

Inferring Recent Demography from Spatial Genetic Structure

by

Harald Ringbauer

*A thesis presented to the Graduate School of the Institute of Science and Technology Austria
in partial fulfillment of the requirements for the degree of Doctor of Philosophy.*

Klosterneuburg, Austria.

January, 2018.



Institute of Science and Technology

The thesis of Harald Ringbauer,

titled

Inferring Recent Demography from Spatial Genetic Structure,

is approved by:

Supervisor: Nicholas Barton, IST Austria, Klosterneuburg, Austria

Signature: _____

Committee Member: Christoph Lampert, IST Austria, Klosterneuburg, Austria

Signature: _____

Committee Member: John Pannell, University of Lausanne, Switzerland

Signature: _____

Defense Chair: Vladimir Kolmogorov, IST Austria, Klosterneuburg, Austria

Signature: _____

© by Harald Ringbauer, January, 2018
All Rights Reserved

I hereby declare that this thesis is my own work and that it does not contain other people's work without this being so stated; this thesis does not contain my previous work without this being stated, and the bibliography contains all the literature that I used in writing the dissertation.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee, and that this thesis has not been submitted for a higher degree to any other university or institution.

I certify that any republication of materials presented in this thesis has been approved by the relevant publishers and co-authors.

Signature: _____

Harald Ringbauer
January, 2018

Abstract

This thesis is concerned with the inference of current population structure based on geo-referenced genetic data. The underlying idea is that population structure affects its spatial genetic structure. Therefore, genotype information can be utilized to estimate important demographic parameters such as migration rates. These indirect estimates of population structure have become very attractive, as genotype data is now widely available. However, there also has been much concern about these approaches. Importantly, genetic structure can be influenced by many complex patterns, which often cannot be disentangled. Moreover, many methods merely fit heuristic patterns of genetic structure, and do not build upon population genetics theory. Here, I describe two novel inference methods that address these shortcomings.

In Chapter 2, I introduce an inference scheme based on a new type of signal, identity by descent (IBD) blocks. Recently, it has become feasible to detect such long blocks of genome shared between pairs of samples. These blocks are direct traces of recent coalescence events. As such, they contain ample signal for inferring recent demography. I examine sharing of IBD blocks in two-dimensional populations with local migration. Using a diffusion approximation, I derive formulas for an isolation by distance pattern of long IBD blocks and show that sharing of long IBD blocks approaches rapid exponential decay for growing sample distance. I describe an inference scheme based on these results. It can robustly estimate the dispersal rate and population density, which is demonstrated on simulated data. I also show an application to estimate mean migration and the rate of recent population growth within Eastern Europe.

Chapter 3 is about a novel method to estimate barriers to gene flow in a two dimensional population. This inference scheme utilizes geographically localized allele frequency fluctuations - a classical isolation by distance signal. The strength of these local fluctuations increases on average next to a barrier, and there is less correlation across it. I again use a framework of diffusion of ancestral lineages to model this effect, and provide an efficient numerical implementation to fit the results to geo-referenced biallelic SNP data. This inference scheme is able to robustly estimate strong barriers to gene flow, as tests on simulated data confirm.

Acknowledgments

Finishing a PhD has been the hardest, but also the most productive undertaking of my life so far. There are many people which made it possible that I completed this rough journey. It would have been impossible to make it alone, and I am really grateful for all the support I have received.

Most of all I am indebted to my parents for their continuous support throughout my studies, and also to the rest of my family: My grandparents, my fun sister and furry Kitty, as well as my cousins and their significant others - and by now also Lorenz, the very definition of a best friend. I was also lucky to have an amazing group of friends - thanks to Dominik, Christoph, David, Johannes and Mato for all the fun game and cinema nights. I also want to thank Julia, who supported me when I needed it the most, and was the best thing that happened to me in the last years.

I also want to thank my support group at IST. It is impossible to list all the great connections I made. But I try to pick out a few shining examples. Anton, with whom I have set out on so many adventures and ski trips. To give just one example: Right now, I am on the way to meet him in Istanbul and heading for a powder adventure in Hokkaido. Alex K., my great table soccer sparring partner who pushed my game to the next level. He made many of my days much more fun (fist pump man). Michal, my man for the enjoyable nights out at ...classic Viennese places and deep after-football conversations. And last but not least, Jamie and Alex Z., for countless fun pub quiz, sports bar and table soccer nights.

I have also been lucky to be part of an amazing research group. Maria, who grew a close friend, and David, whom I had so many great discussions about Science Fiction with, deserve a special mentioning. So do also Himani, Barbora, Gemma and Katka, for all the interesting lunch conversations, and for giving me a lot of perspective which hopefully helped me grow. I also want to thank the Coop lab. Visiting them for two months in California has been a truly fun experience, and it was also a very productive time. And most importantly, I want to express my gratitude towards my supervisor, Nick, for the freedom to explore my own interests, and for giving me a lot of flexibility throughout my PhD.

