and J. Hermisson. MSMS: A coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, 26(16):2064–2065, 2010.

[4] I. Shlyakhter, P. C. Sabeti, and S. F. Schaffner. Cosi2: An efficient simulator of exact and approximate coalescent with selection. *Bioinformatics*, 30(23):3427–3429, 2014.

[5] G. Kaiping, G. Jacobs, S. Cox, and T. Sluckin. Nonequivalence of updating rules in evolutionary games under high mutation rates. *Physical Review E*, 90(4):042726, 2014.

[6] L. Pagani, D. Lawson, E. Jagoda, A. Mörseburg, M. Mitt, et al. Geographical barriers, environmental challenges, and complex migration events during the peopling of Eurasia. In submission as of 09/2015. I contributed some simulation work to this paper and am 20$^{\text{th}}$ on the 96-member author list., 2015.

[7] T. Malthus. *An Essay of the Principle of Population*. St. Paul's Churchyard, London: J. Johnson, 1798. Retrieved from http://www.esp.org/books/malthus/population/malthus.pdf 27/09/2015.

[8] P. F. Verhulst. Notice sur la loi que la population poursuit dans son accroissement. *Correspondance Mathématique et Physique*, 10:113–121, 1838.

[9] A. J. Lotka. *Elements of physical biology*. Williams & Wilkins company, Baltimore, 1925.

[10] K. Pearson and J. Blakeman. *A mathematical theory of random migration*, volume 15. Cambridge University Press, 1906.

[11] R. A. Fisher. The wave of advance of advantageous genes. *Annals of Eugenics*, 7(4):355–369, 1937.

[12] R. A. Fisher. *The genetical theory of natural selection.* Clarendon Press, 1930.

[13] S. Wright. Evolution in Mendelian populations. *Genetics*, 16(2):97, 1931.

[14] J. B. S. Haldane. *The causes of evolution.* Longmans, Green and Co., 1932.

[15] R. Levins. The strategy of model building in population biology. *American scientist*, pages 421–431, 1966.

[16] N. Oreskes, K. Shrader-Frechette, and K. Belitz. Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, 263(5147):641–646, 1994.

[17] M. R. Lissack and K. A. Richardson. When modeling social systems, models$\neq$ the modeled: Reacting to Wolfram's "A New Kind of Science". *Emergence, A Journal of Complexity Issues in Organizations and Management*, 3(4):95–111, 2001.

[18] W. Wagner, E. Fisher, and P. Pascual. Misunderstanding models in environmental and public health regulation. *NYU Envtl. LJ*, 18:293, 2010.

[19] R. Frigg and S. Hartmann. Models in science. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Fall 2012 edition, 2012.

[20] R. Frigg. Models and fiction. *Synthese*, 172(2):251–268, 2010.

[21] R. N. Giere. Why scientific models should not be regarded as works of fiction. In M. Suárez, editor, *Fictions in Science: Philosophical Essays on Modelling and Idealization*, pages 248–258. Routledge, 2009.

[22] P. Suppes. A comparison of the meaning and uses of models in mathematics and the empirical sciences. Technical Report 33, Stanford University, August 1961.

[23] M. W. Wartofsky. The model muddle: Proposals for an immodest realism. In *Models*, pages 1–11. Springer, 1979.

[24] G. E. Box. Robustness in the strategy of scientific model building. *Robustness in statistics*, 1:201–236, 1979.

[25] A. Maria. Introduction to modeling and simulation. In *Proceedings of the 29th conference on Winter simulation*, pages 7–13. IEEE Computer Society, 1997.

[26] R. N. Giere. How models are used to represent reality. *Philosophy of Science*, 71(5):742–752, 2004.

[27] S. French. A model-theoretic account of representation (Or, I don't know much about art... but I know it involves isomorphism). *Philosophy of Science*, 70(5):1472–1483, 2003.

[28] K. Craik. *The nature of exploration.* Cambridge, England: Cambridge University Press, 1943.

[29] P. Johnson-Laird. The history of mental models. In K. Manktelow and M. C. Chung, editors, *Psychology of reasoning: Theoretical and historical perspectives*, pages 179–212. New York: Psychology Press, 2004.

[30] P. N. Johnson-Laird. Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, 107(43):18243–18250, 2010.

[31] M. D. Braine. On the relation between the natural logic of reasoning and standard logic. *Psychological Review*, 85(1):1, 1978.

[32] M. Oaksford and N. Chater. The probabilistic approach to human reasoning. *Trends in cognitive sciences*, 5(8):349–357, 2001.

[33] C. Swoyer. Structural representation and surrogative reasoning. *Synthese*, 87(3):449–508, 1991.

[34] L. Boltzmann. On the fundamental principles and equations of mechanics. In B. McGuinness, editor, *Ludwig Boltzmann: Theoretical Physics and Philosophical Problems*, pages 101–128. Boston: Reidel, 1974.

[35] R. I. Hughes. Models and representation. *Philosophy of Science*, pages S325–S336, 1997.

[36] P. Teller. Twilight of the perfect model model. *Erkenntnis*, 55(3):393–415, 2001.

[37] S. H. Orzack. The philosophy of modelling or does the philosophy of biology have any use? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1586):170–180, 2012.

[38] J. B. S. Haldane. A defense of beanbag genetics. *Perspectives in Biology and Medicine*, 7(3):343–360, 1964.

[39] J. F. Crow. Commentary: Haldane and beanbag genetics. *International Journal of Epidemiology*, 37(3):442–445, 2008.

[40] W. J. Ewens. Commentary: On Haldane's 'Defense of beanbag genetics'. *International Journal of Epidemiology*, 37(3):447–451, 2008.

[41] D. Hume. *A treatise of human nature.* Courier Corporation, Dover edition, 2012. Book 1, Part 3, Section 1. First published in three volumes, 1738-1740.

[42] G. De Pierris. Hume's skepticism and inductivism concerning space and geometry. In *Mathematizing Space*, pages 255–274. Springer, 2015.

[43] K. J. Arrow. *Mathematical models in the social sciences.* Cowles Commission for Research in Economics, 1952.

[44] R. H. Johnson. Informal logic and its contribution to argumentation theory. In H. J. Ribeiro, editor, *Inside arguments: Logic and the study of argumentation*. Cambridge Scholars Publishing, 2012.

[45] A. N. Whitehead and B. Russell. *Principia Mathematica*, volume 1. University Press, 1910.

[46] M. Kot, M. A. Lewis, and P. van den Driessche. Dispersal data and the spread of invading organisms. *Ecology*, 77(7):2027–2042, 1996.

[47] M. Krkošek, J. S. Lauzon-Guay, and M. A. Lewis. Relating dispersal and range expansion of California sea otters. *Theoretical Population Biology*, 71(4):401–407, 2007.

[48] D. G. Kendall. Mathematical models of the spread of infection. In *Mathematics and computer science in biology and medicine*, pages 213–225. Medical Research Council London, 1965.

[49] V. Méndez, T. Pujol, and J. Fort. Dispersal probability distributions and the wave-front speed problem. *Phys. Rev. E*, 65(4; Part 1):041109, 2002.

[50] S. H. Orzack and E. Sober. A critical assessment of Levins's "The strategy of model building in population biology" (1966). *Quarterly Review of Biology*, pages 533–546, 1993.

[51] M. Weisberg. Forty years of 'The strategy': Levins on model building and idealization. *Biology and Philosophy*, 21(5):623–645, 2006.

[52] R. Levins. A response to Orzack and Sober: Formal analysis and the fluidity of science. *Quarterly Review of Biology*, pages 547–555, 1993.

[53] J. Odenbaugh. The strategy of "The strategy of model building in population biology". *Biology and Philosophy*, 21(5):607–621, 2006.

[54] K. E. Watt. Use of mathematics in population ecology. *Annual Review of Entomology*, 7(1):243–260, 1962.

[55] M. R. Evans, K. J. Norris, and T. G. Benton. Predictive ecology: Systems approaches. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1586):163–169, 2012.

[56] Z. Pirtle, R. Meyer, and A. Hamilton. What does it mean when climate models agree? A case for assessing independence among general circulation models. *Environmental Science & Policy*, 13(5):351–361, 2010.

[57] C. Tebaldi and R. Knutti. The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 365(1857):2053–2075, 2007.

[58] G. Abramowitz and H. Gupta. Toward a model space and model independence metric. *Geophysical Research Letters*, 35(5), 2008.

[59] D. Mollison. Spatial contact models for ecological and epidemic spread. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 283–326, 1977.

[60] J. E. Bailey. Mathematical modeling and analysis in biochemical engineering: Past accomplishments and future opportunities. *Biotechnology Progress*, 14(1):8–20, 1998.

[61] E. E. Leamer. Let's take the con out of econometrics. *The American Economic Review*, pages 31–43, 1983.

[62] E. E. Leamer. Sensitivity analyses would help. *The American Economic Review*, pages 308–313, 1985.

[63] N. Smith. Simulation article. In *Encyclopedia of Computer Science*, pages 1578–1587. John Wiley and Sons Ltd, Chichester, UK, 2003.

[64] J. Banks. Principles of simulation. In J. Banks, editor, *Handbook of simulation*, pages 3–30. John Wiley and Sons, Inc., 1998.

[65] E. Winsberg. Computer simulations in science. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Summer 2015 edition, 2015.

[66] Y. Frégnac and G. Laurent. Neuroscience: Where is the brain in the human brain project? *Nature*, 513(7516):27–9, 2014.

[67] R. Boumans, R. Costanza, J. Farley, M. A. Wilson, R. Portela, et al. Modeling the dynamics of the integrated earth system and the value of global ecosystem services using the GUMBO model. *Ecological Economics*, 41(3):529–560, 2002.

[68] J. W. Forrester. *World dynamics*, volume 59. Wright-Allen Press Cambridge, MA, 1971.

[69] R. Boumans, J. Roman, I. Altman, and L. Kaufman. The Multiscale Integrated Model of Ecosystem Services (MIMES): Simulating the interactions of coupled human and natural systems. *Ecosystem Services*, 12:30–41, 2015.

[70] V. Grimm. Ten years of individual-based modelling in ecology: What have we learned and what could we learn in the future? *Ecological Modelling*, 115(2):129–148, 1999.

[71] A. J. McLane, C. Semeniuk, G. J. McDermid, and D. J. Marceau. The role of agent-based models in wildlife ecology and management. *Ecological Modelling*, 222(8):1544–1556, 2011.

[72] J. Odenbaugh. The “structure” of population ecology: Philosophical reflections on unstructured and structured models. In K. Cuddington and B. E. Beisner, editors, *Ecological Paradigms Lost*, pages 63–77. Elsevier, 2005.

[73] J. D. Millington, D. O’Sullivan, and G. L. Perry. Model histories: Narrative explanation in generative simulation modelling. *Geoforum*, 43(6):1025–1034, 2012.

[74] P. E. Hulme. Trade, transport and trouble: Managing invasive species pathways in an era of globalization. *Journal of Applied Ecology*, 46(1):10–18, 2009.

[75] R. Guimerà, S. Mossa, A. Turtschi, and L. A. N. Amaral. The worldwide air transportation network: Anomalous centrality, community structure , and cities’ global roles. *Proc. Natl. Acad. Sci. USA*, 102(22):7794–7799, 2005.

[76] K. Khan, J. Arino, W. Hu, P. Raposo, J. Sears, et al. Spread of a novel influenza A (H1N1) virus via global airline transportation. *New England Journal of Medicine*, 361(2):212–214, 2009.

[77] H. Weinberger. Long-time behavior of a class of biological models. *SIAM Journal on Mathematical Analysis*, 13(3):353–396, 1982.

[78] R. R. Veit and M. A. Lewis. Dispersal, population growth, and the Allee effect: dynamics of the house finch invasion of eastern North America. *American Naturalist*, pages 255–274, 1996.

[79] M. G. Neubert and H. Caswell. Demography and dispersal: Calculation and sensitivity analysis of invasion speed for structured populations. *Ecology*, 81(6):1613–1628, 2000.

[80] F. Takasu, N. Yamamoto, K. Kawasaki, K. Togashi, Y. Kishi, et al. Modeling the expansion of an introduced tree disease. *Biological Invasions*, 2(2):141–150, 2000.

[81] D. R. Lockwood, A. Hastings, and L. W. Botsford. The effects of dispersal patterns on marine reserves: Does the tail wag the dog? *Theoretical Population Biology*, 61(3):297–309, 2002.

[82] P. Schofield. Spatially explicit models of Turelli-Hoffmann *Wolbachia* invasive wave fronts. *Journal of Theoretical Biology*, 215(1):121–131, 2002.

[83] S. Dewhirst and F. Lutscher. Dispersal in heterogeneous habitats: Thresholds, spatial scales, and approximate rates of spread. *Ecology*, 90(5):1338–1345, 2009.

[84] V. Méndez, D. Campos, and F. Bartumeus. *Stochastic Foundations in Movement Ecology*. Springer, 2013.

[85] A. Kolmogorov, I. Petrovsky, and N. Piskunov. Etude de l’équation de la diffusion avec croissance de la quantité de matiere et son application a un probleme biologique. *Mosc. Univ. Bull. Math*, 1:1–25, 1937.

[86] J. G. Skellam. Random dispersal in theoretical populations. *Biometrika*, 38:196–218, 1951.

[87] J.-P. Bouchaud and A. Georges. Anomalous diffusion in disordered media: Statistical mechanisms, models and physical applications. *Physics Reports*, 195(4):127–293, 1990.

[88] J. Brown and M. Hovmøller. Aerial dispersal of pathogens on the global and continental scales and its impact on plant disease. *Science*, 297(5581):537–441, 2002.

[89] J. Bullock and R. Clark. Long distance seed dispersal by wind: Measuring and modelling the tail of the curve. *Oecologia*, 124(4):506–521, 2000.

[90] G. D. Sutherland, A. S. Harestad, K. Price, and K. P. Lertzman. Scaling of natal dispersal distances in terrestrial birds and mammals. *Conserv. Ecol.*, 4(1):16, 2000.

[91] P. Lévy. *Théorie de l'addition des variables aléatoires.* Paris: Gauthier-Villars, 1937.

[92] B. Hughes, M. Shlesinger, and E. Montroll. Random walks with self-similar clusters. *Proc. Natl. Acad. Sci. USA*, 78(6):3287–3291, 1981.

[93] D. del Castillo-Negrete, B. Carreras, and V. Lynch. Front dynamics in reaction-diffusion systems with Lévy flights: A fractional diffusion approach. *Physical Review Letters*, 91(1):018302, 2003.

[94] G. M. Viswanthan, V. Afanasyev, S. V. Buldyrev, E. J. Murphy, P. A. Prince, et al. Lévy flight search patterns of wandering albatross. *Nature*, 381(6581):413–415, 1996.

[95] D. Mollison. The rate of spatial propagation of simple epidemics. In L. Le Cam, J. Neyman, and E. Scott, editors, *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, vol 3*, pages 579–614, 1972.

[96] S. Fedotov. Front propagation into an unstable state of reaction-transport systems. *Physical Review Letters*, 86(5):926, 2001.

[97] R. Mancinelli, D. Vergni, and A. Vulpiani. Superfast front propagation in reactive systems with non-Gaussian diffusion. *Europhysics Letters*, 60(4):532–538, 2002.

[98] E. Hanert. Front dynamics in a two-species competition model driven by Lévy flights. *Journal of Theoretical Biology*, 300:134–142, 2012.

[99] D. A. Andow, P. M. Kareiva, S. A. Levin, and A. Okubo. Spread of invading organisms: Patterns of spread. In K. C. Kim and B. A. McPherson, editors,

Evolution of insect pests: The pattern of variations, pages 219–241. New York: John Wiley and Sons, 1993.

[100] R. N. Mack. Invasion of *Bromus tectorum* L. into western North America: An ecological chronicle. *Agro-Ecosystems*, 7:145–165, 1981.

[101] E. Weber. The dynamics of plant invasions: A case study of three exotic goldenrod species (*Solidago* L.) in Europe. *Journal of Biogeography*, 25(1):147–154, 1998.

[102] N. Shigesada and K. Kawasaki. *Biological Invasions: Theory and Practice.* Oxford University Press, 1997.

[103] C. C. Mundt, K. E. Sackett, L. D. Wallace, C. Cowger, and J. P. Dudley. Long-distance dispersal and accelerating waves of disease: Empirical relationships. *The American Naturalist*, 173(4):456–466, 2009.

[104] J. Medlock and M. Kot. Spreading disease: Integro-differential equations old and new. *Mathematical Biosciences*, 184(2):201–222, 2003.

[105] C. Robinet, H. Kehlenbeck, D. J. Kriticos, R. H. Baker, A. Battisti, et al. A suite of models to support the quantitative assessment of spread in pest risk analysis. *PLOS ONE*, 7(10):e43366, 2012.

[106] J. Clark, M. Lewis, and L. Horvath. Invasion by extremes: Population spread with variation in dispersal and reproduction. *American Naturalist*, 157(5):537–554, 2001.

[107] R. E. Snyder. How demographic stochasticity can slow biological invasions. *Ecology*, 84(5):1333–1339, 2003.

[108] D. Brockmann and L. Hufnagel. Front propagation in reaction-superdiffusion dynamics: Taming Lévy flights with fluctuations. *Phys. Rev. Lett.*, 98:178301, Apr 2007.

[109] K. Kawasaki, F. Takasu, H. Caswell, and N. Shigesada. How does stochasticity in colonization accelerate the speed of invasion in a cellular automaton model? *Ecological Research*, 21(3):334–345, 2006.

[110] M. Kot, J. Medlock, T. Reluga, and D. B. Walton. Stochasticity, invasions, and branching random walks. *Theoretical Population Biology*, 66(3):175–184, 2004.

[111] M. Lewis and S. Pacala. Modeling and analysis of stochastic invasion processes. *Journal of Mathematical Biology*, 41(5):387–429, 2000.

[112] E. Brunet and B. Derrida. Shift in the velocity of a front due to a cutoff. *Physical Review E*, 56:2597–2604, 1997.

[113] E. Cohen and D. A. Kessler. Front propagation dynamics with exponentially-distributed hopping. *Journal of Statistical Physics*, 122(5):925–948, 2006.

[114] F. Bartumeus, L. Giuggioli, M. Louzao, V. Bretagnolle, D. Oro, et al. Fishery discards impact on seabird movement patterns at regional scales. *Current Biology*, 20(3):215–222, 2010.

[115] N. E. Humphries, N. Queiroz, J. R. Dyer, N. G. Pade, M. K. Musyl, et al. Environmental context explains Lévy and Brownian movement patterns of marine predators. *Nature*, 465(7301):1066–1069, 2010.

[116] A. Mashanova, T. H. Oliver, and V. A. Jansen. Evidence for intermittency and a truncated power law from highly resolved aphid movement data. *Journal of The Royal Society Interface*, 7(42):199–208, 2010.

[117] H. Kierstead and L. B. Slobodkin. The size of water masses containing plankton blooms. *J. Mar. Res*, 12(1):141–147, 1953.

[118] D. Campos, V. Méndez, and V. Ortega-Cejas. Lattice models for invasions through patchy environments. *Bulletin of Mathematical Biology*, 70(7):1937–1956, 2008.

[119] A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[120] F. van den Bosch, J. Metz, and O. Diekmann. The velocity of spatial population expansion. *Journal of Mathematical Biology*, 28(5):529–565, 1990.

[121] D. Mollison. Dependence of epidemic and population velocities on basic parameters. *Mathematical Biosciences*, 107(2):255–287, 1991.

[122] G. Dee and J. S. Langer. Propagating pattern selection. *Phys. Rev. Lett.*, 50(6):383–386, 1983.

[123] K. Binder and D. Stauffer. A Simple Introduction to Monte Carlo Simulation and Some Specialized Topics (With 6 Figures). In K. Binder, editor, *Applications of the Monte Carlo Method in Statistical Physics*, page 1, 1984.

[124] W. van Saarloos. Front propagation into unstable states: Marginal stability as a dynamical selection mechanism for velocity selection. *Phys. Rev. A*, 37(1):211–229, 1988.

[125] M. T. Subbotin. On the law of frequency of error. *Matematicheskii Sbornik*, 31(2):296–301, 1923.

[126] M. Shlesinger, G. Zaslavsky, and U. Frisch, editors. *Lévy flights and related topics in physics*, volume 450 of *Lecture Notes in Physics*. Springer Berlin Heidelberg, 1995.

[127] J. M. Travis, C. M. Harris, K. J. Park, and J. M. Bullock. Improving prediction and management of range expansions by combining analytical and individual-based modelling approaches. *Methods in Ecology and Evolution*, 2(5):477–488, 2011.

[128] D. del Castillo-Negrete. Truncation effects in superdiffusive front propagation with Lévy flights. *Physical Review E*, 79(3):031120, 2009.

[129] J. Gibbs. Fourier's series. *Nature*, 59:200, 1898.

[130] M. Plank and M. Simpson. Models of collective cell behaviour with crowding effects: comparing lattice-based and lattice-free approaches. *J. Roy. Soc. Interface*, 9(76):2983–2996, 2012.

[131] H. P. Grossart, C. Dziallas, F. Leunert, and K. W. Tang. Bacteria dispersal by hitchhiking on zooplankton. *Proc. Natl. Acad. Sci. USA*, 107(26):11959–11964, 2010.

[132] B. Vanschoenwinkel, A. Waterkeyn, T. Nhiwatiwa, T. Pinceel, E. Spooren, et al. Passive external transport of freshwater invertebrates by elephant and other mud-wallowing mammals in an African savannah habitat. *Freshwater Biology*, 56(8):1606–1619, 2011.

[133] J. S. Clark, M. Silman, R. Kern, E. Macklin, and J. HilleRisLambers. Seed dispersal near and far: Generalized patterns across temperate and tropical forests. *Ecology*, 80:1475–1494, 1999.

[134] O. Hallatschek and D. S. Fisher. Acceleration of evolutionary spread by long-range dispersal. *Proceedings of the National Academy of Sciences*, 111(46):E4911–E4919, 2014.

[135] S. Chatterjee and P. S. Dey. Multiple phase transitions in long-range first-passage percolation on square lattices. *Communications on Pure and Applied Mathematics*, 2015.