List of Publications

This thesis is partly based on the following two publications:

- Ringbauer, H., Coop, G., and Barton, N. H. (2017). Inferring recent demography from isolation by distance of long shared sequence blocks. *Genetics*, 205(3):1335–1351.

Contributions: HR wrote the paper, developed the method, implemented the simulations and inference scheme, analyzed the data; GC contributed ideas and gave feedback, NHB contributed ideas and gave feedback.

Chapter 2 is based on this paper.

- Ringbauer, H, Kolesnikov, A., Field, D., and Barton, N.H (2018). Estimating barriers to gene flow from distorted isolation by distance patterns. *Genetics*, genetics-300638.

Contributions: HR wrote the paper, developed the method, implemented the simulations and inference scheme, analyzed the data; AK helped with implementation and computational speedup, DF helped with data processing and gave feedback, NHB contributed ideas and gave feedback.

Chapter 3 is based on this paper.

Contents

Abstract	i
Acknowledgments	iii
List of Publications	v
List of Symbols	xi
1. Introduction	1
1.1 Demographic Inference from Genetic Data	1
1.2 Inference of Population Structure - a Brief Review	2
1.3 Contributions of this Thesis	10
2. Inferring recent demography from isolation by distance of long shared sequence blocks	13
2.1 Materials and Methods	17
2.1.1 The Model	17
2.1.2 IBD Sharing in the Model	19
2.1.3 Fitting the Model to Data	22
2.1.4 Simulations	25
2.1.5 Application to Eastern European Data	27
2.2 Results	29
2.2.1 Block-sharing in Simulated Data	29
2.2.2 Inference in Simulated Data	31
2.2.3 Sampling Guidelines	35
2.2.4 Analysis of POPRES Data	36
2.3 Discussion	38
2.3.1 Implications for Demographic Inference	40
2.3.2 Analysis of Human Data	40
2.3.3 Outlook	41
2.4 Appendix	42

2.4.1	Appendix A: Effective Density	42
2.4.2	Appendix B: Chromosomal Edge Effects	43
2.4.3	Appendix C: Likelihood	43
2.5	Supplementary Material	45
2.5.1	Supplementary Information 1: POPRES analysis	45
2.5.2	Supplementary Information 2: Simulation Details	47
2.5.3	Supplementary Information 3: Simulations	50
3.	Estimating barriers to gene flow from distorted isolation by distance patterns	59
3.1	Materials and Methods	61
3.1.1	Model	61
3.1.2	Fitting the model to data	67
3.1.3	Implementation	70
3.1.4	Simulations	70
3.1.5	<i>Antirrhinum majus</i> Data	72
3.1.6	Data Availability	72
3.2	Results	72
3.2.1	Inference on Simulated Data	72
3.2.2	Hybrid Zone Analysis	76
3.3	Discussion	80
3.4	Outlook	81
3.5	Appendix	83
3.6	Supplementary Material	84
3.6.1	Supplementary Information 1: Fitting Allele Frequencies	84
3.6.2	Supplementary Information 2: Comparison to Geneland	95
3.6.3	Supplementary Information 3: Data Cleaning for Hybrid Zone	100
4.	Future Directions	109
4.1	Generalized Parametric Inference for IBD Blocks	109
4.1.1	Parametric Model for Spread of Ancestry	110
4.1.2	Example: Heterogeneous Gene Flow and Population Density	111
4.1.3	Further Directions	114
4.2	Software Package for IBD Block Analysis	115
4.3	Outlook	116
	Bibliography	119

List of Tables

2.1	Maximum Likelihood Estimates for Eastern European IBD data.	37
2.2	Location and number of samples used for inference	46

List of Figures

2.1	Example of an IBD block coinherited from a common ancestor	15
2.2	Diffusion model visualized in one spatial dimension	19
2.3	Simulated IBD block sharing compared with theoretical expectations	23
2.4	Various population density scenarios	30
2.5	Maximum likelihood estimates	32
2.6	Likelihood estimates for various population density scenarios	33
2.7	Observable IBD block sharing compared with theoretical predictions	34
2.8	Fit of models to Eastern European block sharing data	37
2.9	The timing of IBD blocks	38
3.1	Geographic fluctuations of allele frequencies	62
3.2	Comparison of analytical diffusion formulas with discrete simulations	65
3.3	Decay of identity by descent in presence of a strong barrier to gene flow	68
3.4	Model used to generate synthetic data sets	71
3.5	Parameter estimates on simulated data	73
3.6	Bootstrap estimates on simulated data	74
3.7	Parameter estimates in a simulation of secondary contact	75
3.8	Testing various putative barrier positions on synthetic data	76
3.9	Decay of pairwise homozygosity with geographic distance	78
3.10	Inference of putative barrier positions within a <i>Antirrhinum</i> population	79
3.11	Comparison of fitting methods on simulated data	89
3.12	Inference with varying numbers of individuals	90