[136] D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[137] J. M. Fryxell, M. Hazell, L. Börger, B. D. Dalziel, D. T. Haydon, et al. Multiple movement modes by large herbivores at multiple spatiotemporal scales. *Proceedings of the National Academy of Sciences*, 105(49):19114–19119, 2008.

[138] G. Viswanathan, S. V. Buldyrev, S. Havlin, M. Da Luz, E. Raposo, et al. Optimizing the success of random searches. *Nature*, 401(6756):911–914, 1999.

[139] A. M. Reynolds. Beating the odds in the aerial lottery: Passive dispersers select conditions at takeoff that maximize their expected fitness on landing. *The American Naturalist*, 181(4):555–561, 2013.

[140] G. H. Pyke. Understanding movements of organisms: It's time to abandon the Lévy foraging hypothesis. *Methods in Ecology and Evolution*, 6(1):1–16, 2015.

[141] A. Reynolds. Liberating Lévy walk research from the shackles of optimal foraging. *Physics of Life Reviews*, 2015.

[142] R. Kelly, M. G. Lundy, F. Mineur, C. Harrod, C. A. Maggs, et al. Historical data reveal power-law dispersal patterns of invasive aquatic species. *Ecography*, 37(6):581–590, 2014.

[143] C. Hui, N. Roura-Pascual, L. Brotons, R. A. Robinson, and K. L. Evans. Flexible dispersal strategies in native and non-native ranges: Environmental quality and the 'good–stay, bad–disperse' rule. *Ecography*, 35(11):1024–1032, 2012.

[144] B. D. Fitt, P. Gregory, A. Todd, H. McCartney, and O. Macdonald. Spore dispersal and plant disease gradients; a comparison between two empirical models. *Journal of Phytopathology*, 118(3):227–242, 1987.

[145] E. A. Fronhofer, E. B. Sperr, A. Kreis, M. Ayasse, H. J. Poethke, et al. Picky hitch-hikers: Vector choice leads to directed dispersal and fat-tailed kernels in a passively dispersing mite. *Oikos*, 122(8):1254–1264, 2013.

[146] A. J. Green, M. I. Sánchez, F. Amat, J. Figuerola, F. Hontoria, et al. Dispersal of invasive and native brine shrimps *Artemia* (Anostraca) via waterbirds. *Limnology and Oceanography*, 50(2):737–742, 2005.

[147] M. Ansong and C. Pickering. Are weeds hitchhiking a ride on your car? A systematic review of seed dispersal on cars. *PLoS ONE*, 8(11):e80275, 2013.

[148] D. Barnes and P. Milner. Drifting plastic and its consequences for sessile organism dispersal in the Atlantic Ocean. *Marine Biology*, 146(4):815–825, 2005.

[149] A. Locke, D. Reid, H. Van Leeuwen, W. Sprules, and J. Carlton. Ballast wafer exchange as a means of controlling dispersal of freshwater organisms by ships. *Canadian Journal of Fisheries and Aquatic Sciences*, 50(10):2086–2093, 1993.

[150] D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006.

[151] I. Rhee, M. Shin, S. Hong, K. Lee, S. J. Kim, et al. On the Lévy-walk nature of human mobility. *IEEE/ACM Transactions on Networking (TON)*, 19(3):630–643, 2011.

[152] P. Marmottant, A. Ponomarenko, and D. Bienaimé. The walk and jump of *Equisetum* spores. *Proceedings of the Royal Society of London B: Biological Sciences*, 280(1770):20131465, 2013.

[153] J. O. Washburn and L. Washburn. Active aerial dispersal of minute wingless arthropods: exploitation of boundary-layer velocity gradients. *Science*, 223(4640):1088–1089, 1984.

[154] J. Bell, D. Bohan, E. Shaw, and G. Weyman. Ballooning dispersal using silk: World fauna, phylogenies, genetics and models. *Bulletin of Entomological Research*, 95(02):69–114, 2005.

[155] F. J. Ferrandino. Dispersive epidemic waves: I. Focus expansion within a linear planting. *Phytopathology*, 83(8):795–802, 1993.

[156] H. Scherm. On the velocity of epidemic waves in model plant disease epidemics. *Ecological Modelling*, 87(1):217–222, 1996.

[157] A. Reynolds. Exponential and power-law contact distributions represent different atmospheric conditions. *Phytopathology*, 101(12):1465–1470, 2011.

[158] A. M. Reynolds. Signatures of active and passive optimized Lévy searching in jellyfish. *Journal of The Royal Society Interface*, 11(99):20140665, 2014.

[159] G. U. Yule. A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, pages 21–87, 1925.

[160] H. A. Simon. On a class of skew distribution functions. *Biometrika*, pages 425–440, 1955.

[161] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[162] H. Takayasu, A.-H. Sato, and M. Takayasu. Stable infinite variance fluctuations in randomly amplified Langevin systems. *Physical Review Letters*, 79(6):966, 1997.

[163] T. S. Biró and A. Jakovác. Power-law tails from multiplicative noise. *Physical Review Letters*, 94(13):132302, 2005.

[164] A. Reynolds. Can spontaneous cell movements be modelled as Lévy walks? *Physica A: Statistical Mechanics and its Applications*, 389(2):273–277, 2010.

[165] A. M. Reynolds. Bridging the gulf between correlated random walks and Lévy walks: Autocorrelation as a source of Lévy walk movement patterns. *Journal of the Royal Society Interface*, page rsif20100292, 2010.

[166] S. Petrovskii and A. Morozov. Dispersal in a statistically structured population: Fat tails revisited. *The American Naturalist*, 173(2):278–289, 2009.

[167] S. Benhamou. How many animals really do the Lévy walk? *Ecology*, 88(8):1962–1969, 2007.

[168] A. M. Reynolds. Mussels realize Weierstrassian Lévy walks as composite correlated random walks. *Scientific Reports*, 4, 2014.

[169] A. Reynolds. Animals that randomly reorient at cues left by correlated random walkers do the Lévy walk. *The American Naturalist*, 175(5):607–613, 2010.

[170] R. Taylor. The relationship between density and distance of dispersing insects. *Ecological Entomology*, 3(1):63–70, 1978.

[171] A. Rieux, S. Soubeyrand, F. Bonnot, E. K. Klein, J. E. Ngando, et al. Long-distance wind-dispersal of spores in a fungal plant pathogen: Estimation of anisotropic dispersal kernels from an extensive field experiment. *PLoS ONE*, 9(8):e103225, 2014.

[172] S. Bazazi, F. Bartumeus, J. J. Hale, and I. D. Couzin. Intermittent motion in desert locusts: Behavioural complexity in simple environments. *PLoS Computational Biology*, 8(5):e1002498, 2012.

[173] J. M. Bullock, S. J. Galsworthy, P. Manzano, P. Poschlod, C. Eichberg, et al. Process-based functions for seed retention on animals: A test of improved descriptions of dispersal using multiple data sets. *Oikos*, 120(8):1201–1208, 2011.

[174] M. C. Wichmann, M. J. Alexander, M. B. Soons, S. Galsworthy, L. Dunne, et al. Human-mediated dispersal of seeds over long distances. *Proceedings of the Royal Society of London B: Biological Sciences*, 276(1656):523–532, 2009.

[175] J.-R. Luévano. Statistical features of the stretched exponentials densities. In *Journal of Physics: Conference Series*, volume 475, page 012008. IOP Publishing, 2013.

[176] P. Kailasnath, K. Sreenivasan, and G. Stolovitzky. Probability density of velocity increments in turbulent flows. *Physical Review Letters*, 68(18):2766, 1992.

[177] R. G. Palmer, D. L. Stein, E. Abrahams, and P. W. Anderson. Models of hierarchically constrained dynamics for glassy relaxation. *Physical Review Letters*, 53(10):958, 1984.

[178] P. Grassberger and I. Procaccia. The long time properties of diffusion in a medium with static traps. *The Journal of Chemical Physics*, 77(12):6281–6284, 1982.

[179] B. Sturman, E. Podivilov, and M. Gorkunov. Origin of stretched exponential relaxation for hopping-transport models. *Physical Review Letters*, 91(17):176602, 2003.

[180] J. Klafter and M. F. Shlesinger. On the relationship among three theories of relaxation in disordered systems. *Proceedings of the National Academy of Sciences*, 83(4):848–851, 1986.

[181] U. Frisch and D. Sornette. Extreme deviations and applications. *Journal de Physique I*, 7(9):1155–1171, 1997.

[182] J. Laherrere and D. Sornette. Stretched exponential distributions in nature and economy:"fat tails" with characteristic scales. *The European Physical Journal B-Condensed Matter and Complex Systems*, 2(4):525–539, 1998.

[183] T. J. Perkins, E. Foxall, L. Glass, and R. Edwards. A scaling law for random walks on networks. *Nature Communications*, 5, 2014.

[184] V. Macaulay, C. Hill, A. Achilli, C. Rengo, D. Clarke, et al. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science*, 308(5724):1034–1036, 2005.

[185] J. M. Erlandson, M. L. Moss, and M. Des Lauriers. Life on the edge: Early maritime cultures of the Pacific Coast of North America. *Quaternary Science Reviews*, 27(23):2232–2245, 2008.

[186] P. Mellars, K. C. Gori, M. Carr, P. A. Soares, and M. B. Richards. Genetic and archaeological perspectives on the initial modern human colonization of southern Asia. *Proceedings of the National Academy of Sciences*, 110(26):10699–10704, 2013.

[187] J. Steele, J. Adams, and T. Sluckin. Modelling paleoindian dispersals. *World Archaeology*, 30(2):286–305, 1998.

[188] E. Mathias and P. Mundy. Herd movements: The exchange of livestock breeds and genes between North and South, 2005.

[189] M. Thibier and H.-G. Wagner. World statistics for artificial insemination in cattle. *Livestock Production Science*, 74(2):203–212, 2002.

[190] D. Gollin, E. Van Dusen, and H. Blackburn. Animal genetic resource trade flows: Economic assessment. *Livestock Science*, 120(3):248–255, 2009.

[191] A. Malafosse. Propagation of improved breeds: The role of artificial insemination and embryo transfer. *Revue scientifique et technique (International Office of Epizootics)*, 9(3):795–824, 1990.

[192] D. Funk. Major advances in globalization and consolidation of the artificial insemination industry. *Journal of Dairy Science*, 89(4):1362–1368, 2006.

[193] P. Taberlet, A. Valentini, H. Rezaei, S. Naderi, F. Pompanon, et al. Are cattle, sheep, and goats endangered species? *Molecular Ecology*, 17(1):275–284, 2008.

[194] C. B. Wollny. The need to conserve farm animal genetic resources in Africa: Should policy makers be concerned? *Ecological Economics*, 45(3):341–351, 2003.

[195] J. Morton and D. Barton. Destocking as a drought–mitigation strategy: Clarifying rationales and answering critiques. *Disasters*, 26(3):213–228, 2002.

[196] A. Islam and P. Maitra. Health shocks and consumption smoothing in rural households: Does microcredit have a role to play? *Journal of Development Economics*, 97(2):232–243, 2012.

[197] M. D. Turner. Capital on the move: The changing relation between livestock and labor in Mali, West Africa. *Geoforum*, 40(5):746–755, 2009.

[198] S. S. Acharya, S. A. the Working Group on Agricultural Marketing Infrastructure, P. R. for Internal, and E. Trade. Report of the working group on agricultural marketing infrastructure, secondary agriculture and policy required for internal and external trade for the xi five year plan 2007-2012, 2007.

[199] M. I. Barbaruah and V. H. S. Team. Research study: To define and analyse cross-border and in-country livestock and livestock-products market systems in India for control of transboundary animal diseases, 2013.

[200] T. Williams, B. Spycher, and I. Okike. *Improving livestock marketing and intra-regional trade in West Africa: Determining appropriate economic incentives and policy framework.* ILRI, 2006.

[201] IIRR. Moving herds, moving markets: Making markets work for African pastoralists. Technical report, International Institute for Rural Reconstruction, 2013.

[202] B. W. Dwight. *Modern philology: Its discoveries, history, and influence.* Charles Schribner, New York, 2nd series edition, 1864.

[203] F. L. Pryor. *Capitalism Reassessed.* Cambridge University Press, 2013. Appendix 2-1.

[204] M. Weber. *Economy and society: An outline of interpretive sociology*, volume 1. Univ of California Press, 1978 [1922].

[205] S. Turner. *The History of the Anglo-Saxons, Comprising the History of England from the Earliest Period to the Norman Conquest*, volume 2. Longman, Hurst, Rees, Orme, and Brown, 1823.

[206] T. Lewis. *A glossary of mediaeval Welsh law, based upon the Black book of Chirk.* Manchester University Press, 1913.

[207] P. Steinkeller. The renting of fields in early Mesopotamia and the development of the concept of" interest" in Sumerian. *Journal of the Economic and Social History of the Orient/Journal de l'Histoire Economique et Sociale de l'Orient*, pages 113–145, 1981.

[208] L. Allen. *The Encyclopedia of Money.* ABC-CLIO, 2nd edition, 2009.

[209] A. R. Burns. *Money and monetary policy in early times.* Routledge, 2013 [1927].

[210] U. Thakur. A study in barter and exchange in ancient India. *Journal of the Economic and Social History of the Orient/Journal de l'Histoire Economique et Sociale de l'Orient*, pages 297–315, 1972.

[211] F. McCormick. The decline of the cow: Agricultural and settlement change in early medieval Ireland. *Peritia*, 20(1):209–224, 2008.

[212] M. J. Herskovits. The cattle complex in East Africa. *American Anthropologist*, 28(1):230–272, 1926.

[213] G. McGuire. By coin or by kine? Barter and pastoral production in Kazakhstan. *Ethnos*, pages 1–22, 2014.

[214] B. R. Foster. Commercial activity in sargonic Mesopotamia. *Iraq*, 39(01):31–43, 1977.

[215] H. de Genouillac. *Inventaire des tablettes de Tello conservées au Musée Impérial Ottoman*, volume Tome 2, Pt. 2, chapter Textes de l'époque d'Agadé et de l'époque d'Ur, page 49. E. Leroux, 1910. Available online as of 08/2015 at https://archive.org/details/inventairedestab02consuoft.

[216] W. M. Ramsay. *Asianic Elements in Greek Civilization: The Gifford Lectures in the University of Edinburgh, 1915-16.* Ams Press, 1927.

[217] M. Chandra. *Trade and trade routes in ancient India.* Abhinav Publications, 1977.

[218] J. Tainter. *The collapse of complex societies.* Cambridge University Press, 1990.

[219] J.-D. Vigne, I. Carrere, F. Briois, and J. Guilaine. The early process of mammal domestication in the Near East. *Current Anthropology*, 52(S4):S255–S271, 2011.

[220] M. A. Zeder. Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. *Proceedings of the National Academy of Sciences*, 105(33):11597–11604, 2008.

[221] C. Knipper. chapter Mobility in a sedentary society: Insights from isotope analysis of LBK human and animal teeth, pages 142–158. 2009.

[222] A. Bogaard, E. Henton, J. Evans, K. Twiss, M. Charles, et al. Locating land use at Neolithic Çatalhöyük, Turkey: The implications of 87Sr/86Sr signatures in plants and sheep tooth sequences. *Archaeometry*, 56(5):860–877, 2014.

[223] F. Feulnera, L. Kootkerb, H. Hollundb, G. Daviesc, and O. Craiga. Combined isotope analysis indicate restricted mobility of cattle at the Neolithic causewayed enclosure of Champ-Durand, Vendée (France). In R. Joussaume, editor, *L'enceinte Néolithique de Champ-Durand à Nieul-sur-l'Autise (Vendée).*

[224] K.-G. Sjögren and T. D. Price. A complex Neolithic economy: Isotope evidence for the circulation of cattle and sheep in the TRB of western Sweden. *Journal of Archaeological Science*, 40(1):690–704, 2013.

[225] S. Viner, J. Evans, U. Albarella, and M. P. Pearson. Cattle mobility in prehistoric Britain: Strontium isotope analysis of cattle teeth from Durrington Walls (Wiltshire, Britain). *Journal of Archaeological Science*, 37(11):2812–2820, 2010.

[226] R. Madgwick, J. Mulville, and J. Evans. Investigating diagenesis and the suitability of porcine enamel for strontium (87Sr/86Sr) isotope analysis. *J. Anal. At. Spectrom.*, 27:733–742, 2012.

[227] L. Babcock and E. Arnold. Identifying merchants in the Early Bronze Age city of Tell es-Safi. isotopic analysis of a sacrificial ass. *Student Summer Scholars*, (129), 2012.

[228] M. Balasse, S. H. Ambrose, A. B. Smith, and T. D. Price. The seasonal mobility model for prehistoric herders in the south-western Cape of South Africa assessed by isotopic analysis of sheep tooth enamel. *Journal of Archaeological Science*, 29(9):917–932, 2002.

[229] C. Minniti, S. Valenzuela-Lamas, J. Evans, and U. Albarella. Widening the market. Strontium isotope analysis on cattle teeth from Owslebury (Hampshire, UK) highlights changes in livestock supply between the Iron Age and the Roman period. *Journal of Archaeological Science*, 42:305–314, 2014.

[230] J. E. Robb and R. H. Farr. Substances in motion: Neolithic Mediterranean ‘trade’. *The Archaeology of Mediterranean Prehistory*, 1:24–46, 2005.

[231] R. Carter. Boat remains and maritime trade in the Persian Gulf during the sixth and fifth millennia BC. *Antiquity*, 80(307):52–63, 2006.

[232] J. Connan and T. Van de Velde. An overview of bitumen trade in the Near East from the Neolithic (c. 8000 BC) to the early Islamic period. *Arabian Archaeology and Epigraphy*, 21(1):1–19, 2010.

[233] O. Williams-Thorpe. Obsidian in the Mediterranean and the Near East: A provenancing success story. *Archaeometry*, 37(2):217–248, 1995.

[234] E. G. Ravenstein. The laws of migration. *Journal of the Statistical Society of London*, pages 167–235, 1885.

[235] D. B. Grigg. EG Ravenstein and the laws of migration. *Journal of Historical Geography*, 3(1):41–54, 1977.

[236] G. K. Zipf. The P1 P2/D hypothesis: On the intercity movement of persons. *American Sociological Review*, pages 677–686, 1946.

[237] G. Monge. Mémoire sur la théorie des déblais et de remblais. *Histoire de lAcadémie Royale des Sciences de Paris, avec les Mémoires de Mathématique et de Physique pour la même année*, pages 666–704, 1781.

[238] S. A. Stouffer. Intervening opportunities: A theory relating mobility and distance. *American Sociological Review*, 5(6):845–867, 1940.

[239] F. Simini, M. C. González, A. Maritan, and A.-L. Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100, 2012.

[240] E. A. Codling, M. J. Plank, and S. Benhamou. Random walk models in biology. *Journal of the Royal Society Interface*, 5(25):813–834, 2008.

[241] C. S. Patlak. A mathematical contribution to the study of orientation of organisms. *The Bulletin of Mathematical Biophysics*, 15(4):431–476, 1953.

[242] P. Kareiva and N. Shigesada. Analyzing insect movement as a correlated random walk. *Oecologia*, 56(2-3):234–238, 1983.

[243] M. Kot and W. M. Schaffer. Discrete-time growth-dispersal models. *Mathematical Biosciences*, 80(1):109–136, 1986.

[244] J. Haldane. The theory of a cline. *Journal of Genetics*, 48(3):277–284, 1948.

[245] J. B. S. Haldane. A mathematical theory of natural and artificial selection.(part VI, Isolation.). In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 26, pages 220–230. Cambridge Univ Press, 1930.

[246] J. Felsenstein. The theoretical population genetics of variable selection and migration. *Annual Review of Genetics*, 10(1):253–280, 1976.

[247] H. Levene. Genetic equilibrium when more than one ecological niche is available. *American Naturalist*, pages 331–333, 1953.

[248] J. M. Smith. Sympatric speciation. *American Naturalist*, pages 637–650, 1966.

[249] M. Kimura. "stepping-stone" models of population. Technical report, Technical report 3, Institute of Genetics, Japan, 1953.

[250] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo. A tale of many cities: Universal patterns in human urban mobility. *PloS ONE*, 7(5):e37027, 2012.

[251] D. A. Raichlen, B. M. Wood, A. D. Gordon, A. Z. Mabulla, F. W. Marlowe, et al. Evidence of Lévy walk foraging patterns in human hunter–gatherers. *Proceedings of the National Academy of Sciences*, 111(2):728–733, 2014.

[252] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11623–11628, 2005.

[253] J. Clobert, M. Baguette, T. G. Benton, J. M. Bullock, and S. Ducatez. *Dispersal ecology and evolution.* Oxford University Press, 2012.

[254] R. Nathan, W. M. Getz, E. Revilla, M. Holyoak, R. Kadmon, et al. A movement ecology paradigm for unifying organismal movement research. *Proceedings of the National Academy of Sciences*, 105(49):19052–19059, 2008.

[255] S. D. Fretwell and J. S. Calver. On territorial behavior and other factors influencing habitat distribution in birds. *Acta Biotheoretica*, 19(1):37–44, 1969.

[256] N. Shigesada and E. Teramoto. A consideration on the theory of environmental density. *Japanese Journal of Ecology*, 1978.

[257] P. E. Lutscher, F., , and M. A. Lewis. The effect of dispersal patterns on stream populations. *Siam Review*, 47(4):749–772, 2005.

[258] J. E. Byers and J. M. Pringle. Going against the flow: Retention, range limits and invasions in advective environments. *Marine Ecology Progress Series*, 313:27–41, 2006.

[259] D. E. Aylor. Biophysical scaling and the passive dispersal of fungus spores: Relationship to integrated pest management strategies. *Agricultural and Forest Meteorology*, 97(4):275–292, 1999.

[260] B. J. Balkau and M. W. Feldman. Selection for migration modification. *Genetics*, 74(1):171–174, 1973.

[261] D. Roff. Population stability and the evolution of dispersal in a heterogeneous environment. *Oecologia*, 19(3):217–237, 1975.

[262] J. Bull, C. Thompson, D. Ng, and R. Moore. A model for natural selection of genetic migration. *American Naturalist*, pages 143–157, 1987.

[263] A. Hastings. Can spatial variation alone lead to selection for dispersal? *Theoretical Population Biology*, 24(3):244–251, 1983.

[264] R. M. May. Dispersal in stable habitats. *Nature*, 269(5629):578–581, 1977.

[265] R. D. Holt. Population dynamics in two-patch environments: Some anomalous consequences of an optimal habitat distribution. *Theoretical Population Biology*, 28(2):181–208, 1985.

[266] D. N. Ngoc, R. B. De La Parra, M. A. Zavala, and P. Auger. Competition and species coexistence in a metapopulation model: Can fast asymmetric migration reverse the outcome of competition in a homogeneous environment? *Journal of Theoretical Biology*, 266(2):256–263, 2010.

[267] D. I. Bolnick and S. P. Otto. The magnitude of local adaptation under genotype-dependent dispersal. *Ecology and Evolution*, 3(14):4722–4735, 2013.

[268] V. Křivan, R. Cressman, and C. Schneider. The ideal free distribution: A review and synthesis of the game-theoretic perspective. *Theoretical Population Biology*, 73(3):403–425, 2008.

[269] M. A. McPeek and R. D. Holt. The evolution of dispersal in spatially and temporally varying environments. *American Naturalist*, pages 1010–1027, 1992.

[270] R. Boyd and P. J. Richerson. Voting with your feet: Payoff biased migration and the evolution of group beneficial behavior. *Journal of Theoretical Biology*, 257(2):331 – 339, 2009.

[271] C. Zhang, J. Zhang, and G. Xie. Evolution of cooperation among game players with non-uniform migration scopes. *Chaos, Solitons & Fractals*, 59:103–111, 2014.

[272] R. Eftimie. Hyperbolic and kinetic models for self-organized biological aggregations and movement: A brief review. *Journal of Mathematical Biology*, 65(1):35–75, 2012.

[273] P. Edelaar and D. I. Bolnick. Non-random gene flow: An underappreciated force in evolution and ecology. *Trends in Ecology & Evolution*, 27(12):659–665, 2012.

[274] J. L. Richardson, M. C. Urban, D. I. Bolnick, and D. K. Skelly. Microgeographic adaptation and the spatial scale of evolution. *Trends in Ecology & Evolution*, 29(3):165–176, 2014.

[275] I. J. Wang and G. S. Bradburd. Isolation by environment. *Molecular Ecology*, 23(23):5649–5662, 2014.

[276] D. Lamouroux, J. Nagler, T. Geisel, and S. Eule. Paradoxical effects of coupling infectious livestock populations and imposing transport restrictions. *Proceedings of the Royal Society of London B: Biological Sciences*, 282(1800):20142805, 2015.

[277] R. T. Forman and L. E. Alexander. Roads and their major ecological effects. *Annual Review of Ecology and Systematics*, pages 207–C2, 1998.

[278] S. Jennings and M. J. Kaiser. The effects of fishing on marine ecosystems. *Advances in Marine Biology*, 34:201–352, 1998.

[279] E. M. Fèvre, B. M. d. C. Bronsvoort, K. A. Hamilton, and S. Cleaveland. Animal movements and the spread of infectious diseases. *Trends in Microbiology*, 14(3):125–131, 2006.

[280] L. Mansley, P. Dunlop, S. Whiteside, and R. Smith. Early dissemination of foot-and-mouth disease virus through sheep marketing in February 2001. *The Veterinary Record*, 153(2):43–50, 2003.

[281] R. J. S. Magalhães, A. Ortiz-Pelaez, K. L. Thi, Q. H. Dinh, J. Otte, et al. Associations between attributes of live poultry trade and HPAI H5N1 outbreaks: A descriptive and network analysis study in northern Vietnam. *BMC Veterinary Research*, 6(1):10, 2010.

[282] M. D. Van Kerkhove, S. Vong, J. Guitian, D. Holl, P. Mangtani, et al. Poultry movement networks in Cambodia: Implications for surveillance and control of highly pathogenic avian influenza (hpai/h5n1). *Vaccine*, 27(45):6345–6352, 2009.

[283] M. Gilbert, A. Mitchell, D. Bourn, J. Mawdsley, R. Clifton-Hadley, et al. Cattle movements and bovine tuberculosis in Great Britain. *Nature*, 435(7041):491–496, 2005.

[284] D. M. Green, I. Z. Kiss, A. P. Mitchell, and R. R. Kao. Estimates for local and movement-based transmission of bovine tuberculosis in British cattle. *Proceedings of the Royal Society B: Biological Sciences*, 275(1638):1001–1005, 2008.

[285] R. Kao, L. Danon, D. Green, and I. Kiss. Demographic structure and pathogen dynamics on the network of livestock movements in Great Britain. *Proceedings of the Royal Society B: Biological Sciences*, 273(1597):1999–2007, 2006.

[286] S. Robinson, M. Everett, and R. Christley. Recent network evolution increases the potential for large epidemics in the British cattle population. *Journal of the Royal Society Interface*, 4(15):669–674, 2007.

[287] V. V. Volkova, R. Howey, N. J. Savill, and M. E. Woolhouse. Sheep movement networks and the transmission of infectious diseases. *PLoS One*, 5(6):e11185, 2010.

[288] International fund for agricultural development. http://www.ifad.org/index.htm, Accessed: 05-06-2015. Searching website for 'artificial insemination' gave 1320 results, including many case studies describing the application of artificial insemination as a development tool.

[289] Artificial insemination in armenian village producing positive results. http://www.heifer.org/join-the-conversation/blog/2013/April/artificial-insemination-in-armenian-village-producing-positive-results.html, Accessed: 01-03-2015.

[290] C. Peacock. Dairy goat development in East Africa: A replicable model for smallholders? *Small Ruminant Research*, 77(2):225–238, 2008.

[291] M. Hamid and K. Hossain. Role of private sector in the development of dairy industry in Bangladesh. *Growth*, 3:22–5, 2002.

[292] N. Marongwe and K. Chatiza. Evaluation of the livestock fairs intervention project in Zimbabwe. Technical report, Oxfam, 2007. http://policy-practice.oxfam.org.uk/publications/evaluation-of-the-livestock-fairs-intervention-project-in-zimbabwe-119464.

[293] W. F. Bodmer and L. L. Cavalli-Sforza. A migration matrix model for the study of random genetic drift. *Genetics*, 59(4):565, 1968.

[294] H. Steinfeld, T. Wassenaar, and S. Jutzi. Livestock production systems in developing countries: status, drivers, trends. *Rev Sci Tech*, 25(2):505–516, 2006.

[295] G. Seidel. Economics of selecting for sex: The most important genetic trait. *Theriogenology*, 59(2):585–598, 2003.

[296] P. K. Thornton. Livestock production: Recent trends, future prospects. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1554):2853–2867, 2010.

[297] M. Fafchamps, C. Udry, and K. Czukas. Drought and saving in West Africa: Are livestock a buffer stock? *Journal of Development Economics*, 55(2):273–305, 1998.

[298] M. D. Turner and T. O. Williams. Livestock market dynamics and local vulnerabilities in the Sahel. *World Development*, 30(4):683–705, 2002.

[299] M. Verpoorten. Household coping in war-and peacetime: Cattle sales in Rwanda, 1991–2001. *Journal of Development Economics*, 88(1):67–86, 2009.

[300] R. Scarpa, E. S. Ruto, P. Kristjanson, M. Radeny, A. G. Drucker, et al. Valuing indigenous cattle breeds in Kenya: An empirical comparison of stated and revealed preference value estimates. *Ecological Economics*, 45(3):409–426, 2003.

[301] E. Ouma, A. Abdulai, and A. Drucker. Measuring heterogeneous preferences for cattle traits among cattle-keeping households in East Africa. *American Journal of Agricultural Economics*, 89(4):1005–1019, 2007.

[302] D. J. Nagurney, A. and D. Zhang. A supply chain network equilibrium model. *Transportation Research Part E: Logistics and Transportation Review*, 38(5):281 – 303, 2002.

[303] R. E. Kranton and D. F. Minehart. A theory of buyer-seller networks. In *Networks and Groups*, pages 347–378. Springer, 2003.

[304] F. Nava. Efficiency in decentralized oligopolistic markets. *Journal of Economic Theory*, 157(0):315 – 348, 2015.

[305] D. H. Glueck, A. Karimpour-Fard, J. Mandel, and K. E. Muller. Probabilities for separating sets of order statistics. *Statistics*, 44(2):145–153, 2010.

[306] R. J. Beverton and S. J. Holt. *On the dynamics of exploited fish populations*, volume 11 of *Fisheries Investigation Series 2, volume 19*. UK Ministery of Agriculture and Fisheries, London, UK, 1957.

[307] P. Kashyap. Traditional haats and melas in India. *MART (Marketing and Research Team), New Delhi*, 24, 1995.

[308] G. Patnaik, S. A. the Working Group on Agricultural Marketing Infrastructure, P. R. for Internal, and E. Trade. Report of the working group on agricultural marketing infrastructure, secondary agriculture and policy required for internal and external trade for the xii five year plan 2012-2017, 2011.

[309] G. Patnaik. Status of agricultural marketing reforms. In *Workshop on Policy Options and Investment Priorities for Accelerating Agricultural Productivity and Development in India, New Delhi, India*, 2011.

[310] S. Anagol. Adverse selection in asset markets: Theory and evidence from the Indian market for cows. 2009.

[311] V. B. Bairagi. Marketing of cows in cattle market Loni. *Lokavishkar International E-Journal*, 2.

[312] M. Ali. Livestock trade in semi-subsistence type of rural economy: A case study from Uttar Pradesh, India. *International Journal of Management, IT and Engineering*, 2(9):353–368, 2012.

[313] D. O. Lodrick. A cattle fair in Rajasthan: The Kharwa Mela. *Current Anthropology*, pages 218–225, 1984.

[314] M. R. Rosenzweig and K. I. Wolpin. Credit market constraints, consumption smoothing, and the accumulation of durable production assets in low-income countries: Investments in bullocks in India. *Journal of Political Economy*, pages 223–244, 1993.

[315] A. K. Gauraha and B. C. Jain. Marketing of livestock in rural areas of Madhya Pradesh. In J. Prasad, editor, *Encyclopaedia of Agricultural Marketing Vol. 7*, pages 27–36. Mittal Publications, New Delhi, India, 2001.

[316] G. of India. Indian 18th livestock census 2007. http://dahd.nic.in/DistrictWiseReport/HTML/Chattisgarh.htm.

[317] G. Das, D. Jain, and J. Dhaka. Analysis of price spread and marketing efficiency of milch cow marketing in the state level cattle fairs of Rajasthan, India. *SAARC Journal of Agriculture*, 12(1):33–47, 2014.

[318] T. R. Shanmugan and V. Balakrishnan. Livestock marketing. In J. Prasad, editor, *Encyclopaedia of Agricultural Marketing Vol. 7*, pages 11–26. Mittal Publications, New Delhi, India, 2001.

[319] A. Pandit. Efficiency of dairy cattle markets in central alluvial plains of West Bengal. *Agricultural Marketing*, 48(1):44, 2005.

[320] A. S. Babu. A study of cattle markets in Kolar district of Karnataka. *Asian Journal of Research in Business Economics and Management*, 1(3):300–304, 2011.

[321] P. Sagar and H. Patange. *Indian Streams Research Journal*, 1, 2011.

[322] N. T. S. P. Singh, R. and K. Kumar. Functioning of livestock markets and buyers' perspectives on voluntary versus mandatory disclosure of information: Evidence from cattle markets in Uttar Pradesh. *Indian Journal of Agricultural Economics*, 69(3), 2014.

[323] R. A. Bentley. Scale-free network growth and social inequality. In R. A. Bentley and H. D. Maschner, editors, *Complex systems and archaeology.* University of Utah Press Salt Lake City, 2003.

[324] R. A. Bentley, M. W. Lake, and S. J. Shennan. Specialisation and wealth inequality in a model of a clustered economic network. *Journal of Archaeological Science*, 32(9):1346–1356, 2005.

[325] C. Smith. Patterns of wealth concentration. *Human Organization*, 50(1):50–60, 1991.

[326] H. Lopez and L. Servén. A normal relationship? Poverty, growth, and inequality. *Poverty, Growth, and Inequality (January 2006). World Bank Policy Research Working Paper*, (3814), 2006.

[327] L. M. Rouse and L. Weeks. Specialization and social inequality in Bronze Age SE Arabia: Analyzing the development of production strategies and economic networks using agent-based modeling. *Journal of Archaeological Science*, 38(7):1583–1590, 2011.

[328] D. F. Sieff. The effects of wealth on livestock dynamics among the Datoga pastoralists of Tanzania. *Agricultural Systems*, 59(1):1–25, 1999.

[329] S. Harrison and F. Moog. Livestock dispersal programs in developing countries: social-economic benefits for human resource development in rural sector. In R. Ghosh, R. Gabbay, and A. Siddique, editors, *Human resources and gender issues in poverty eradication*, page 269. Atlantic, Delhi, India, 2001.

[330] B. Gebremedhin, D. Hoekstra, A. Tegegne, K. Shiferaw, and A. Bogale. Factors determining household market participation in small ruminant production in the highlands of Ethiopia. 2015.

[331] A. Negassa and M. Jabbar. *Livestock ownership, commercial off-take rates and their determinants in Ethiopia.* ILRI, 2008.

[332] M. Lubungu, A. Chapoto, and G. Tembo. Smallholder farmers participation in livestock markets: The case of Zambian farmers. Technical report, Citeseer, 2012.

[333] P. M. Mwanyumba. *Pastoralist liveslihoods, livestock herd dynamics and trade in Garissa County, Kenya.* PhD thesis, University of Nairobi, 2014. Chapter 6, 'Garissa livestock market structure, conduct and performance, Kenya.

[334] M. A. Jabbar. Buyer preferences for sheep and goats in southern Nigeria: A hedonic price analysis. *Agricultural Economics*, 18(1):21–30, 1998.

[335] M. A. Jabbar and M. Diedhiou. Does breed matter to cattle farmers and buyers? Evidence from West Africa. *Ecological Economics*, 45(3):461–472, 2003.

[336] G. Das and D. Jain. Factors affecting the price of bullocks in the organised cattle fairs of Rajisthan. *Indian Journal of Agricultural Economics*, 68(4):594–599, 2013.

[337] T. O. Williams, I. Okike, and B. Spycher. A hedonic analysis of cattle prices in the central corridor of West Africa: Implications for production and marketing decisions. In *Contributed paper prepared for presentation at the International Association of Agricultural Economists Conference. Gold Coast, Australia*, 2006.

[338] E. Verbeek, E. Kanis, R. Bett, and I. Kosgey. Socio-economic factors influencing small ruminant breeding in Kenya. 2007.

[339] R. Singh. *The Indian Cow*, pages 29–37, 2006.

[340] C. Roncoli, K. Ingram, and P. Kirshen. The costs and risks of coping with drought: Livelihood impacts and farmers' responses in Burkina Faso. *Climate Research*, 19(2):119–132, 2001.

[341] M. R. Carter, P. D. Little, T. Mogues, and W. Negatu. Shocks, sensitivity and resilience: Tracking the economic impacts of environmental disaster on assets in Ethiopia and Honduras. *BASIS Research Program on Poverty, Inequality and Development. Washington, DC: US Agency for International Development*, 2004.

[342] J. Hoddinott. Shocks and their consequences across and within households in rural Zimbabwe. *The Journal of Development Studies*, 42(2):301–321, 2006.

[343] S. Vetter. Rangelands at equilibrium and non-equilibrium: Recent developments in the debate. *Journal of Arid Environments*, 62(2):321–341, 2005.

[344] D. Nkedianye, J. de Leeuw, J. O. Ogutu, M. Y. Said, T. L. Saidimu, et al. Mobility and livestock mortality in communally used pastoral areas: The impact of the 2005-2006 drought on livestock mortality in Maasailand. *Pastoralism*, 1(1):1–17, 2011.

[345] J. T. McCabe. Success and failure: The breakdown of traditional drought coping institutions among the pastoral Turkana of Kenya. *Journal of Asian and African Studies*, 25(3-4):146–160, 1990.

[346] C. B. Barrett, M. F. Bellemare, and S. M. Osterloh. Household-level livestock marketing behavior among northern Kenyan and southern Ethiopian pastoralists. *Available at SSRN 716301*, 2004.

[347] P. D. Little, M. P. Stone, T. Mogues, A. P. Castro, and W. Negatu. 'moving in place': Drought and poverty dynamics in South Wollo, Ethiopia. *The Journal of Development Studies*, 42(2):200–225, 2006.

[348] N. Alemayehu, G. Gebru, and W. Ayalew. Genetic dilution of the Ethiopian Boran cattle. *Challenges and Opportunities of Livestock Marketing in Ethiopia*, page 377, 2002.

[349] A. Angassa and G. Oba. Relating long-term rainfall variability to cattle population dynamics in communal rangelands and a government ranch in southern Ethiopia. *Agricultural Systems*, 94(3):715–725, 2007.

[350] G. Kibreab. The consequences of non-participatory planning: Lessons from a livestock provision project to returnees in Eritrea. *Journal of Refugee Studies*, 12(2):135–160, 1999.

[351] L. M. Ausubel. auctions (theory). In S. N. Durlauf and L. E. Blume, editors, *The New Palgrave Dictionary of Economics*. Palgrave Macmillan, Basingstoke, 2008.

[352] H. Varian. *Intermediate Microeconomics: A Modern Approach*. W.W. Norton & Company, 2010.

[353] G. Bordes, D. E. Campbell, and M. Le Breton. Arrow's theorem for economic domains and Edgeworth hyperboxes. *International Economic Review*, pages 441–454, 1995.

[354] G. A. Akerlof. The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, pages 488–500, 1970.

[355] J. F. Kingman. On the genealogy of large populations. *Journal of Applied Probability*, pages 27–43, 1982.

[356] Y. Kim and W. Stephan. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*, 160(2):765–777, 2002.

[357] Y. Kim and R. Nielsen. Linkage disequilibrium as a signature of selective sweeps. *Genetics*, 167(3):1513–1524, 2004.

[358] R. Nielsen, S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark, et al. Genomic scans for selective sweeps using SNP data. *Genome Research*, 15(11):1566–1575, 2005.

[359] B. F. Voight, S. Kudaravalli, X. Wen, and J. K. Pritchard. A map of recent positive selection in the human genome. *PLoS Biology*, 4(3):446, 2006.

[360] J. D. Jensen, K. R. Thornton, C. D. Bustamante, and C. F. Aquadro. On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in nonequilibrium populations. *Genetics*, 176(4):2371–2379, 2007.

[361] J. K. Pickrell, G. Coop, J. Novembre, S. Kudaravalli, J. Z. Li, et al. Signals of recent positive selection in a worldwide sample of human populations. *Genome Research*, 19(5):826–837, 2009.

[362] S. R. Grossman, I. Shylakhter, E. K. Karlsson, E. H. Byrne, S. Morales, et al. A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science*, 327(5967):883–886, 2010.

[363] P. Pavlidis, J. D. Jensen, and W. Stephan. Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. *Genetics*, 185(3):907–922, 2010.

[364] J. L. Crisci, Y.-P. Poh, S. Mahajan, and J. D. Jensen. The impact of equilibrium assumptions on tests of selection. *Frontiers in Genetics*, 4, 2013.

[365] A. Ferrer-Admetlla, M. Liang, T. Korneliussen, and R. Nielsen. On detecting incomplete soft or hard selective sweeps using haplotype structure. *Molecular Biology and Evolution*, 31(5):1275–1291, 2014.

[366] M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson, and E. S. Lander. High-resolution haplotype structure in the human genome. *Nature Genetics*, 29(2):229–232, 2001.

[367] T. D. Petes. Meiotic recombination hot spots and cold spots. *Nature Reviews Genetics*, 2(5):360–369, 2001.

[368] C. Mezard. Meiotic recombination hotspots in plants. *Biochemical Society Transactions*, 34(4):531–534, 2006.

[369] K. Paigen and P. Petkov. Mammalian recombination hot spots: Properties, control and evolution. *Nature Reviews Genetics*, 11(3):221–233, 2010.

[370] J. M. Akey. Constructing genomic maps of positive selection in humans: Where do we go from here? *Genome Research*, 19(5):711–722, 2009.

[371] R. D. Barrett and H. E. Hoekstra. Molecular spandrels: Tests of adaptation at the genetic level. *Nature Reviews Genetics*, 12(11):767–780, 2011.

[372] M. Nei, Y. Suzuki, and M. Nozawa. The neutral theory of molecular evolution in the genomic era. *Annual Review of Genomics and Human Genetics*, 11:265–289, 2010.

[373] S. J. Gould and R. C. Lewontin. The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society of London B: Biological Sciences*, 205(1161):581–598, 1979.

[374] R. Nielsen. Adaptionism30 years after Gould and Lewontin. *Evolution*, 63(10):2487–2490, 2009.

[375] G. A. McVean, S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley, et al. The fine-scale structure of recombination rate variation in the human genome. *Science*, 304(5670):581–584, 2004.

[376] A. Kong, G. Thorleifsson, D. F. Gudbjartsson, G. Masson, A. Sigurdsson, et al. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, 467(7319):1099–1103, 2010.

[377] H. Li, U. B. Gyllensten, X. Cui, R. K. Saiki, H. A. Erlich, et al. Amplification and analysis of dna sequences in single human sperm and diploid cells. *Nature*, 335(6189):414–417, 1988.

[378] R. Lewontin. The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics*, 49(1):49, 1964.

[379] W. Hill and A. Robertson. Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, 38(6):226–231, 1968.

[380] J. K. Kelly. A test of neutrality based on interlocus associations. *Genetics*, 146(3):1197–1206, 1997.

[381] J. M. Smith and J. Haigh. The hitch-hiking effect of a favourable gene. *Genetical research*, 23(01):23–35, 1974.

[382] P. Sabeti, S. Schaffner, B. Fry, J. Lohmueller, P. Varilly, et al. Positive natural selection in the human lineage. *Science*, 312(5780):1614–1620, 2006.

[383] T. K. Oleksyk, M. W. Smith, and S. J. O'Brien. Genome-wide scans for footprints of natural selection. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1537):185–205, 2010.

[384] S. Wright. The genetical structure of populations. *Annals of Eugenics*, 15(1):323–354, 1949.

[385] X. Yi, Y. Liang, E. Huerta-Sanchez, X. Jin, Z. X. P. Cuo, et al. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*, 329(5987):75–78, 2010.

[386] D. Charlesworth. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics*, 2(4):e64, 04 2006.

[387] J. C. Fay and C.-I. Wu. Hitchhiking under positive Darwinian selection. *Genetics*, 155(3):1405–1413, 2000.

[388] F. Tajima. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):585–595, 1989.

[389] P. C. Sabeti, D. E. Reich, J. M. Higgins, H. Z. Levine, D. J. Richter, et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909):832–837, 2002.

[390] K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861, 2007.

[391] D. R. Schrider and A. D. Kern. S/HIC: Robust identification of soft and hard sweeps using machine learning. *bioRxiv*, page 024547, 2015.

[392] M. Kimura. Evolutionary rate at the molecular level. *Nature*, 217(5129):624–626, 1968.

[393] J. Lachance and S. A. Tishkoff. Population genomics of human adaptation. *Annual review of ecology, evolution, and systematics*, 44:123, 2013.

[394] T. Bersaglieri, P. C. Sabeti, N. Patterson, T. Vanderploeg, S. F. Schaffner, et al. Genetic signatures of strong recent positive selection at the lactase gene. *The American Journal of Human Genetics*, 74(6):1111–1120, 2004.

[395] F. J. Simoons. Primary adult lactose intolerance and the milking habit: A problem in biologic and cultural interrelations. *The American Journal of Digestive Diseases*, 15(8):695–710, 1970.

[396] N. S. Enattah, T. Sahi, E. Savilahti, J. D. Terwilliger, L. Peltonen, et al. Identification of a variant associated with adult-type hypolactasia. *Nature Genetics*, 30(2):233–237, 2002.

[397] S. A. Tishkoff, F. A. Reed, A. Ranciaro, B. F. Voight, C. C. Babbitt, et al. Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genetics*, 39(1):31–40, 2007.

[398] C. J. Ingram, M. F. Elamin, C. A. Mulcare, M. E. Weale, A. Tarekegn, et al. A novel polymorphism associated with lactose tolerance in Africa: Multiple causes for lactase persistence? *Human Genetics*, 120(6):779–788, 2007.

[399] N. S. Enattah, T. G. Jensen, M. Nielsen, R. Lewinski, M. Kuokkanen, et al. Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture. *The American Journal of Human Genetics*, 82(1):57–72, 2008.

[400] L. Fang, J. K. Ahn, D. Wodziak, and E. Sibley. The human lactase persistence-associated SNP-13910* T enables *in vivo* functional persistence of lactase promoter–reporter transgene expression. *Human Genetics*, 131(7):1153–1159, 2012.

[401] I. Mathieson, I. Lazaridis, N. Rohland, S. Mallick, B. Llamas, et al. Eight thousand years of natural selection in Europe. *bioRXiv*, 2015.

[402] R. L. Lamason, M.-A. P. Mohideen, J. R. Mest, A. C. Wong, H. L. Norton, et al. *SLC24A5*, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science*, 310(5755):1782–1786, 2005.

[403] C. B. Mallick, F. M. Iliescu, M. Möls, S. Hill, R. Tamang, et al. The light skin allele of *SLC24A5* in South Asians and europeans shares identity by descent. 2013.

[404] P. C. Sabeti, P. Varilly, B. Fry, J. Lohmueller, E. Hostetter, et al. Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449(7164):913–918, 2007.

[405] J. Bryk, E. Hardouin, I. Pugach, D. Hughes, R. Strotmann, et al. Positive selection in East Asians for an *EDAR* allele that enhances NF-κB activation. *PLoS One*, 3(5):e2209–e2209, 2008.

[406] A. Fujimoto, R. Kimura, J. Ohashi, K. Omi, R. Yuliwulandari, et al. A scan for genetic determinants of human hair morphology: *EDAR* is associated with Asian hair thickness. *Human Molecular Genetics*, 17(6):835–843, 2008.

[407] R. Kimura, T. Yamaguchi, M. Takeda, O. Kondo, T. Toma, et al. A common variation in *EDAR* is a genetic determinant of shovel-shaped incisors. *The American Journal of Human Genetics*, 85(4):528–535, 2009.

[408] Y. G. Kamberov, S. Wang, J. Tan, P. Gerbault, A. Wark, et al. Modeling recent human evolution in mice by expression of a selected *EDAR* variant. *Cell*, 152(4):691–702, 2013.

[409] C. M. Beall, G. L. Cavalleri, L. Deng, R. C. Elston, Y. Gao, et al. Natural selection on *EPAS1* (*HIF*2α) associated with low hemoglobin concentration in Tibetan highlanders. *Proceedings of the National Academy of Sciences*, 107(25):11459–11464, 2010.

[410] M. P. Donnelly, P. Paschou, E. Grigorenko, D. Gurwitz, C. Barta, et al. A global view of the *OCA2-HERC2* region and pigmentation. *Human Genetics*, 131(5):683–696, 2012.

[411] J. Lachance, B. Vernot, C. C. Elbers, B. Ferwerda, A. Froment, et al. Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell*, 150(3):457–469, 2012.

[412] G. Casella and R. L. Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

[413] G. Bohm and G. Zech. *Introduction to statistics and data analysis for physicists.* DESY, 2010.

[414] International HapMap Consortium et al. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, 2005.

[415] H. M. Cann, C. De Toma, L. Cazes, M.-F. Legrand, V. Morel, et al. A human genome diversity cell line panel. *Science*, 296(5566):261, 2002.

[416] 1000 Genomes Project Consortium et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.

[417] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

[418] L. Cavalli-Sforza. Population structure and human evolution. 164(995):362–379, 1966.

[419] P. C. Sham and S. M. Purcell. Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics*, 15(5):335–346, 2014.

[420] J. Felsenstein. Theoretical evolutionary genetics. January 2015 edition. Note that this is sometimes updated with corresponding slight changes to equation numbers., 2015.

[421] R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, and C. D. Bustamante. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, 5(10):e1000695, 2009.

[422] S. Gravel, B. M. Henn, R. N. Gutenkunst, A. R. Indap, G. T. Marth, et al. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences*, 108(29):11983–11988, 2011.

[423] N. Bierne, D. Roze, and J. J. Welch. Pervasive selection or is it…? Why are F_{ST} outliers sometimes so frequent? *Molecular Ecology*, 22(8):2061–2064, 2013.

[424] A. Hodgkinson and A. Eyre-Walker. Variation in the mutation rate across mammalian genomes. *Nature Reviews Genetics*, 12(11):756–766, 2011.

[425] A. G. Hinch, A. Tandon, N. Patterson, Y. Song, N. Rohland, et al. The landscape of recombination in African Americans. *Nature*, 476(7359):170–175, 2011.

[426] P. A. P. Moran. Random processes in genetics. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 54, pages 60–71. Cambridge Univ Press, 1958.

[427] X. Yuan, D. J. Miller, J. Zhang, D. Herrington, and Y. Wang. An overview of population genetic data simulation. *Journal of Computational Biology*, 19(1):42–54, 2012.

[428] F. Tajima. Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105(2):437–460, 1983.

[429] R. R. Hudson. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23(2):183–201, 1983.

[430] J. F. Kingman. Origins of the coalescent: 1974–1982. *Genetics*, 156(4):1461–1463, 2000.

[431] P. Donnelly and S. Tavaré. Coalescents and genealogical structure under neutrality. *Annual Review of Genetics*, 29(1):401–421, 1995.

[432] J. Wakeley. *Coalescent theory: An introduction*, volume 1, chapter 3. Roberts & Company Publishers Greenwood Village, Colorado, 2009.

[433] C. Cannings. The latent roots of certain Markov chains arising in genetics: A new approach, I. Haploid models. *Advances in Applied Probability*, pages 260–290, 1974.

[434] R. R. Hudson. Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, 7(1):44, 1990.

[435] R. R. Hudson. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002.

[436] G. A. Huttley, M. W. Smith, M. Carrington, and S. J. OBrien. A scan for linkage disequilibrium across the human genome. *Genetics*, 152(4):1711–1722, 1999.

[437] J. L. Kelley, J. Madeoy, J. C. Calhoun, W. Swanson, and J. M. Akey. Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Research*, 16(8):980–989, 2006.

[438] K. R. Thornton and J. D. Jensen. Controlling the false-positive rate in multilocus genome scans for selection. *Genetics*, 175(2):737–750, 2007.

[439] S. Mallick, S. Gnerre, P. Muller, and D. Reich. The difficulty of avoiding false positives in genome scans for natural selection. *Genome Research*, 19(5):922–933, 2009.

[440] W. Stephan, Y. S. Song, and C. H. Langley. The hitchhiking effect on linkage disequilibrium between linked neutral loci. *Genetics*, 172(4):2647–2663, 2006.

[441] G. McVean. The structure of linkage disequilibrium around a selective sweep. *Genetics*, 175(3):1395–1406, 2007.

[442] P. Pfaffelhuber, A. Lehnert, and W. Stephan. Linkage disequilibrium under genetic hitchhiking in finite populations. *Genetics*, 179(1):527–537, 2008.

[443] J. Rozas, M. Gullaud, G. Blandin, and M. Aguadé. Dna variation at the rp49 gene region of *Drosophila simulans*: Evolutionary inferences from an unusual haplotype structure. *Genetics*, 158(3):1147–1155, 2001.

[444] N. Alachiotis, A. Stamatakis, and P. Pavlidis. OmegaPlus: A scalable tool for rapid detection of selective sweeps in whole-genome datasets. *Bioinformatics*, 28(17):2274–2275, 2012.

[445] P. F. OReilly, E. Birney, and D. J. Balding. Confounding between recombination and selection, and the Ped/Pop method for detecting selection. *Genome Research*, 18(8):1304–1313, 2008.

[446] P. S. Pennings and J. Hermisson. Soft sweeps III: The signature of positive selection from recurrent mutation. *PLoS Genetics*, 2(12):e186–e186, 2006.

[447] N. Alachiotis, P. Pavlidis, and A. Stamatakis. Exploiting multi-grain parallelism for efficient selective sweep detection. In *Algorithms and Architectures for Parallel Processing*, pages 56–68. Springer, 2012.

[448] A. Catalán, S. Hutter, and J. Parsch. Population and sex differences in drosophila melanogaster brain gene expression. *BMC Genomics*, 13(1):654, 2012.

[449] T. Lee, S. Cho, K. S. Seo, J. Chang, H. Kim, et al. Genetic variants and signatures of selective sweep of Hanwoo population (Korean native cattle). *BMB reports*, 46(7):346–351, 2013.

[450] N. Renzette, T. F. Kowalik, and J. D. Jensen. On the relative roles of background selection and genetic hitchhiking in shaping human cytomegalovirus genetic diversity. *Molecular Ecology*, 2015.

[451] A. Kong, M. L. Frigge, G. Masson, S. Besenbacher, P. Sulem, et al. Rate of de novo mutations and the importance of father's age to disease risk. *Nature*, 488(7412):471–475, 2012.

[452] C. E. Metz. Basic principles of ROC analysis. In *Seminars in nuclear medicine*, volume 8, pages 283–298. Elsevier, 1978.

[453] D. K. McClish. Analyzing a portion of the ROC curve. *Medical Decision Making*, 9(3):190–195, 1989.

[454] C. S. Carlson, D. J. Thomas, M. A. Eberle, J. E. Swanson, R. J. Livingston, et al. Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Research*, 15(11):1553–1565, 2005.

[455] T. K. Oleksyk, K. Zhao, M. Francisco, D. A. Gilbert, S. J. O'Brien, et al. Identifying selected regions from heterozygosity and divergence using a light-coverage genomic dataset from two human populations. *PLoS ONE*, 3(3):e1712, 2008.

[456] H. Chen, N. Patterson, and D. Reich. Population differentiation as a test for selective sweeps. *Genome Research*, 20(3):393–402, 2010.

[457] R. Ronen, N. Udpa, E. Halperin, and V. Bafna. Learning natural selection from the site frequency spectrum. *Genetics*, 195(1):181–193, 2013.

[458] V. Colonna, Q. Ayub, Y. Chen, L. Pagani, P. Luisi, et al. Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. *Genome Biology*, 15(6):R88, 2014.

[459] Y. Itan, A. Powell, M. A. Beaumont, J. Burger, M. G. Thomas, et al. The origins of lactase persistence in Europe. *PLoS Comput. Biol.*, 5(8):e1000491–e1000491, 2009.

[460] L. B. Barreiro, M. Ben-Ali, H. Quach, G. Laval, E. Patin, et al. Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. *PLoS Genetics*, 5(7):e1000562, 2009.

[461] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[462] M. Florio, M. Albert, E. Taverna, T. Namba, H. Brandl, et al. Human-specific gene *ARHGAP11B* promotes basal progenitor amplification and neocortex expansion. *Science*, 347(6229):1465–1470, 2015.

[463] A. P. Abreu, A. Dauber, D. B. Macedo, S. D. Noel, V. N. Brito, et al. Central precocious puberty caused by mutations in the imprinted gene *MKRN3*. *New England Journal of Medicine*, 368(26):2467–2475, 2013.

[464] J. R. Perry, F. Day, C. E. Elks, P. Sulem, D. J. Thompson, et al. Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature*, 514(7520):92–97, 2014.

[465] D. Enard, P. W. Messer, and D. A. Petrov. Genome-wide signals of positive selection in human evolution. *Genome Research*, 24(6):885–895, 2014.

[466] S. F. Schaffner, C. Foo, S. Gabriel, D. Reich, M. J. Daly, et al. Calibrating a coalescent simulation of human genome sequence variation. *Genome Research*, 15(11):1576–1583, 2005.

[467] J. Wakeley. Natural selection and coalescent theory. In M. A. Bell, D. J. Futuyma, W. F. Eanes, and J. S. Levinton, editors, *Evolution since Darwin: The first 150 years*, pages 119–149. Sinauer Associates Sunderland, Massachusetts, 2010.

[468] R. R. Hudson and N. L. Kaplan. The coalescent process in models with selection and recombination. *Genetics*, 120(3):831–840, 1988.

[469] Y. Kim and T. Wiehe. Simulation of DNA sequence evolution under models of recent directional selection. *Briefings in bioinformatics*, 10(1):84–96, 2009.

[470] R. Durrett and J. Schweinsberg. Approximating selective sweeps. *Theoretical Population Biology*, 66(2):129–138, 2004.

[471] Y.-X. Fu. Exact coalescent for the Wright–Fisher model. *Theoretical Population Biology*, 69(4):385–394, 2006.

[472] J. Wang, H. C. Fan, B. Behr, and S. R. Quake. Genome-wide single-cell analysis of recombination activity and *de novo* mutation rates in human sperm. *Cell*, 150(2):402–412, 2012.

[473] K. Harris and R. Nielsen. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genetics*, 9(6):e1003521, 2013.

[474] J. Hermisson and P. S. Pennings. Soft sweeps molecular population genetics of adaptation from standing genetic variation. *Genetics*, 169(4):2335–2352, 2005.

[475] D. R. Schrider, F. K. Mendes, M. W. Hahn, and A. D. Kern. Soft shoulders ahead: Spurious signatures of soft and partial selective sweeps result from linked hard sweeps. *Genetics*, 200(1):267–284, 2015.

[476] A. Wollstein and W. Stephan. Inferring positive selection in humans from genomic data. *Investigative Genetics*, 6(1):5, 2015.

[477] P. C. Phillips. Epistasisthe essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, 9(11):855–867, 2008.

[478] R. D. Hernandez, J. L. Kelley, E. Elyashiv, S. C. Melton, A. Auton, et al. Classic selective sweeps were rare in recent human evolution. *Science*, 331(6019):920–924, 2011.

[479] E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. [Online; accessed 2015-08-19].

[480] J. Schweinsberg. Coalescents with simultaneous multiple collisions. *Electronic Journal of Probability*, 5:1–50, 2000.

[481] J. Pitman. Coalescents with multiple collisions. *Annals of Probability*, pages 1870–1902, 1999.

[482] P. Sjödin, I. Kaj, S. Krone, M. Lascoux, and M. Nordborg. On the meaning and existence of an effective population size. *Genetics*, 169(2):1061–1070, 2005.

[483] J. Wakeley. Coalescent theory has many new branches. *Theoretical Population Biology*, (87):1–4, 2013.

[484] T. Ohta. Slightly deleterious mutant substitutions in evolution. *Nature*, 246(5428):96–98, 1973.

[485] G. Sella, D. A. Petrov, M. Przeworski, and P. Andolfatto. Pervasive natural selection in the *Drosophila* genome. *PLoS Genet*, 5(6):e1000495, 2009.

[486] K. Lindblad-Toh, M. Garber, O. Zuk, M. F. Lin, B. J. Parker, et al. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478(7370):476–482, 2011.

[487] W. Fu and J. M. Akey. Selection and adaptation in the human genome. *Annual Review of Genomics and Human Genetics*, 14:467–489, 2013.

[488] B. Charlesworth, M. Morgan, and D. Charlesworth. The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134(4):1289–1303, 1993.

[489] B. Charlesworth. The effects of deleterious mutations on evolution at linked sites. *Genetics*, 190(1):5–22, 2012.

[490] L. E. Nicolaisen and M. M. Desai. Distortions in genealogies due to purifying selection and recombination. *Genetics*, 195(1):221–230, 2013.

[491] R. D. Hernandez. A flexible forward simulator for populations subject to selection and demography. *Bioinformatics*, 24(23):2786–2787, 2008.

[492] K. Zeng and B. Charlesworth. The joint effects of background selection and genetic recombination on local gene genealogies. *Genetics*, 189(1):251–266, 2011.

[493] W. Allee and E. S. Bowen. Studies in animal aggregations: Mass protection against colloidal silver among goldfishes. *Journal of Experimental Zoology*, 61(2):185–207, 1932.

[494] A. P. Kirman. Whom or what does the representative individual represent? *The Journal of Economic Perspectives*, pages 117–136, 1992.

[495] M. Levy, H. Levy, and S. Solomon. Microscopic simulation of the stock market: The effect of microscopic diversity. *Journal de Physique I*, 5(8):1087–1107, 1995.

[496] D. Colander, M. Goldberg, A. Haas, K. Juselius, A. Kirman, et al. The financial crisis and the systemic failure of the economics profession. *Critical Review*, 21(2-3):249–267, 2009.

[497] F. Black and M. Scholes. The pricing of options and corporate liabilities. *The Journal of Political Economy*, pages 637–654, 1973.

[498] R. C. Merton. Theory of Rational Option Pricing. *Bell Journal of Economics*, 4(1):141–183, Spring 1973.

[499] B. Eichengreen. Origins and responses to the crisis, 2008.

[500] E. G. Haug and N. N. Taleb. Option traders use (very) sophisticated heuristics, never the Black–Scholes–Merton formula. *Journal of Economic Behavior & Organization*, 77(2):97–106, 2011.

[501] A. M. Cowan and C. D. Cowan. Default correlation: An empirical investigation of a subprime lender. *Journal of Banking & Finance*, 28(4):753–771, 2004.

[502] N. Silver. *The signal and the noise: Why so many predictions fail-but some don't.* Penguin, 2012.

[503] SNP FAQ archive: Ascertainment bias, 2005. Bethesda (MD): National Center for Biotechnology Information (US); available online from http://www.ncbi.nlm.nih.gov/books/NBK9792/.

Appendices

A1 Thesis Appendix 1: Presentation of work at conferences

Much of the work included has been presented as talks or posters in seminars and conferences.

Presentations at conferences

2nd International Conference on New Frontiers in Industrial and Applied Biotechnology, Bareilly, India, 2015 (Plenary lecture) - “Simulation in population biology and evolutionary genetics”

Modelling Biological Evolution conference, Leicester, 2013 - “Finite populations and long-distance dispersal”

Student Conference on Complexity Science, Gloucester, 2012 - ‘’Modelling species invasions: accelerating waves of advance in finite populations”

Posters at conferences

XIV Congress of the European Society for Evolutionary Biology, Lisbon, Portugal, 2013 “Linkage-based method for identifying selected loci and characterising selection intensity”

Wessex Life Sciences Alliance conference (Institute for Life Sciences), Southampton, 2012 - “Linkage-based genetic signatures of recent human evolution”

Student Conference on Complexity Science, Gloucester 2012 - “Population diffusion and long distance dispersal in finite size populations”

Seminars

Centre for the Archaeology of Human Origins seminar series, Southampton, 2013 - “Long-distance dispersal and human population diffusion: why modelling the spread of populations in archaeology is harder than we think”

Applied Mathematics Postgraduate seminar series, Southampton, 2012 - “Accelerating species invasions”

Human Evolutionary Genetics, Kivisild Lab seminar, Cambridge 2013 - “Linkage simulations”

A2 Thesis Appendix 2: Simple derivation of a diffusive limit for mean-field Model 2

We here offer a basic derivation of a diffusive limit for our mean field Model 2. This is intended to highlight differences between our model and the Fisher-Kolmogorov equation, rather than as a rigorous indication that this is the exact diffusive limit.

We begin with our Model 2 equation,

$$n(x,t+1) = (1-d)n(x,t) + (1-d)b[1-n(x,t)] \sum_{l=-\infty}^{+\infty} K(|l|)n(x+l,t), \tag{A2.1}$$

noting that in the temporal continuum limit

$$\big[n(x,t+1) - n(x,t)\big] \rightarrow \frac{\partial n(x)}{\partial t}, \tag{A2.2}$$

leading, with $d = 0$ and $b = 1$, to the integrodifference equation

$$\frac{\partial n(x,t)}{\partial t} = [1-n(x,t)] \sum_{l=-\infty}^{+\infty} K(|l|)n(x+l,t). \tag{A2.3}$$

We take the lattice sum on the right hand side to its spatially continuous limit,

$$\sum_{l\neq 0} K(|l|)n(x+l,t) \rightarrow \int_{-\infty}^{\infty} K(|x-y|)n(y,t)\mathrm{d}y, \tag{A2.4}$$

where the normalisation conditions

$$\sum_{l=-\infty}^{\infty} K(|l|) = \int_{-\infty}^{\infty} K(|x-y|)\mathrm{d}y = 1. \tag{A2.5}$$

ensure that the kernels represent probabilities of propagules at particular points.

We now suppose the integral to be one dimensional and follow the traditional approach (eg. [48] in a rather similar model of epidemic spread) of expanding $n(x')$ in a Taylor series around x and substituting $n(x')$ into Eq. (A2.4), omitting forgotten terms. A similar method can be applied to integrodifference equations with the non-linear behaviour applied to the growth term, and retrieves the Fisher-Kolmogorov equation (eg. [49]).

Some caution is advised here - [49], following [96], find that complications can arise when describing the discrete-time behaviour as a continuous-time system, particularly when a significant time-delay is involved. This is relevant for many biological systems. However, as our intention is to present a qualitative comparison to the Fisher-Kolmogorov model rather than to retrieve a diffusive approximation for further analysis we trust this derivation will suffice. We obtain

$$n(x') = n(x) + (x' - x)\frac{\partial n(x)}{\partial x} + \frac{1}{2}(x - x')^2\frac{\partial^2 n(x)}{\partial x^2} + \ldots. \tag{A2.6}$$

and hence

$$\frac{\partial n(x,t)}{\partial t} = n(x,t)(1 - n(x,t)) + D(1 - n(x,t))\frac{\partial^2 n(x,t)}{\partial x^2}, \tag{A2.7}$$

with

$$D = \frac{1}{2}\left(\int_{-\infty}^{\infty} l^2 K(|l|)\,\mathrm{d}l\right),$$

which is the diffusion approximation of Mollison's simple epidemic [59]. Indeed, we find that several analytic results for this model [59] hold when we reduce the spatial and temporal scale of our system, §2.2.4.1.

This equation resembles the Fisher-Kolmogorov equation, though is not identical to it. Although generally our simulations have $d = 0$ and $b = 1$, we do explore the situation where $d > 0$, and re-introducing both terms modifies Eq. (A2.7) to yield

$$\frac{\partial n(x,t)}{\partial t} = \alpha n(x,t)\left(1 - \frac{n(x,t)}{\kappa}\right) + \tilde{D}[1 - n(x,t)]\frac{\partial^2 n(x,t)}{\partial x^2}. \tag{A2.8}$$

where the equilibrium occupation is given by $\kappa = \frac{\alpha}{(1-d)b}$, the Malthusian constant is $\alpha = (1-d)b - d$, and the effective diffusion constant is also modified, yielding $\tilde{D} = b(1-d)D$.

On inspection, our diffusive approximation Eq. (A2.8) behaves in a similar manner to the stochastic algorithm of Model 1 and our mean-field equation Eq. (2.6). As a result of the detailed form of reproduction in our model, and in contrast to the Fisher-Kolmogorov equation, diffusion neither occurs when $b = 0$ or when the system is full. In the latter case, increasing the death rate d leads to a carrying capacity $\kappa < 1.0$, and diffusion is resumed. The population dies out when $d > (1-d)b$, as the number deaths a site experiences outweighs its contribution to population growth through surviving births, even in an empty environment. These behaviours are biologically reasonable in many cases. For example, the main crowding effects for certain plant species are likely to arise through limiting the survival of recently dispersed saplings rather than in reducing the number of propagules for adult organisms.

We discuss the relevance of these differences given the Linear Conjecture in the main text. Results from simulations that investigate the effect of implementing a logistic effect based on home-site filling can be found in §2.2.4.6.

A3 Thesis Appendix 3: Marginal stability analysis of Model 2 with a nearest-neighbour kernel

Here we offer an argument based on standard methods to predict the velocity implied by our mean-field Model 2 for short-range kernels. The full recurrence relation of Model 2 is given by Eq.(2.6). Far ahead of the wave of advance, all $\{n\}$ are small, and we can linearise, yielding:

$$n(x,t+1) = n(x,t) + b\sum_{y\neq x} K_{xy}\; n(y,t). \tag{A3.1}$$

If the kernel is not fat-tailed, we can parameterise the asymptotic behaviour as:

$$n \approx e^{-(kx-\omega t)} \tag{A3.2}$$

where k is unknown. Eq.(A3.1) yields a dispersion relation $\omega(k)$. It is a standard result that the dominant $k = k_c$ is given by the minimum value of $c(k) = \dfrac{\omega(k)}{k}$ [122, 124], where $v = c(k_c)$ and $c(k)$ is the speed of the wave with wave number k. For general kernels, substituting into Eq.(A3.1) yields:

$$e^{-[(kx-\omega(t+1)]} = e^{-(kx-\omega t)} + b\sum_{y\neq x} K_{xy} e^{-(ky-\omega t)}. \tag{A3.3}$$

and hence, given the symmetry of $K_{xy} = K_{yx} = K(|x-y|)$,

$$\begin{aligned} e^{\omega} &= 1 + b\sum_{y\neq x} K(|y-x|)e^{k(x-y)} && \text{(A3.4)}\\ &= 1 + b\sum_{n\neq 0} K(|n|)e^{kn}. && \text{(A3.5)} \end{aligned}$$

Taking logarithms and then applying the definition of the dispersion relation yields an expression for the velocity as a function of wave number.

$$c(k) = \frac{1}{k}\ln\left(1 + b\sum_{n\neq 0} K(n)e^{kn}\right). \tag{A3.6}$$

Differentiating (A3.4) in terms of k then implies

$$\begin{aligned} \frac{d}{dk}(ck)e^{ck_c} &= b\sum_{n\neq 0} nK(n)e^{k_c n} \\ \frac{\omega}{k_c} &= \frac{b\sum_{n\neq 0} nK(n)e^{k_c n}}{1+b\sum_{n\neq 0} K(n)e^{kn}} \\ \ln\left(1+b\sum_{n\neq 0} K(n)e^{k_c n}\right) &= \frac{b\sum_{n\neq 0}(nk_c)K(n)e^{k_c n}}{1+b\sum_{n\neq 0} K(n)e^{k_c n}}. \end{aligned} \tag{A3.7}$$

The value of k_c can be determined using Eq. (A3.7), and $c = c(k_c)$ is obtained by substituting this into Eq. (A3.6). If $b < 1$, one can consider this a change in the time scale by normalising the resulting velocity by a factor $\frac{1}{b}$.

In the nearest neighbour kernel case, Eq. (A3.4) reduces to

$$e^{\omega(k)} = 1 + \cosh k. \tag{A3.8}$$

Substituting Eq. (A3.8) into Eq.(A3.7), we obtain

$$\ln\left(1+\cosh k_c\right) = \frac{k_c \sinh k_c}{1+\cosh k_c},$$

with

$$c(k) = \frac{\ln\left(1+\cosh k_c\right)}{k_c} \tag{A3.9}$$

following Eq.(A3.6). Solving this equation numerically also yields $c(k_c) = v = 0.78$, which serves as a important check on the accuracy of the simulation results.

A4 Thesis Appendix 4: Selection candidate windows

Complete lists of the selection candidates identified from other studies [358, 454, 455, 361, 456, 457, 458, 6] are listed in the Tables A4, A4 and A4. For each candidate, a core position was defined as either the reported SNP position or the centre of the reported window. If necessary, this core position was convered to NCBI b36 coordinates using the UCSC Liftover tool. The selection candidate window was defined as the 200kb winow including this SNP. Genes whose transcripts overlap each window were identified using the UCSC Table Browser (track: RefSeq Genes).

StartReported	EndReported	Build	Core (NCBI b36)	Type	Selected Population	Population comparison	Method	Source	Candidate window	Genes
136000000		NCBI b36	136000000	WINDOW	CEU		CLR	[358]	136000000	ZRANB3,R3HDM1
84700000		NCBI b36	84700000	WINDOW	CEU		CLR	[358]	84600000	DNAH6
196000000		NCBI b36	196000000	WINDOW	CEU		CLR	[358]	196000000	
158000000		NCBI b36	158000000	WINDOW	CEU		CLR	[358]	158000000	CYTIP,ACVR1C
122000000		NCBI b36	122000000	WINDOW	CEU		CLR	[358]	122000000	CLASP1,RNU4ATAC
21800000		NCBI b36	21800000	WINDOW	CEU		CLR	[358]	21800000	
74700000		NCBI b36	74700000	WINDOW	CEU		CLR	[358]	74600000	DQX1,AUP1
24500000		NCBI b36	24500000	WINDOW	CEU		CLR	[358]	24400000	ITSN2
139000000		NCBI b36	139000000	WINDOW	CEU		CLR	[358]	139000000	SPOPL,NXPH2
51700000		NCBI b36	51700000	WINDOW	CEU		CLR	[358]	51600000	
148000000		NCBI b36	148000000	WINDOW	CEU		CLR	[358]	148000000	
39400000		NCBI b36	39400000	WINDOW	CEU		CLR	[358]	39400000	MAP4K3,LOC728730
135000000		NCBI b36	135000000	WINDOW	CEU		CLR	[358]	135000000	TMEM163
228000000		NCBI b36	228000000	WINDOW	CEU		CLR	[358]	228000000	MIR5703
176000000		NCBI b36	176000000	WINDOW	CEU		CLR	[358]	176000000	
68800000		NCBI b36	68800000	WINDOW	CEU		CLR	[358]	68800000	ARHGAP25
144000000		NCBI b36	144000000	WINDOW	CEU		CLR	[358]	144000000	ARHGAP15
193000000		NCBI b36	193000000	WINDOW	CEU		CLR	[358]	193000000	
223000000		NCBI b36	223000000	WINDOW	CEU		CLR	[358]	223000000	SGPP2,FARSB
163000000		NCBI b36	163000000	WINDOW	CEU		CLR	[358]	163000000	KCNH7
132000000		NCBI b36	132000000	WINDOW	CEU		CLR	[358]	132000000	RNU6-81P,CCDC74A
164000000		NCBI b36	164000000	WINDOW	CEU		CLR	[358]	164000000	FIGN
116000000		NCBI b36	116000000	WINDOW	CEU		CLR	[358]	116000000	DPP10,DPP10,DPP10
72368190		GR37	72221698	SNP	Europe 1000G	Africa 1000G	DDAF	[458]	72200000	CYP26B1
158126458		GR37	157834704	SNP	Europe 1000G	Africa 1000G	DDAF	[458]	157800000	GALNT5
215975232		GR37	215683477	SNP	Europe 1000G	Africa 1000G	DDAF	[458]	215600000	ABCA12
10618570		GR37	10536021	SNP	Europe 1000G	East Asian 1000G	DDAF	[458]	10400000	HPCAL1,ODC1
104960655		GR37	104327087	SNP	Europe 1000G	Africa 1000G	DDAF	[458]	104200000	LOC100287010
29980408		GR37	29833912	SNP	Europe 1000G	Africa 1000G	DDAF	[458]	29800000	ALK
236886196		GR37	236550935	SNP	Europe 1000G	Africa 1000G	DDAF	[458]	236400000	AGAP1
71314521		GR37	71168029	SNP	Europe 1000G	East Asian 1000G	DDAF	[458]	71000000	VAX2,ATP6V1B1
112190331		GR37	111906802	SNP	CEU 1000G	FINland	DDAF	[458]	111800000	MIR4435-2HG
43237764		GR37	43091268	SNP	CEU 1000G	FINland	DDAF	[458]	43000000	LOC102723854
225843666		GR37	225551910	INDEL	CEU 1000G	FINland	DDAF	[458]	225400000	DOCK10,MIR4439
87947360		GR37	87728475	SNP	CEU 1000G	FINland	DDAF	[458]	87600000	LINC00152,MIR4435-1
193251166		GR37	192959411	INDEL	CEU 1000G	FINland	DDAF	[458]	192800000	
11220562		GR37	11138013	SNP	CEU 1000G	FINland	DDAF	[458]	11000000	FLJ33534
87929798		GR37	87710913	SNP	CEU 1000G	GBR	DDAF	[458]	87600000	LINC00152,MIR4435-1
237701947		GR37	237366686	INDEL	CEU 1000G	TSI	DDAF	[458]	237200000	
181566975		GR37	181275220	SNP	CEU 1000G	FINland	DDAF	[458]	181200000	SCHLAP1
71816602		GR37	71670110	SNP	CEU 1000G	TSI	DDAF	[458]	71600000	DYSF
221351731		GR37	221059975	INDEL	CEU 1000G	GBR	DDAF	[458]	221000000	
13690000	13900000	GR37	13712451	WINDOW	CEU	YRI	XP-SFselect	[457]	13600000	
150390000	150490000	GR37	150148246	WINDOW	CEU	YRI	XP-SFselect	[457]	150000000	LYPD6,MMADHC
97680000	97850000	GR37	97128727	WINDOW	CEU	YRI	XP-SFselect	[457]	97000000	FAM178B,FAHD2B
104760000	104830000	GR37	104161432	WINDOW	CEU	YRI	XP-SFselect	[457]	104000000	
167500000	167600000	GR37	167258246	WINDOW	CEU	YRI	XP-SFselect	[457]	167200000	
39,341,697		NCBI b35	39283550	WINDOW	European American	African American	S2Fst	[455]	39200000	SOS1,CDKL4
71,951,007		NCBI b35	71892860	WINDOW	European American	African American	S2Fst	[455]	71800000	
72,023,156		NCBI b35	71965009	WINDOW	European American	African American	S2Fst	[455]	71800000	
74,684,965		NCBI b35	74626818	WINDOW	European American	African American	S2Fst	[455]	74600000	DQX1,AUP1
84,900,804		NCBI b35	84842657	WINDOW	European American	African American	S2Fst	[455]	84800000	DNAH6,TRABD2A
114,364,004		NCBI b35	114364244	WINDOW	European American	African American	S2Fst	[455]	114200000	SLC35F5,LOC101060091
121,770,798		NCBI b35	121771038	WINDOW	European American	African American	S2Fst	[455]	121600000	TFCP2L1
152,484,853		NCBI b35	152367591	WINDOW	European American	African American	S2Fst	[455]	152200000	NEB,ARL5A
158,397,794		NCBI b35	158280532	WINDOW	European American	African American	S2Fst	[455]	158200000	ACVR1
163,109,385		NCBI b35	162992124	WINDOW	European American	African American	S2Fst	[455]	162800000	FAP,IFIH1
219,312,501		NCBI b35	219195240	WINDOW	European American	African American	S2Fst	[455]	219000000	VIL1,USP37
232,679,638		NCBI b35	232562377	WINDOW	European American	African American	S2Fst	[455]	232400000	MIR1471
238,452,816		NCBI b35	238335555	WINDOW	European American	African American	S2Fst	[455]	238200000	LRRFIP1
74300000	74870000	NCBI b36	74585000	WINDOW	CEU	YRI	XP-CLR	[456]	74400000	SLC4A5,DCTN1
72400000	73050000	NCBI b36	72725000	WINDOW	CEU	YRI	XP-CLR	[456]	72600000	EXOC6B
121000000	121400000	NCBI b36	121200000	WINDOW	CEU	YRI	XP-CLR	[456]	121200000	GLI2
21400000	21800000	NCBI b36	21600000	WINDOW	CEU		CLR	[361]	21600000	LOC645949
84470000	84810000	NCBI b35	84581853	WINDOW	European CEPH		Tajima's D	[454]	84400000	SUCLG1
162820000	163240000	NCBI b35	162912739	WINDOW	European CEPH		Tajima's D	[454]	162800000	FAP,IFIH1
163000000	163199999	GR37	162758246	WINDOW	SWEuropean		Tajima's D	[6]	162600000	DPP4,LOC101929532
182600000	182799999	GR37	182358245	WINDOW	SWEuropean		Tajima's D	[6]	182200000	CERKL,NEUROD1
92000000	92199999	GR37	91413727	WINDOW	SWEuropean		Tajima's D	[6]	91400000	ACTR3BP2
194800000	194999999	GR37	194558245	WINDOW	SWEuropean		Tajima's D	[6]	194400000	
163200000	163399999	GR37	162958246	WINDOW	SWEuropean		Tajima's D	[6]	162800000	FAP,IFIH1
158200000	158399999	GR37	157958246	WINDOW	SWEuropean		Tajima's D	[6]	157800000	GALNT5

Table A1: Selection candidates used for the CEU population, Chromosome 2

StartReported	EndReported	Build	Core (NCBI b36)	Type	Selected Population	Population comparison	Method	Source	Candidate window	Genes
48426484		GR37	46213776	SNP	Europe 1000G	East Asian 1000G	DDAF	[458]	46200000	SLC24A5
48392165		GR37	46179457	SNP	Europe 1000G	African 1000G	DDAF	[458]	46000000	
34258834		GR37	32046126	SNP	Europe 1000G	African 1000G	DDAF	[458]	32000000	AVEN,CHRM5
93585347		GR37	91386351	SNP	Europe 1000G	East Asian 1000G	DDAF	[458]	91200000	LINC01578
90513342		GR37	88314346	SNP	Europe 1000G	African 1000G	DDAF	[458]	88200000	AP3S2,C15orf38-AP3S2,AP3S2,ARPIN
44290000	44390000	GR37	42127292	WINDOW	CEU	YRI	XP-SFselect (SFS)	[457]	42000000	FRMD5
28190000	28270000	GR37	25903595	WINDOW	CEU	YRI	XP-SFselect (SFS)	[457]	25800000	OCA2
23,097,946		NCBIb35	23097946	WINDOW	European American	African American	S2Fst	[455]	23000000	SNORD115-19,SNORD115-18
27,184,680		NCBIb35	27184680	WINDOW	European American	African American	S2Fst	[455]	27000000	APBA2
42,231,163		NCBIb35	42231163	WINDOW	European American	African American	S2Fst	[455]	42200000	FRMD5,CASC4
67,155,853		NCBIb35	67155853	WINDOW	European American	African American	S2Fst	[455]	67000000	SPESP1
70,466,010		NCBIb35	70466010	WINDOW	European American	African American	S2Fst	[455]	70400000	HEXA
46200000	46330000	NCBI b36	46265000	WINDOW	CEU	YRI	XP-CLR	[456]	46200000	SLC24A5
27000000	27190000	NCBI b36	27095000	WINDOW	CEU	YRI	XP-CLR	[456]	27000000	APBA2
42900000	43210000	NCBI b36	43055000	WINDOW	CEU	YRI	XP-CLR	[456]	43000000	C15orf43
25900000	26230000	NCBI b36	26065000	WINDOW	CEU	YRI	XP-CLR	[456]	26000000	OCA2,HERC2
67400000	67800000	NCBI b36	67600000	WINDOW	CEU		CLR	[361]	67600000	DRAIC
44200000	44399999	GR37	42037292	WINDOW	SWEuropean		Tajima's D	[6]	42000000	FRMD5
44600000	44799999	GR37	42437292	WINDOW	SWEuropean		Tajima's D	[6]	42400000	CASC4,CTDSPL2
72600000	72799999	GR37	70437054	WINDOW	SWEuropean		Tajima's D	[6]	70400000	HEXA
44800000	44999999	GR37	42637292	WINDOW	SWEuropean		Tajima's D	[6]	42600000	CTDSPL2,EIF3J-AS1

Table A2: Selection candidates used for the CEU population, Chromosome 15

StartReported	EndReported	Build	Core (NCBI b36)	Type	Selected Population	Population comparison	Method	Source	Candidate window	Genes
72826665		GR37	72680173	SNP	East Asian 1000G	African 1000G	DDAF	[458]	72600000	EXOC6B
9551707		GR37	9469158	SNP	East Asian 1000G	African 1000G	DDAF	[458]	9400000	ASAP2,ITGB1BP1
109543883		GR37	108910315	SNP	East Asian 1000G	Europe 1000G	DDAF	[458]	108800000	CCDC138,EDAR
53181448		GR37	53034952	SNP	East Asian 1000G	African 1000G	DDAF	[458]	53000000	
109006665		GR37	108373097	SNP	East Asian 1000G	African 1000G	DDAF	[458]	108200000	LINC01594,SULT1C3
214009667		GR37	213717912	SNP	East Asian 1000G	African 1000G	DDAF	[458]	213600000	IKZF2
145715873		GR37	145432343	SNP	East Asian 1000G	African 1000G	DDAF	[458]	145400000	TEX41
26113913		GR37	25967417	SNP	East Asian 1000G	Europe 1000G	DDAF	[458]	25800000	ASXL2
216321788		GR37	216030033	SNP	East Asian 1000G	Europe 1000G	DDAF	[458]	216000000	FN1,LOC102724849
125859777		GR37	125576247	SNP	East Asian 1000G	Europe 1000G	DDAF	[458]	125400000	
242087712		GR37	241736385	SNP	East Asian 1000G	Europe 1000G	DDAF	[458]	241600000	SNED1,MTERF4
154896323		GR37	154604569	SNP	East Asian 1000G	Europe 1000G	DDAF	[458]	154600000	GALNT13
72330000	73030000	NCBI b36	72680000	WINDOW	CHB+JPT	YRI	XP-CLR	[456]	72600000	EXOC6B
177260000	177700000	NCBI b36	177480000	WINDOW	CHB+JPT	YRI	XP-CLR	[456]	177400000	
121390000	121440000	NCBI b36	121415000	WINDOW	CHB+JPT	YRI	XP-CLR	[456]	121400000	GLI2
177200000	177600000	NCBI b36	177400000	WINDOW	CHB+JPT		CLR	[361]	177400000	
84540000	84910000	NCBIb35	84666853	WINDOW	Chinese CEPH		Tajima's D	[454]	84600000	DNAH6
108350000	109120000	NCBIb35	108642914	WINDOW	Chinese CEPH		Tajima's D	[454]	108600000	LIMS1,RANBP2
177390000	177730000	NCBIb35	177442739	WINDOW	Chinese CEPH		Tajima's D	[454]	177400000	
189140000	189570000	NCBIb35	189237739	WINDOW	Chinese CEPH		Tajima's D	[454]	189200000	DIRC1
194650000	194990000	NCBIb35	194702739	WINDOW	Chinese CEPH		Tajima's D	[454]	194600000	

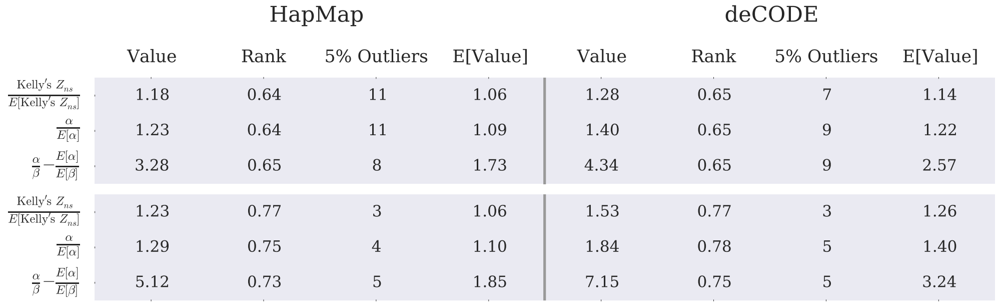
Table A3: Selection candidates used for the CHB+JPT population, Chromosome 2

A5 Thesis Appendix 5: Further selection scans and replication data

In the main text, we only presented replication results for a subset of the selection statistics tested. Full data on the selection candidates used and their resampled p-values for selection statistics is provided in Supplementary File 1. We here compare several more scans and present further results on the ability of different statistics to replicate selection candidate signals.

Comparing the Kong and HapMap genetic maps

Some studies have emphasised the potential for positive selection to distort estimates of the genetic map based on linkage disequilibrium [445]. We therefore expected selection statistics that controlled for expected LD based on the HapMap genetic map [390] to lose signals that might be recovered when using the doCODE map [376]. As illustrated in Table A4, and in Supplementary File 1, this was not obviously the case. While the value of each statistic was larger when using the doCODE map, the expected maximum 200kb window value also increased and the number of replicated signals (5% outliers) usually remained stable or decreased. We often observed a slight increase in statistic focussing on the Along region and controlling for LD when using the Kong map. We attribute this to the lower resolution of the deCODE map. When the genetic profile was being estimated, many pairs of SNPs separated by a relatively large physical distance, and with relatively low pairwise LD, were assigned a very small genetic map distance. This lead to substantially reduced estimates of average LD in short-range bins, which tended to create the impression of inflated deviations from expected LD for some statistics.



	HapMap				deCODE			
	Value	Rank	5% Outliers	E[Value]	Value	Rank	5% Outliers	E[Value]
$\frac{\text{Kelly's } Z_{ns}}{E[\text{Kelly's } Z_{ns}]}$	1.18	0.64	11	1.06	1.28	0.65	7	1.14
$\frac{\alpha}{E[\alpha]}$	1.23	0.64	11	1.09	1.40	0.65	9	1.22
$\frac{\alpha}{\beta} - \frac{E[\alpha]}{E[\beta]}$	3.28	0.65	8	1.73	4.34	0.65	9	2.57
$\frac{\text{Kelly's } Z_{ns}}{E[\text{Kelly's } Z_{ns}]}$	1.23	0.77	3	1.06	1.53	0.77	3	1.26
$\frac{\alpha}{E[\alpha]}$	1.29	0.75	4	1.10	1.84	0.78	5	1.40
$\frac{\alpha}{\beta} - \frac{E[\alpha]}{E[\beta]}$	5.12	0.73	5	1.85	7.15	0.75	5	3.24

Table A4: Replication of previously identified selection candidates when controlling for expected LD based on the HapMap and deCODE genetic maps

Example selection scans using the $\frac{\alpha}{E[\alpha]}$ statistic and both the deCODE and HapMap genetic maps are shown in Fig. A1. As expected, the two scans are highly correlated. There is no clear effect on the strength of the high-confidence selection signals (*MCM6*, *HERC2*, *SLC25A4*).

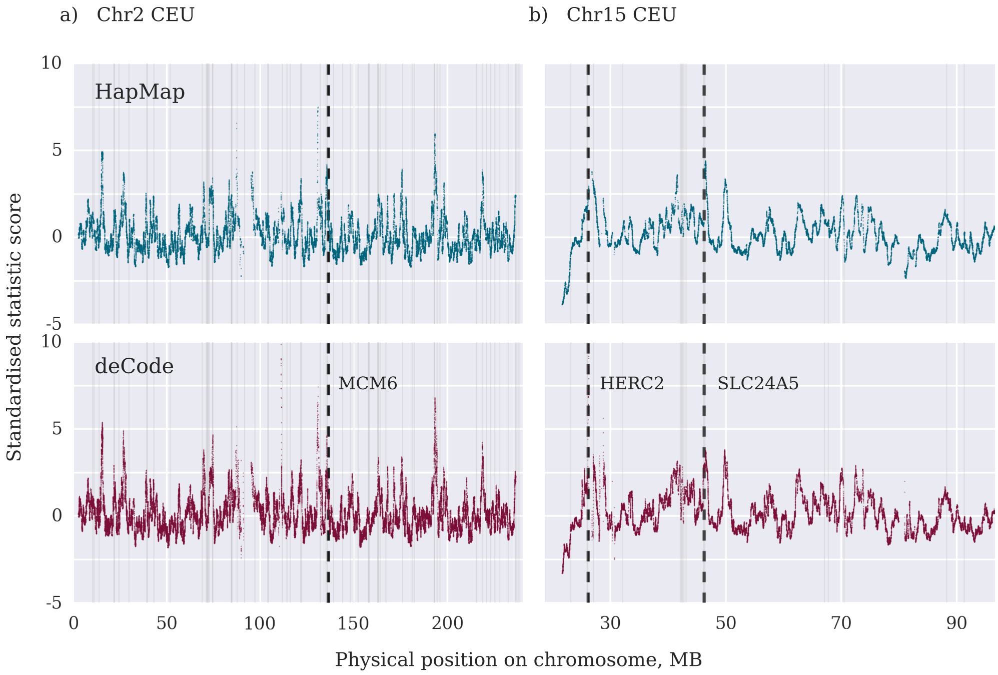


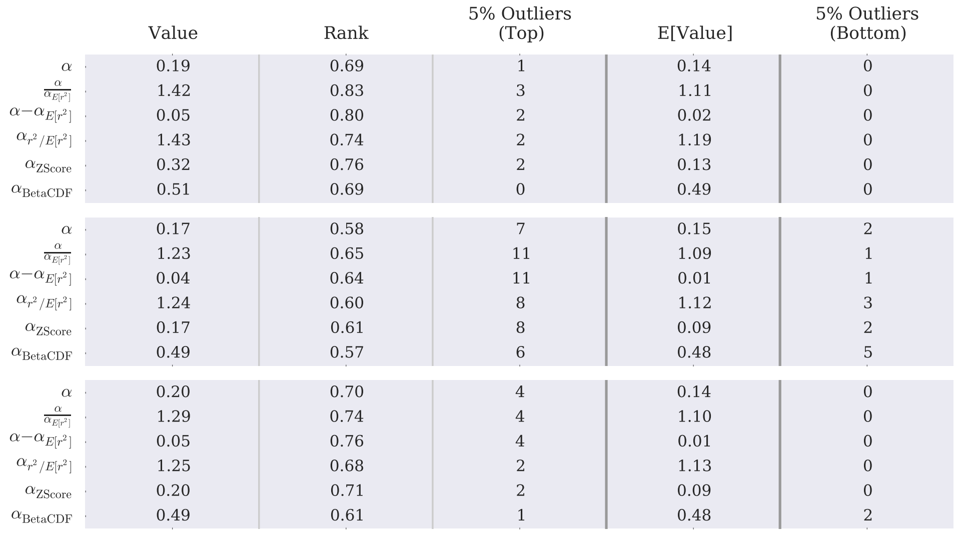
Figure A1: Standardised selection scan using the $\frac{\alpha}{E[\alpha]}$ statistic, using both the deCODE and HapMap genetic maps as labelled.

Comparing methods of controlling for expected LD

As mentioned in the Extended Methods, we investigated a range of different approaches to modifying selection statistics such that expected LD was controlled for. We present details on the ability of different approaches to replicate signals at selection candidates in Tables A5 (based on LD in the Along region) and A6 (based on LD in the Over region). The most interesting result is the extent to which different approaches differ in the number of signals they replicate, despite being based on similar data. The selection scans themselves are shown in Figures A2 and A3 respectively, which give an indication of the extent of correlation between the different statistics.

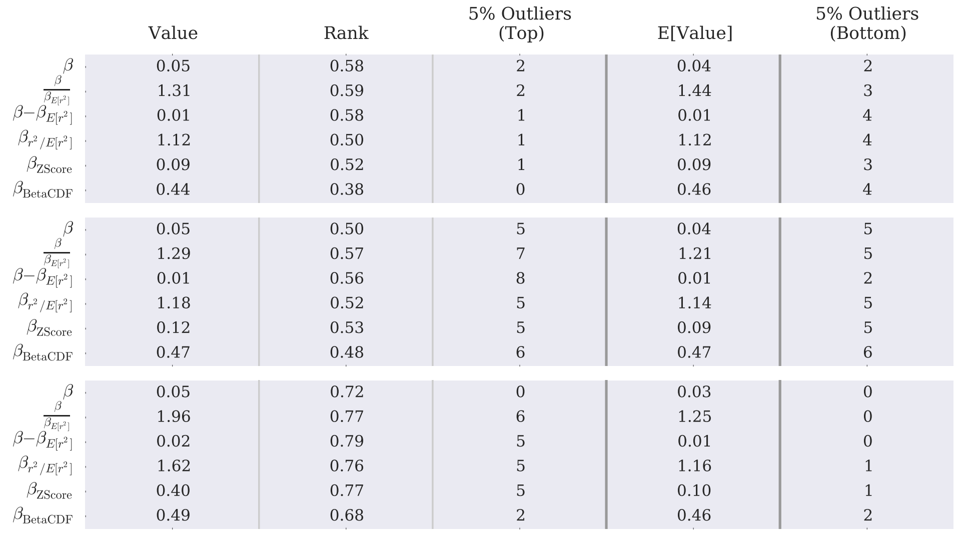
Controlling for expected LD and assessing the impact of window size on OmegaPlus signal replication

The ability of OmegaPlus to replicate previously identified selection candidates using different window sizes is shown in Table A7, as is the performance of the $\frac{\alpha}{\beta} - E[\frac{\alpha}{\beta}]$ statistic which uses a fixed 400kb window size to detects a similar signal but attempts to control for expected LD. The different maximum and minimum window sizes have a strong impact on the raw maximum value of $\omega_{\max}$ in a 200kb window. Larger values occur when the minimum window size is smaller, which likely reflects both observations made in Appendix 2 and the potential for greater variance in the denominator of ω to generate spikes in its value. Nevertheless, the ability of the OmegaPlus algorithm to



	Value	Rank	5% Outliers (Top)	E[Value]	5% Outliers (Bottom)
α	0.19	0.69	1	0.14	0
$\frac{\alpha}{\alpha_{E[r^2]}}$	1.42	0.83	3	1.11	0
$\alpha - \alpha_{E[r^2]}$	0.05	0.80	2	0.02	0
$\alpha_{r^2/E[r^2]}$	1.43	0.74	2	1.19	0
α_{ZScore}	0.32	0.76	2	0.13	0
α_{BetaCDF}	0.51	0.69	0	0.49	0
α	0.17	0.58	7	0.15	2
$\frac{\alpha}{\alpha_{E[r^2]}}$	1.23	0.65	11	1.09	1
$\alpha - \alpha_{E[r^2]}$	0.04	0.64	11	0.01	1
$\alpha_{r^2/E[r^2]}$	1.24	0.60	8	1.12	3
α_{ZScore}	0.17	0.61	8	0.09	2
α_{BetaCDF}	0.49	0.57	6	0.48	5
α	0.20	0.70	4	0.14	0
$\frac{\alpha}{\alpha_{E[r^2]}}$	1.29	0.74	4	1.10	0
$\alpha - \alpha_{E[r^2]}$	0.05	0.76	4	0.01	0
$\alpha_{r^2/E[r^2]}$	1.25	0.68	2	1.13	0
α_{ZScore}	0.20	0.71	2	0.09	0
α_{BetaCDF}	0.49	0.61	1	0.48	2

TABLE A5: Ability of statistics based on LD in the Along region to replicate selection signals, with expected LD controlled for according to a variety of methods using the HapMap genetic map. From top to bottom, the tables correspond to CHB+JPT (Chr2), CEU (Chr2) and CEU (Chr15).



	Value	Rank	5% Outliers (Top)	E[Value]	5% Outliers (Bottom)
β	0.05	0.58	2	0.04	2
$\frac{\beta}{\beta_{E[r^2]}}$	1.31	0.59	2	1.44	3
$\beta - \beta_{E[r^2]}$	0.01	0.58	1	0.01	4
$\beta_{r^2/E[r^2]}$	1.12	0.50	1	1.12	4
β_{ZScore}	0.09	0.52	1	0.09	3
β_{BetaCDF}	0.44	0.38	0	0.46	4
β	0.05	0.50	5	0.04	5
$\frac{\beta}{\beta_{E[r^2]}}$	1.29	0.57	7	1.21	5
$\beta - \beta_{E[r^2]}$	0.01	0.56	8	0.01	2
$\beta_{r^2/E[r^2]}$	1.18	0.52	5	1.14	5
β_{ZScore}	0.12	0.53	5	0.09	5
β_{BetaCDF}	0.47	0.48	6	0.47	6
β	0.05	0.72	0	0.03	0
$\frac{\beta}{\beta_{E[r^2]}}$	1.96	0.77	6	1.25	0
$\beta - \beta_{E[r^2]}$	0.02	0.79	5	0.01	0
$\beta_{r^2/E[r^2]}$	1.62	0.76	5	1.16	1
β_{ZScore}	0.40	0.77	5	0.10	1
β_{BetaCDF}	0.49	0.68	2	0.46	2

TABLE A6: Ability of statistics based on LD in the Over region to replicate selection signals, with expected LD controlled for according to a variety of methods using the HapMap genetic map. From top to bottom, the tables correspond to CHB+JPT (Chr2), CEU (Chr2) and CEU (Chr15).

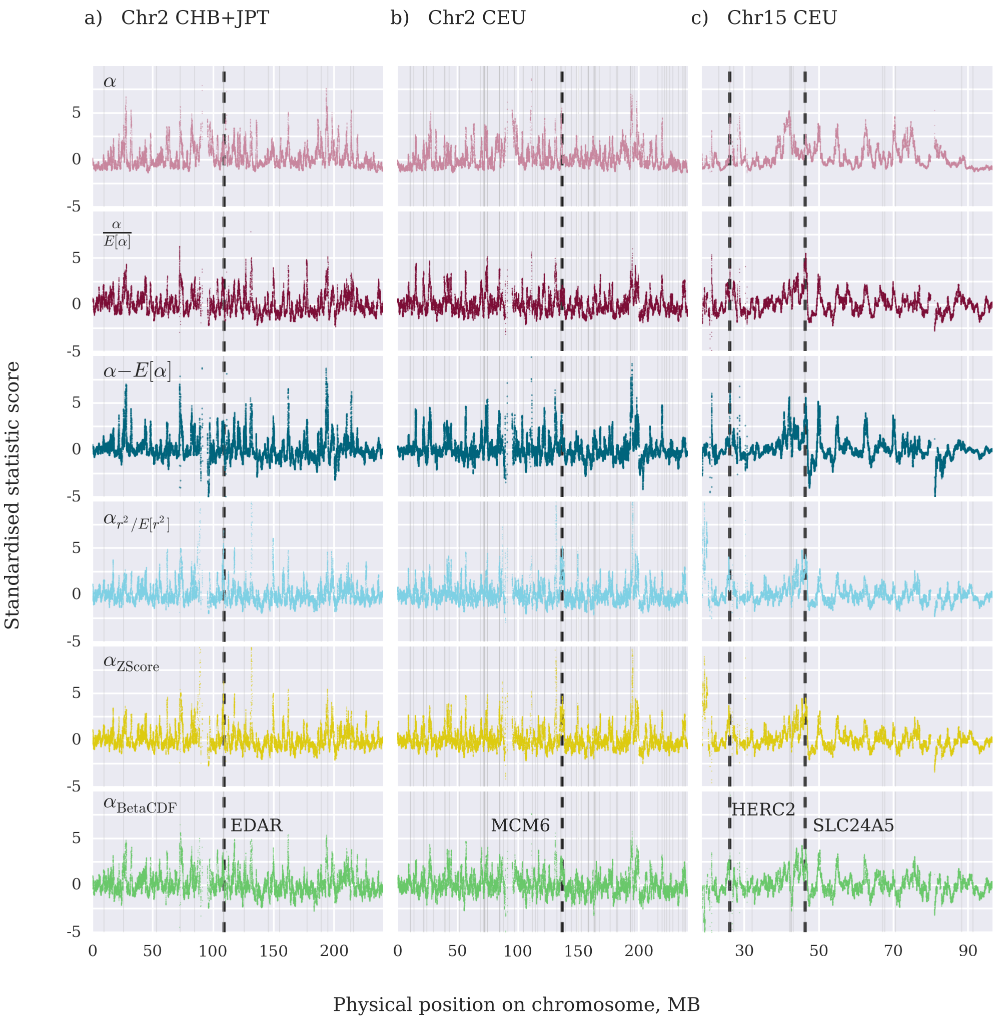


Figure A2: Standardised selection scans for statistics based on LD in the Along region, with expected LD controlled for in various ways using the HapMap genetic map

replicate signals was not strongly affected.

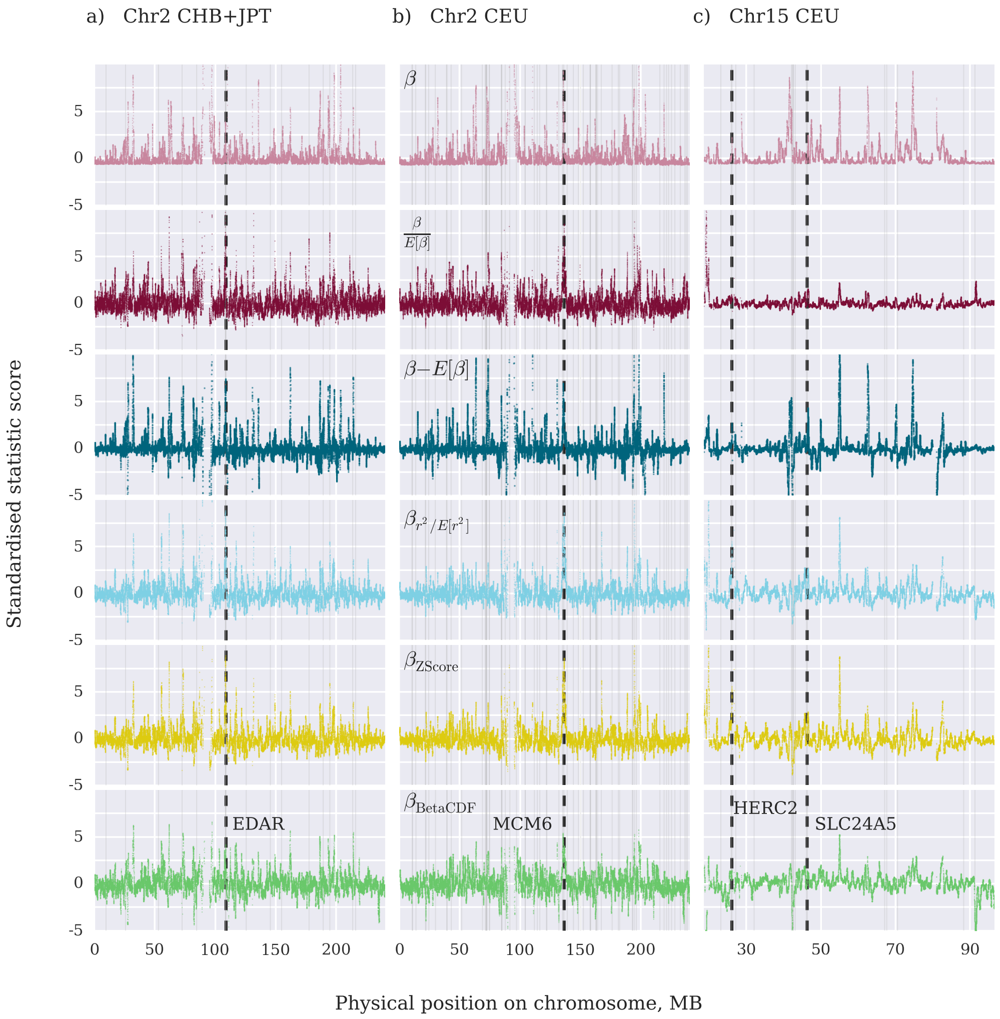


FIGURE A3: Standardised selection scans for statistics based on LD in the Over region, with expected LD controlled for in various ways using the HapMap genetic map. The ‘compressed’ signal for $\frac{\beta}{E[\beta]}$ in CEU Chr15 is caused by an extreme peak centered at position 19244303.

	Value Max	5% Outliers (Top)	E[Value Max]	Value Min	5% Outliers (Bottom)	E[Value Min]
$\frac{\alpha}{\beta} - E[\frac{\alpha}{\beta}]$	2.23	1	2.80	-2.44	3	-1.63
$\omega_{\max}^{10,400}$	98.65	3	43.16	15.76	0	5.23
$\omega_{\max}^{50,400}$	60.80	2	28.44	15.76	0	5.23
$\omega_{\max}^{100,400}$	52.95	3	22.96	15.76	0	5.20
$\omega_{\max}^{200,200}$	14.72	2	14.62	2.23	3	2.82
$\frac{\alpha}{\beta} - E[\frac{\alpha}{\beta}]$	5.12	5	1.84	-0.67	2	-1.84
$\omega_{\max}^{10,400}$	109.72	4	34.41	4.16	1	4.08
$\omega_{\max}^{50,400}$	76.70	5	20.55	4.16	1	4.09
$\omega_{\max}^{100,400}$	38.56	5	15.69	4.15	1	4.06
$\omega_{\max}^{200,200}$	15.60	4	9.39	1.93	0	2.44
$\frac{\alpha}{\beta} - E[\frac{\alpha}{\beta}]$	3.25	8	1.71	-0.71	3	-0.92
$\omega_{\max}^{10,400}$	53.75	8	30.30	4.37	5	3.92
$\omega_{\max}^{50,400}$	26.36	8	17.87	4.37	5	3.91
$\omega_{\max}^{100,400}$	19.52	8	13.97	4.36	5	3.88
$\omega_{\max}^{200,200}$	11.98	7	9.47	2.46	5	2.33

TABLE A7: Ability of the OmegaPlus program to replicate selection signals using different window sizes, and for statistic that searches for a similar signal but controls for expected LD (with window size fixed at 400kb) to do so. From top to bottom, the tables correspond to CHB+JPT (Chr2), CEU (Chr2) and CEU (Chr15).

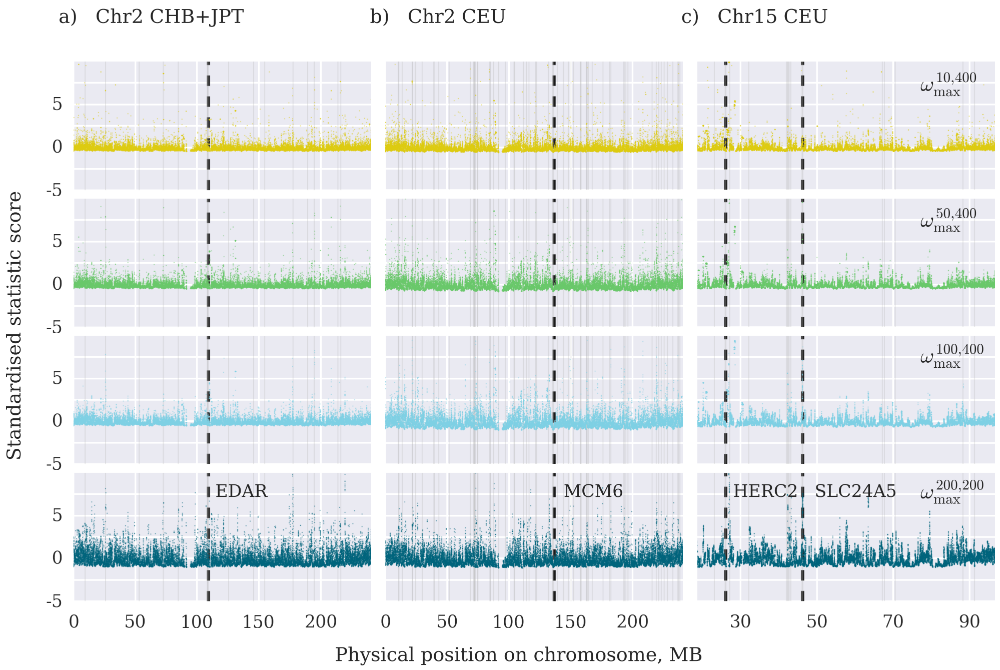


FIGURE A4: Standardised selection scans using $\omega_{\max}$ and a variety of variable and fixed window sized.

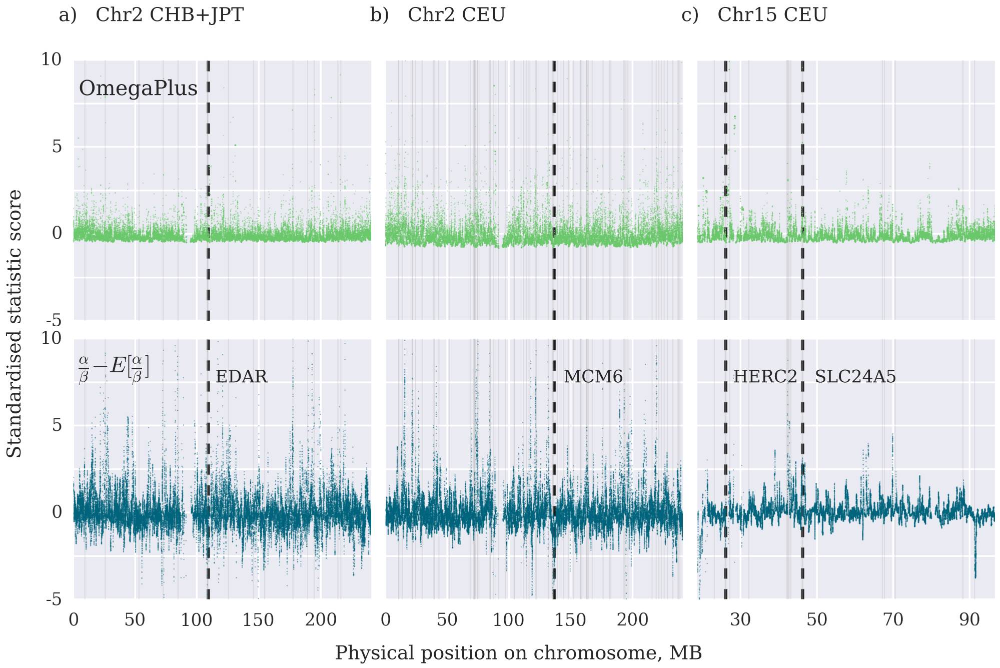


FIGURE A5: Standardised selection scans using $\omega_{\text{max}}^{50,400}$ and a fixed window-size statistic expected to capture the same signal of selection with expected LD controlled for.

Detailed replication data

Detailed data on selection candidate replication by a set of LD-based statistics is shown in Table A8. Approximate p-values were obtained by resampling 10,000 random 200kb windows for the relevant chromosome and population, and obtaining the maximum (or minimum for $|L||R|$) statistic value each time. These p-values correspond closely to those used when constructing Tables 4.3, 4.5, A4, A5, A6 and A7 but are not identical.

Table A8: Replication of selection candidate windows in HapMap Phase II data. Windows containing protein-coding genes in Chromosome 2 are highlighted in green; positions indicate the starting NCBI b36 window position and are in MB

Pop	Chr	Pos	Source	Kelly's	HapMap	Kong	α	HapMap	Kong	$\lvert L\rvert\lvert R\rvert$	$\omega_{\max}^{50,400}$	HapMap	Kong
				Z_{nS}	$\frac{Z_{nS}}{E[Z_{nS}]}$	$\frac{Z_{nS}}{E[Z_{nS}]}$		$\frac{\alpha}{E[\alpha]}$	$\frac{\alpha}{E[\alpha]}$			$\frac{\alpha}{\beta}-E[\frac{\alpha}{\beta}]$	$\frac{\alpha}{\beta}-E[\frac{\alpha}{\beta}]$
CEU	2	10.4	[458]	0.730	0.347	0.163	0.829	0.426	0.241	0.718	0.152	0.518	0.342
CEU	2	11	[458]	0.674	0.308	0.245	0.675	0.342	0.355	0.670	0.318	0.329	0.252
CEU	2	13.6	[457]	0.538	0.266	0.395	0.526	0.360	0.563	0.627	0.425	0.647	0.932
CEU	2	21.6	[361]	0.202	0.048	0.147	0.207	0.049	0.049	0.195	0.011	0.024	0.049
CEU	2	21.8	[358]	0.179	0.050	0.164	0.136	0.027	0.111	0.149	0.284	0.007	0.016
CEU	2	24.4	[358]	0.267	0.671	0.447	0.225	0.625	0.436	0.261	0.127	0.056	0.052
CEU	2	29.8	[458]	0.683	0.265	0.371	0.731	0.279	0.396	0.835	0.378	0.310	0.516
CEU	2	39.2	[455]	0.104	0.110	0.235	0.085	0.063	0.171	0.176	0.701	0.145	0.181
CEU	2	39.4	[358]	0.155	0.105	0.267	0.214	0.082	0.130	0.247	0.865	0.384	0.312
CEU	2	43	[458]	0.548	0.043	0.068	0.514	0.059	0.113	0.773	0.313	0.486	0.847
CEU	2	51.6	[358]	0.693	0.791	0.881	0.710	0.734	0.879	1.000	0.572	0.270	0.497
CEU	2	68.8	[358]	0.837	0.228	0.323	0.750	0.211	0.230	0.783	0.262	0.509	0.457
CEU	2	71	[458]	0.669	0.621	0.583	0.630	0.836	0.718	0.811	0.554	0.913	0.874
CEU	2	71.6	[458]	0.839	0.737	0.593	0.820	0.729	0.619	0.810	0.479	0.524	0.447
CEU	2	71.8	[455]	0.894	0.310	0.601	0.896	0.354	0.643	0.807	0.665	0.535	0.536
CEU	2	72.2	[458]	0.126	0.193	0.693	0.097	0.210	0.338	0.410	0.198	0.107	0.076
CEU	2	72.6	[456]	0.015	0.129	0.083	0.035	0.189	0.180	0.190	0.997	0.835	0.878
CEU	2	74.4	[456]	0.133	0.009	0.030	0.062	0.006	0.020	0.101	0.845	0.211	0.256
CEU	2	74.6	[358, 455]	0.275	0.029	0.047	0.058	0.004	0.014	0.172	0.667	0.026	0.031
CEU	2	84.4	[454]	0.082	0.007	0.069	0.075	0.020	0.093	0.546	0.559	0.073	0.082
CEU	2	84.6	[358]	0.067	0.005	0.060	0.076	0.034	0.062	0.374	0.916	0.790	0.737
CEU	2	84.8	[455]	0.079	0.006	0.092	0.113	0.018	0.110	0.364	0.042	0.059	0.060
CEU	2	87.6	[458]	0.237	0.215	0.015	0.296	0.433	0.035	0.004	0.020	0.409	0.037
CEU	2	91.4	[6]								0.988		
CEU	2	97	[457]	0.038	0.199	0.331	0.029	0.135	0.139	0.023	0.990	0.644	0.713
CEU	2	104	[457]	0.593	0.888	0.410	0.648	0.790	0.379	0.441	0.535	0.264	0.221
CEU	2	104.2	[458]	0.762	0.528	0.222	0.757	0.625	0.318	0.436	0.393	0.511	0.612
CEU	2	1118	[458]	0.651	0.722	0.472 .	0.588	0.736	0.636	0.090	0.170	0.339	0.248
CEU	2	114.2	[455]	0.282	0.692	0.886	0.364	0.829	0.869	0.123	0.921	0.574	0.646
CEU	2	116	[358]	0.223	0.652	0.742	0.226	0.549	0.699	0.562	0.183	0.053	0.029
CEU	2	121.2	[456]	0.905	0.093	0.123	0.930	0.093	0.147	0.470	0.007	0.197	0.247
CEU	2	121.6	[455]	0.225	0.025	0.030	0.044	0.040	0.037	0.347	0.092	0.000	0.003
CEU	2	122	[358]	0.057	0.077	0.088	0.019	0.089	0.084	0.180	0.667	0.228	0.123
CEU	2	132	[358]	0.277	0.404	0.115	0.261	0.179	0.061	0.045	0.458	0.027	0.035
CEU	2	135	[358]	0.397	0.098	0.027	0.474	0.173	0.068	0.354	0.619	0.976	0.976
CEU	2	136	[358]	0.015	0.088	0.091	0.007	0.140	0.078	0.068	0.961	0.760	0.719
CEU	2	139	[358]	0.636	0.973	0.945	0.570	0.927	0.941	0.529	0.548	0.642	0.915
CEU	2	144	[358]	0.770	0.847	0.879	0.824	0.883	0.903	0.371	0.653	0.743	0.467
CEU	2	148	[358]	0.195	0.201	0.214	0.117	0.162	0.277	0.391	0.498	0.257	0.256
CEU	2	150	[457]	0.554	0.660	0.793	0.495	0.486	0.745	0.425	0.266	0.166	0.225
CEU	2	152.2	[455]	0.277	0.410	0.229	0.277	0.403	0.271	0.243	0.056	0.086	0.061
CEU	2	157.8	[458, 6]	0.452	0.349	0.541	0.479	0.370	0.583	0.128	0.217	0.501	0.382
CEU	2	158	[358]	0.657	0.331	0.542	0.710	0.329	0.542	0.137	0.055	0.148	0.143

CEU	2	158.2	[455]	0.876	0.270	0.524	0.849	0.258	0.508	0.179	0.298	0.175	0.135
CEU	2	162.6	[6]	0.330	0.569	0.077	0.318	0.208	0.044	0.044	0.151	0.033	0.048
CEU	2	162.8	[455, 454, 6]	0.403	0.203	0.077	0.313	0.208	0.044	0.050	0.064	0.035	0.050
CEU	2	163	[358]	0.197	0.147	0.071	0.235	0.100	0.031	0.051	0.012	0.079	0.076
CEU	2	164	[358]	0.628	0.387	0.287	0.477	0.323	0.223	0.210	0.507	0.167	0.129
CEU	2	167.2	[457]	0.322	0.615	0.641	0.395	0.631	0.745	0.380	0.571	0.608	0.639
CEU	2	176	[358]	0.229	0.179	0.081	0.241	0.162	0.099	0.431	0.807	0.300	0.351
CEU	2	181.2	[458]	0.493	0.906	0.717	0.434	0.891	0.767	0.563	0.257	0.213	0.166
CEU	2	182.2	[6]	0.499	0.467	0.382	0.491	0.350	0.261	0.339	0.002	0.090	0.134
CEU	2	192.8	[458]	0.293	0.055	0.035	0.168	0.044	0.026	0.174	0.317	0.087	0.072
CEU	2	193	[358]	0.082	0.032	0.012	0.051	0.025	0.019	0.123	0.024	0.017	0.013
CEU	2	194.4	[6]	0.002	0.024	0.049	0.003	0.006	0.019	0.037	0.281	0.571	0.391
CEU	2	196	[358]	0.210	0.204	0.226	0.178	0.279	0.352	0.224	0.131	0.192	0.110
CEU	2	215.6	[458]	0.850	0.671	0.645	0.833	0.673	0.662	0.538	0.116	0.390	0.276
CEU	2	219	[455]	0.036	0.014	0.023	0.029	0.044	0.033	0.072	0.580	0.146	0.562
CEU	2	221	[458]	0.683	0.535	0.548	0.667	0.462	0.499	0.861	0.111	0.162	0.161
CEU	2	223	[358]	0.890	0.834	0.863	0.859	0.853	0.914	0.579	0.054	0.182	0.373
CEU	2	225.4	[458]	0.595	0.711	0.774	0.547	0.706	0.571	0.565	0.256	0.222	0.095
CEU	2	228	[358]	0.969	0.736	0.668	0.975	0.855	0.742	0.731	0.652	0.953	0.943
CEU	2	232.4	[455]	0.352	0.996	0.969	0.565	0.997	0.985	0.248	0.596	0.881	0.912
CEU	2	236.4	[458]	0.615	0.054		0.671	0.070		0.289	0.025	0.090	
CEU	2	237.2	[458]	0.663	0.141		0.742	0.149		0.341	0.438	0.151	
CEU	2	238.2	[455]	0.757	0.097		0.778	0.172		0.673	0.511	0.760	
CEU	15	23	[455]	0.737	0.909	0.929	0.767	0.938	0.821	0.740	0.753	0.602	0.465
CEU	15	25.8	[457]	0.373	0.165	0.099	0.189	0.180	0.104	0.195	0.587	0.722	0.570
CEU	15	26	[456]	0.231	0.259	0.026	0.021	0.015	0.002	0.025	0.581	0.276	0.270
CEU	15	27	[455, 456]	0.214	0.057	0.015	0.272	0.052	0.016	0.041	0.011	0.054	0.063
CEU	15	32	[458]	0.647	0.270	0.397	0.604	0.367	0.430	0.497	0.242	0.166	0.186
CEU	15	42	[457, 6]	0.031	0.049	0.400	0.006	0.043	0.042	0.122	0.825	0.385	0.199
CEU	15	42.2	[455]	0.029	0.089	0.169	0.018	0.164	0.118	0.066	0.023	0.004	0.005
CEU	15	42.4	[6]	0.035	0.112	0.095	0.048	0.126	0.051	0.073	0.613	0.046	0.058
CEU	15	42.6	[6]	0.195	0.392	0.071	0.111	0.483	0.175	0.071	0.150	0.010	0.012
CEU	15	43	[456]	0.209	0.073	0.156	0.189	0.092	0.172	0.208	0.240	0.154	0.048
CEU	15	46	[458]	0.409	0.071	0.326	0.256	0.012	0.037	0.372	0.002	0.032	0.035
CEU	15	46.2	[458, 456]	0.090	0.004	0.002	0.134	0.010	0.034	0.251	0.001	0.029	0.041
CEU	15	67	[455]	0.352	0.242	0.171	0.401	0.270	0.275	0.320	0.333	0.275	0.258
CEU	15	67.6	[361]	0.518	0.316	0.269	0.494	0.240	0.317	0.657	0.355	0.417	0.451
CEU	15	70.4	[455, 6]	0.040	0.287	0.169	0.038	0.316	0.294	0.099	0.585	0.382	0.612
CEU	15	88.2	[458]	0.595	0.424	0.198	0.613	0.410	0.242	0.445	0.004	0.083	0.081
CEU	15	91.2	[458]	0.959	0.444	0.482	0.926	0.607	0.664	0.824	0.319	0.947	0.981
CHBJPT	2	94	[458]	0.462	0.072		0.537	0.102		0.555	0.679	0.584	
CHBJPT	2	25.8	[458]	0.061	0.023		0.044	0.044		0.081	0.603	0.266	
CHBJPT	2	53	[458]	0.550	0.265		0.652	0.323		0.955	0.552	0.817	
CHBJPT	2	72.6	[458, 456]								0.002		
CHBJPT	2	84.6	[454]	0.112	0.027		0.172	0.081		0.260	0.917	0.921	
CHBJPT	2	108.2	[458]	0.121	0.064		0.101	0.144		0.288	0.371	0.749	
CHBJPT	2	108.6	[454]	0.084	0.006		0.109	0.065		0.249	0.985	0.964	
CHBJPT	2	108.8	[458]	0.171	0.080		0.150	0.145		0.305	0.158	0.997	
CHBJPT	2	121.2	[456]	0.866	0.073		0.864	0.073		0.387	0.081	0.211	
CHBJPT	2	125.4	[458]	0.171	0.207		0.207	0.163		0.388	0.670	0.146	
CHBJPT	2	145.4	[458]	0.293	0.329		0.273	0.235		0.192	0.820	0.449	
CHBJPT	2	154.6	[458]	0.415	0.840		0.550	0.900		0.833	0.495	0.970	
CHBJPT	2	177.4	[456, 361, 454]	0.243	0.001		0.248	0.010		0.369	0.011	0.021	
CHBJPT	2	189.2	[454]	0.134	0.225		0.150	0.229		0.275	0.861	0.828	
CHBJPT	2	194.6	[454]	0.056	0.000		0.051	0.017		0.153	0.600	0.887	
CHBJPT	2	213.6	[458]	0.375	0.655		0.115	0.153		0.138	0.655	0.297	
CHBJPT	2	216	[458]	0.687	0.112		0.720	0.111		0.285	0.049	0.285	
CHBJPT	2	241.6	[458]								0.552		

A6 Thesis Appendix 6: Top 1% hits based on selected statistics

For each of the statistics presented in the main paper (Kelly's Z_{nS}, $\frac{\text{Kelly's}\, Z_{nS}}{E[\text{Kelly's}\, Z_{nS}]}$, α, $\frac{\alpha}{E[\alpha]}$, $|L||R|$ and $\omega_{\text{max}}^{50,400}$), approximate p-values were determined for each 200kb window from the HapMap data by respampling the maximum (or minimum for $|L||R|$) test statistic in 10,000 200kb windows from the relevant chromosome and population. Windows with approximate p-values ¡0.01 were considered selection candidates. For statistics that control for expected LD, the HapMap genetic map was used [390].

To help compare the statistics, approximate p-values for each selection candidate window were then determined for each of these six statistics, and also for $\frac{\text{Kelly's}\, Z_{nS}}{E[\text{Kelly's}\, Z_{nS}]}$ and $\frac{\alpha}{E[\alpha]}$ with the expected LD estimated using the deCODE genetic map. Genes for which the transcript overlapped a window were retrieved from the UCSC Table Browser (track: RefSeq). Full results are shown in Tables A10, A11 and A12.

That this approach is primarily intended to aid comparison between the different LD-based statistics and highlight regions showing persistent LD distortions in multiple statistics, rather than to narrowly identify novel selection candidates.

	Kelly's Z_{nS}	HapMap $\frac{\text{Kelly's } Z_{nS}}{E[\text{Kelly's } Z_{nS}]}$	deCODE $\frac{\text{Kelly's } Z_{nS}}{E[\text{Kelly's } Z_{nS}]}$	α	HapMap $\frac{\alpha}{E[\alpha]}$	deCODE $\frac{\alpha}{E[\alpha]}$	$\|L\|\|R\|$	$\omega_{\max}^{50,400}$	Replication	Genes (Protein coding)
5000000	0.842	0.831	0.650	0.807	0.747	0.651	0.907	0.001		
15600000	0.213	0.005	0.006	0.098	0.009	0.009	0.880	0.015		NBAS,DDX1
15800000	0.231	0.004	0.007	0.265	0.013	0.018	0.707	0.126		MYCN
16000000	0.979	0.379	0.283	0.964	0.428	0.377	0.730	0.006		MYCN
26600000	0.320	0.032	0.023	0.108	0.009	0.014	0.217	0.058		OTOF,C2orf70,CIB4,KCNK3
26800000	0.232	0.017	0.018	0.094	0.007	0.013	0.168	0.253		KCNK3,SLC35F6,CENPA,DPYSL5
73200000	0.059	0.010	0.052	0.041	0.030	0.057	0.192	0.071		NOTO,SMYD5,PRADC1,CCT7,FBXO41,EGR4
73400000	0.026	0.001	0.049	0.039	0.028	0.057	0.217	0.167		ALMS1
73600000	0.050	0.003	0.043	0.059	0.017	0.061	0.305	0.923		ALMS1,NAT8,NAT8B
73800000	0.057	0.005	0.044	0.065	0.017	0.062	0.294	0.194		TPRKB,DUSP11,C2orf78,STAMBP,ACTG2
74200000	0.152	0.011	0.030	0.057	0.006	0.018	0.096	0.154		BOLA3,MOB1A,MTHFD2,SLC4A5
74400000	0.134	0.010	0.032	0.060	0.006	0.021	0.103	0.844	[456]	SLC4A5,DCTN1,C2orf81,WDR54,RTKN, INO80B,WBP1,MOGS,MRPL53,CCDC142, TTC31,LBX2,PCGF1,TLX2,DQX1
74600000	0.274	0.029	0.051	0.058	0.004	0.015	0.168	0.662	[358, 455]	DQX1,AUP1,HTRA2,LOXL3,DOK1,M1AP, SEMA4F
74800000	0.364	0.067	0.057	0.065	0.006	0.018	0.259	0.154		HK2
84200000	0.139	0.046	0.086	0.089	0.039	0.122	0.611	0.003		
84400000	0.085	0.009	0.074	0.074	0.020	0.100	0.544	0.555	[6]	SUCLG1,DNAH6
84600000	0.072	0.007	0.064	0.075	0.033	0.065	0.376	0.916	[358]	DNAH6
84800000	0.081	0.008	0.099	0.109	0.018	0.118	0.365	0.044	[455]	DNAH6,TRABD2A,TMSB10
87200000	0.219	0.051	0.001	0.321	0.127	0.010	0.008	0.921		
87400000	0.417	0.219	0.024	0.395	0.459	0.024	0.007	0.601		
87600000	0.239	0.219	0.014	0.286	0.429	0.037	0.004	0.023	[458]	
89200000	0.032	0.174	0.620	0.042	0.071	0.402	0.009	0.161		
89800000	0.036	0.718	0.591	0.001	0.064	0.040	0.002	0.883		
90800000	0.034	0.102	0.800	0.001	0.014	0.040	0.001	0.998		
91000000	0.038	0.086	0.832	0.029	0.081	0.406	0.003	0.285		
95000000	0.004	0.052	0.022	0.010	0.135	0.044	0.006	0.987		MAL,MRPS5,ZNF514,ZNF2
95200000	0.001	0.044	0.033	0.014	0.160	0.050	0.024	0.893		ZNF2,PROM2,KCNIP3
95600000	0.008	0.239	0.013	0.023	0.425	0.049	0.017	0.944		TRIM43
95800000	0.008	0.239	0.013	0.024	0.436	0.048	0.018	0.989		ANKRD36C
103200000	0.052	0.023	0.191	0.062	0.056	0.139	0.150	0.009		
110200000	0.006	0.112	0.416	0.008	0.192	0.182	0.003	0.757		MALL,NPHP1
111000000	0.006	0.112	0.055	0.000	0.000	0.001	0.003	0.019		RGPD6,RGPD5,BUB1
111200000	0.147	0.479	0.091	0.012	0.002	0.001	0.031	0.246		ACOXL
121200000	0.905	0.093	0.123	0.930	0.093	0.147	0.470	0.007		GLI2
130600000	0.010	0.005	0.001	0.025	0.012	0.006	0.011	0.851		POTEF,CCDC74B,SMPD4,MZT2B,TUBA3E
130800000	0.010	0.007	0.000	0.017	0.003	0.002	0.010	0.125		CCDC115,IMP4,PTPN18,POTEI,CFC1B
131000000								0.008		CFC1B,CFC1,POTEJ
131200000	0.327	0.221	0.097	0.068	0.010	0.004	0.015	0.008		GPR148,AMER3,ARHGEF4
131400000	0.285	0.066	0.079	0.118	0.008	0.053	0.042	0.037		ARHGEF4,FAM168B,PLEKHB2
135800000	0.007	0.138	0.031	0.015	0.156	0.128	0.068	0.996		ZRANB3
136000000	0.015	0.089	0.099	0.010	0.134	0.083	0.070	0.961	[358]	ZRANB3,R3HDM1
182200000	0.496	0.471	0.393	0.486	0.346	0.268	0.336	0.002	[6]	CERKL,NEUROD1
193400000	0.012	0.085	0.002	0.002	0.023	0.005	0.064	0.733		
193800000	0.019	0.124	0.021	0.007	0.049	0.006	0.102	0.545		
194200000	0.005	0.034	0.015	0.004	0.005	0.016	0.052	0.891		
194400000	0.002	0.021	0.053	0.005	0.006	0.019	0.039	0.278	[6]	
194600000	0.037	0.000	0.077	0.024	0.001	0.025	0.048	0.250		
194800000	0.138	0.002	0.093	0.029	0.002	0.028	0.040	0.647		
198000000	0.009	0.036	0.042	0.018	0.100	0.058	0.080	0.994		SF3B1,COQ10B,HSPD1,HSPE1-MOB4,HSPE1,MOB4,RFTN2
198200000	0.000	0.011	0.038	0.006	0.070	0.075	0.066	0.997		RFTN2,MARS2,BOLL,PLCL1
208200000	0.632	0.676	0.336	0.595	0.711	0.435	0.288	0.005		CCNYL1,FZD5,PLEKHM3
239400000	0.917	0.598		0.922	0.642		0.570	0.003		TWIST2

Table A9: Selection candidates identified by representative selection statistics in this study for HapMap Phase II CEU Chromosome 2 data

Position	Kelly's Z_{nS}	HapMap $\frac{\text{Kelly's } Z_{nS}}{E[\text{Kelly's } Z_{nS}]}$	deCODE $\frac{\text{Kelly's } Z_{nS}}{E[\text{Kelly's } Z_{nS}]}$	α	HapMap $\frac{\alpha}{E[\alpha]}$	deCODE $\frac{\alpha}{E[\alpha]}$	$\lvert L\rvert\lvert R\rvert$	$\omega_{\max}^{50,400}$	Supporting Evidence	Genes (Protein coding)
19200000	0.826	0.009		0.823	0.144		0.043	0.542		POTEB3,POTEB2,POTEB
19400000	0.833	0.007		0.535	0.296		0.045	0.780		
21200000	0.630	0.246		0.092	0.006		0.024	0.153		GOLGA6L2,MKRN3
26200000	0.206	0.132	0.027	0.001	0.001	0.003	0.004	0.050		HERC2
26600000	0.035	0.035	0.007	0.001	0.001	0.003	0.001	0.056		GOLGA8M
26800000	0.037	0.034	0.008	0.186	0.161	0.523	0.024	0.008		
28600000	0.076	0.839	0.056	0.010	0.021	0.005	0.006	0.015		GOLGA8H,ARHGAP11B
41400000	0.005	0.233	0.011	0.027	0.362	0.022	0.100	0.979		LCMT2,ADAL,ZSCAN29, TUBGCP4, TP53BP1,MAP1A
41600000	0.010	0.112	0.019	0.033	0.227	0.044	0.126	0.991		MAP1A,PPIP5K1,CKMT1B, STRC, CATSPER2,CKMT1A
41800000	0.020	0.040	0.285	0.008	0.064	0.049	0.138	0.840		PDIA3,ELL3,SERF2,SERINC4,HYPK MFAP1,WDR76,FRMD5
42000000	0.029	0.047	0.402	0.006	0.045	0.037	0.115	0.823	[457, 6]	FRMD5
46000000	0.416	0.065	0.328	0.266	0.011	0.033	0.377	0.002	[458]	SLC24A5,MYEF2,CTXN2,SLC12A1 DUT,FBN1
46200000	0.087	0.004	0.004	0.146	0.009	0.030	0.250	0.001	[458, 456]	
46400000	0.084	0.002	0.001	0.113	0.003	0.011	0.238	0.663		
62400000	0.008	0.026	0.066	0.031	0.123	0.091	0.087	0.964		CSNK1G1,KIAA0101,TRIP4,ZNF609
74600000	0.002	0.122	0.111	0.035	0.489	0.197	0.168	0.995		SCAPER
81000000	0.152	0.511	0.755	0.004	0.418	0.125	0.008	0.957		RPS17,CPEB1,AP3B2
88200000	0.599	0.424	0.198	0.608	0.409	0.227	0.447	0.004	[458]	AP3S2,C15orf38-AP3S2,ARPIN,ZNF710

TABLE A10: Selection candidates identified by representative selection statistics in this study for HapMap Phase II CEU Chromosome 15 data

	Kelly's Z_{nS}	HapMap $\frac{\text{Kelly's } Z_{nS}}{E[\text{Kelly's } Z_{nS}]}$	α	HapMap $\frac{\alpha}{E[\alpha]}$	$\lvert L\rvert\lvert R\rvert$	$\omega_{\max}^{50,400}$	Supporting Evidence	Genes (Protein coding)
22200000	0.306	0.443	0.107	0.116	0.083	0.006		
26000000	0.082	0.040	0.045	0.066	0.093	0.000		KIF3C,RAB10
27600000	0.009	0.099	0.004	0.034	0.057	0.602		GCKR,C2orf16,ZNF512,CCDC121, GPN1,SUPT7L,SLC4A1AP
27800000	0.012	0.059	0.008	0.063	0.068	0.817		MRPL33,RBKS,BRE
31800000	0.006	0.075	0.021	0.163	0.058	0.989		MEMO1
32000000	0.002	0.074	0.010	0.120	0.047	0.940		MEMO1,DPY30,SPAST
72000000	0.499	0.063	0.152	0.002	0.200	0.256		
72200000	0.221	0.013	0.034	0.001	0.013	0.153		CYP26B1,EXOC6B
72400000	0.055	0.005	0.006	0.032	0.012	0.001		EXOC6B
72600000						0.002	[458, 456]	EXOC6B
72800000	0.029	0.022	0.007	0.040	0.010	0.004		EXOC6B,SPR,EMX1
74000000	0.323	0.038	0.336	0.053	0.168	0.003		ACTG2,DGUOK,TET3
87400000	0.344	0.237	0.507	0.374	0.008	0.273		
87600000	0.286	0.125	0.293	0.183	0.006	0.078		
89800000	0.018	0.543	0.018	0.574	0.001	0.993		
90800000	0.007	0.052	0.001	0.007	0.000	0.993		
91000000	0.012	0.057	0.035	0.142	0.005	0.105		
95000000	0.040	0.086	0.024	0.326	0.004	0.886		MAL,MRPS5,ZNF514,ZNF2
97200000	0.008	0.026	0.033	0.145	0.020	0.981		ANKRD36
97400000	0.010	0.039	0.039	0.324	0.023	0.996		ANKRD36B
103200000	0.029	0.009	0.120	0.055	0.060	0.134		
108400000	0.080	0.009	0.100	0.057	0.233	0.981		GCC2,LIMS1
108600000	0.085	0.006	0.108	0.066	0.244	0.987	[454]	LIMS1,RANBP2,CCDC138
110200000	0.016	0.712	0.019	0.443	0.002	0.764		MALL,NPHP1
111000000	0.016	0.712	0.001	0.015	0.002	0.086		RGPD6,RGPD5,BUB1
121400000	0.575	0.101	0.596	0.092	0.328	0.008		GLI2
126800000	0.237	0.041	0.086	0.004	0.276	0.464		
130800000	0.005	0.035	0.041	0.021	0.002	0.017		CCDC115,IMP4,PTPN18,POTEI, CFC1B
131200000	0.335	0.310	0.083	0.000	0.003	0.050		GPR148,AMER3,ARHGEF4
131400000	0.220	0.013	0.146	0.005	0.035	0.149		ARHGEF4,FAM168B,PLEKHB2
131600000	0.162	0.005	0.156	0.003	0.051	0.947		PLEKHB2,POTEE
131800000	0.093	0.006	0.073	0.009	0.042	0.610		WTH3DI,TUBA3D,MZT2A
177000000	0.418	0.007	0.366	0.010	0.449	0.228		
177200000	0.404	0.003	0.381	0.005	0.482	0.010		
177400000	0.247	0.001	0.245	0.012	0.355	0.009	[456, 361, 454]	
177600000	0.148	0.003	0.145	0.015	0.286	0.046		HNRNPA3
193400000	0.000	0.015	0.002	0.024	0.073	0.952		
193600000	0.009	0.053	0.008	0.068	0.140	0.594		
193800000	0.017	0.144	0.003	0.035	0.116	0.578		
194400000	0.034	0.032	0.009	0.020	0.153	0.502		
194600000	0.057	0.000	0.050	0.020	0.149	0.601	[454]	
194800000	0.101	0.002	0.053	0.002	0.131	0.008		
198200000	0.001	0.189	0.015	0.230	0.076	0.981		RFTN2,MARS2,BOLL,PLCL1
198400000	0.010	0.124	0.058	0.265	0.094	0.994		PLCL1
203600000	0.003	0.466	0.045	0.800	0.050	0.999		NBEAL1
215600000	0.404	0.057	0.458	0.085	0.282	0.005		ABCA12
242200000						0.009		THAP4,ATG4B,DTYMK,ING5, D2HGDH,GAL3ST2,NEU4

TABLE A11: Selection candidates identified by representative selection statistics in this study for HapMap Phase II CHB+JPT Chromosome 2 data

A7 Thesis Appendix 7: Glossary of population genetics terms

Allele - A variant at a genetic locus. For example, most SNPs have two alleles, while a 'single allele' would imply an invariant site.

Ascertainment bias - In the context of SNP genotyping, ascertainment bias describes 'systematic deviations from an expected theoretical result attributable to the sampling processes used to find (ascertain) SNPs and measure (estimate) their population-specific allele frequencies' [503]. When obtaining genetic data from a sample, it is common to sequence specific markers (e.g. SNPs) that are known to be frequently polymorphic rather than the entire genome. These markers are identified using detailed sequencing of a small sample of individuals (the 'ascertainment set'). This protocol - finding polymorphic markers in a small sample then genotyping these in a large sample, perhaps from a different population - can introduce several biases to genetic datasets.

Autosomal - The chromosomes that are not sex chromosomes (e.g. in humans, not the X or Y chromosomes).

Balancing selection - Balancing selection occurs when natural selection maintains the presence of multiple alleles. The frequencies of these alleles can be stable over different time scales and be maintained due to several different processes, heterozygote advantage being a simple example [386].

Diploid - Having two copies of each chromosome. One copy is inherited from the father and one from the mother. The somatic (body) cells of a diploid organism are diploid, while the gametes (e.g. sperm and egg cells) are haploid.

Directional selection - Directional selection is selection for a phenotypic trait that is higher or lower than the current average trait value, thus changing the frequency of alleles that have an impact on that trait.

Fitness - The expected reproductive success of an individual with a given genotype or phenotype. When measured in relative terms, fitness is the relative expected number of offspring produced by an individual of one type expressed in terms of the expected number produced by an individual of another type, usually known as the selection coefficient. In absolute terms, fitness describes the expected number of offspring of an individual.

Frequency dependent selection - Frequency dependent selection describes a phenomenon whereby the fitness of a phenotype depends on the frequency of that phenotype in the population.

Genealogy - In coalescent theory, the bifurcating tree or ancestral recombination graph describing the history a set of homologous sequences, indicating how closely related different sequences are and, potentially, their divergence times. Note that the genealogy is not affected by mutation, such that it is possible for two sequences of finite length and identical mutation rates to be distantly related in their genealogy but have few segregating sites due to the random nature of mutation events. A phylogenetic tree is not the genealogy, but may attempt to estimate the genealogy by assuming that genetic divergence can be used as a proxy for branch length in the genealogy.

Genetic map - A map of estimated linkage between loci on a chromosome.

Genetic drift - The random fluctuation in allele frequencies due to random sampling of a finite population. Under many reproductive strategies, only a subset of individuals are expected to reproduce in a generation, even under neutrality. In sexual organisms, the version (paternal or maternal) of a chromosome passed to a child by a parent are also random. This leads to allele frequency fluctuations, which are greatest when the population size is small. Genetic drift reduces genetic variation within a population, but increases genetic distance between populations as populations 'drift' in random, usually different, ways.

Genetic hitchhiking - Genetic hitchhiking describes the change in the frequency of an allele due to linkage with a selected allele at a different locus.

Genetic variation - The extent to which individuals in a population show genetic differences in a genomic region. There are many measures of genetic variation - the number of polymorphic loci, heterozygosity and the number of alleles per locus are examples.

Genome-wide association study (GWAS) - A genome-wide association study attempts to identify genetic loci associated with a phenotype of interest in a sample, by searching for correlations between the occurrence of the phenotype and the allelic state at genetic markers interspersed along the genome.

Genome-wide selection scan (GWSS) - A genome-wide selection scan involves the repeated evaluation of a test statistic at different physical locations along the genome in order to identify locations that do not conform with neutral expectations, termed selection candidates.

Genotype - The heritable 'blueprint' of information contained internally in an organism and encoded genetically. The de-coding and expression of the genotype through the process of protein synthesis ultimately is a major factor influencing the phenotype of an organism. In the context of applied genetics, genotyping involves obtaining some information on an organism's genotype through one of many of laboratory techniques. Recently, 'genotype data' is often used to describe data on the state of the genome of an individual at many SNPs, in contrast to 'whole-genome data' which indicates data covering the vast majority of the genome, including loci that are not polymorphic in the population.

Haplotype - A set of alleles in an organism's genome that were inherited together from a single parent. In the context of signals of selection, haplotype homozygosity is of particular interest, describing the probability of two chromosomes in a sample sharing allelic state in a given genomic region.

Haploid - Having only a single copy of each chromosome. The gametes of a sexual diploid organism are haploid, with each single chromosome copy either selected from the chromosome that individual inherited from the mother or the father. In a haploid species, all somatic (body) cells are haploid also.

Heterozygote advantage - If an organism that is heterozygous at a given locus has greater average reproductive success than either homozyote then that variation at that locus displays heterozygote advantage. This will lead to balancing selection.

Linkage - Loci on the same chromosome that show high linkage have alleles that tend to be inherited together. This corresponds to a low probability of recombination between the loci. When the probability of recombination is greater than 50%, loci are considered unlinked.

Linkage disequilibrium (LD) - The non-random association of alleles at different loci, such that knowledge of the allelic state at one locus gives information on the allelic state at another. Usually, the two-locus case is considered, and is properly termed pairwise linkage disequilibrium. See the main text for details.

Locus - A specific location on the genome, such as a particular gene or base pair position.

Minor allele frequency (MAF) - The frequency of the rarest allele at a locus in a sample.

Monomorphic - A locus is monomorphic if it only has a single allele in a given population. Loci that are not monomorphic are polymorphic.

Neutral theory - The theory that the majority of molecular polymorphism is approximately selectively neutral, such that most allele frequency changes are not due to genetic hitch-hiking with selected alleles. The theory was suggested by Kimura in the 1960s [392], with Ohta proposing a modification such that many novel variants are mildly deleterious [484].

Panmictic - A population is panmictic if all individuals choose their mate at random. This is an unrealistic assumption that is frequently used to simplify mathematical models such as the diploid Wright-Fisher model.

Phasing - A statistical procedure performed on non-haploid genetic data to predict which variants belong to the same copy of a chromosome (i.e. are on the same haplotype). For example, genotype data is usually reported as the observed states at each locus. If the states of a diploid organism at two loci 'a/A' and 'b/B', then it is unclear whether the two haplotypes are 'a + b' and 'A + B' or 'a + B' and 'A + b'. A phasing procedure tries to estimate this, based on the genetic state of family members or the inferred frequency of different haplotypes in population genetic data.

Phenotype - The combined observable properties of a living organism, including its morphology, development, behaviour, metabolism and life cycle - in fact, anything related to its structure, function of behaviour. The phenotype is the product of the interaction between genotype, environment and epigenetic factors.

Polymorphic - In genetics, a polymorphic locus is a locus with multiple alleles in a given population. In practice, only a sample of the population will be available, such that in applied genetics polymorphism is usually determined with reference to the sample.

Population genetic data - A set of samples from a population for which some genetic information is available. An slightly extended description is given in section §4.4.1.

Purifying selection - Selection against alleles that are deleterious, such that they fall in frequency. Purifying selection can be especially intense within functionally important genes, leading novel variants to be rapidly purged. The prevents the accumulation of deleterious variants in the population, which is known as 'genetic load'. The reduction in diversity associated with purifying selection, which is considerably intensified by linkage between negatively selected and neutral alleles, is known as background selection. Purifying selection distorts the site frequency spectrum, leads to reduced diversity, and can create long haplotypes.

Recombination rate - The recombination rate is the probability of recombination at a specific location on a chromosome. The cumulative probability of recombination between two loci is known as their genetic map distance. A description of the genetic map distance between many loci on a chromosome is a genetic map.

Selection candidates - A selection candidate is a region of the genome that has been proposed as evolving in a non-neutral manner in a specific population. GWSSs tend to report lists of selection candidates.

Selective sweep - A pattern whereby a specific allele rises to high frequency due to strong positive directional selection, leading to characteristic distortions in the frequencies of alleles at linked loci and a consequent reduction in local genetic diversity.

Single nucleotide polymorphism (SNP) - A polymorphism whereby more than one nucleotide is observed at a specific locus (and strand) in a population. SNPs have proven particularly popular genetic markers for conducting GWSSs, and many selection statistics are designed with SNPs in mind.

Site frequency spectrum (SFS) - The site frequency spectrum, or allele frequency spectrum, is a vector describing the number of polymorphic sites for which the derived allele is at each possible frequency in a sample. For example, if there are five polymorphic sites for which the derived allele is only observed once, and three for which the derived allele is observed twice, then the first two terms of the SFS will be $[5, 3, \ldots]$, with the length of the vector being $n - 1$. When the derived and ancestral states at a locus are not known, then the minor allele frequency at polymorphic loci is used, yielding a folded site frequency spectrum.

Standing genetic variation - Allelic variation that currently exists in a populations. Selection on standing genetic variation describes a situation whereby a polymorphic allele becomes subject to selection after evolving neutrally for some period of time, and can be contrasted with selection on novel variants.