3.13	Inference with varying number of loci	91
3.14	Fitting only the barrier parameter	93
3.15	Fitting with various binning	94
3.16	Geneland on datasets with no isolation by distance	96
3.18	Geneland on datasets with isolation by distance	98
3.20	Position of SNPs on <i>Antirrhinum</i> linkage map	102
3.21	Distribution of mean allele frequencies	104
3.22	Heterogeneity of isolation by distance estimates	105
3.23	IBD of data simulated with hybrid zone parameters	106
3.24	Power of inference scheme on simulated data set	107
4.1	Spatial heterogeneous model	112
4.2	Spread of ancestry in heterogeneous model	112
4.3	Inference on data simulated under spatial heterogeneous model	113

List of Symbols

β	Population growth rate: $D_e(t) = Dt^{-\beta}$
γ	Scaled barrier strength ($= \frac{2\kappa}{\sigma^2}$)
κ	Barrier strength parameter
μ	Rate of long-distance mutation/migration
σ^2	Dispersal rate (rate of geographic diffusion of ancestral lineages)
Nbh	Neighborhood Size ($= 4\pi D_e \sigma^2$)
D_e	Effective population density
F	Chance of identity by descent (coalescence before mutation/migration)
F_{ST}	Wright's fixation index
G	Chromosome map length (in centimorgan)
h	Pairwise homozygosity
K_γ	Modified Bessel function of the second kind of degree γ
L	Map length of genome block (in centimorgan)
m'	Fraction of migrants each population receives
N_e	Effective population size of local populations
p	Allele frequency
r	Euclidean geographic distance
t	Time back since present (in generations)



1. Introduction

1.1 Demographic Inference from Genetic Data

THE main theme of this thesis is the estimation of population structure from genetic data. The underlying idea is simple: The demographic structure of a population affects its genetic structure. In turn, genetic variation contains information about population structure. Therefore, sampling and genotyping individuals can be utilized to learn about demography of a population.

Typical questions of this discipline are about important demographic parameters:

- **Dispersal:** What is the distribution of distances covered by an individual between birth and reproduction? In particular, what is the variance of this distribution?
- **Barriers to gene flow:** Are there barriers to dispersal, for instance caused by physical barriers or genetic incompatibilities? What is the strength of such barriers?
- **Population size and density:** What is the total size of the population? What is the population density per area?
- **Population dynamics:** Is the population currently growing or declining? Was population structure different in the past?

Answers to such questions are of great interest in conservation and population management, and can also be of much practical utility, for instance for pest control. Moreover, knowledge about demographic structure is important for understanding how various evolutionary forces shape genetic variation - one of the key questions in population genetics and modern genomics. A sound null model for demographic structure is also essential for interpreting the genetic traces of selection and adaptation in the wild, as the lack of such knowledge can lead to seriously misleading conclusions (Meirmans, 2012).

In principle, demographic inference can be based on direct observations, such as radio-tagging or capture-mark-release experiments. However, such observations are often very time- and work-intensive. They are error-prone, since rare but important events, such as long-distance migrations, can be easily missed. Moreover, these methods only provide a snapshot in time. Therefore, the rapidly increasing capacities to genotype genetic markers and to record geographic data has aroused a great deal of interest in inferring demographic structure from genetic data. The idea to extract information about population structure from genetic variation has a long history. First attempts trace back to the founding fathers of the modern synthesis (Wright et al., 1942).

Genetic data can also be used to infer the deep history of a population and large-scale demographic events, for instance large scale colonization patterns. This discipline has been termed “phylogeography” (Knowles, 2009; Nielsen and Beaumont, 2009), and has delivered novel insight into important open questions, such as the origin of our own species, *Homo sapiens*. In contrast, this thesis focuses on the estimation of structure of populations on more recent timescales. The aim is not to draw inference about the history of the population, but to estimate its current structure. Inference is best done on geographical scales that are much smaller than the range of a species, as genetic structure begins to mirror current population structure much quicker and approximate equilibrium is reached more rapidly on geographically localized scales (Slatkin and Barton, 1989).

1.2 Inference of Population Structure - a Brief Review

Demographic inference based on genetic markers has a long history. Interest in this approach traces back all the way to the founders of the modern synthesis of population genetics. Already S. Wright and Th. Dobzhansky studied the population structure and dispersal patterns of the common fruit fly *Drosophila pseudoobscura* (Wright et al., 1942). Back then, they lacked the ability to genotype individuals, i.e. to directly measure genetic variation. But they ingeniously based inference on recessive lethal mutations, as it was possible to detect those by cumbersome counting offspring of crosses in the laboratory. They fit a basic model of population structure to the spatial distribution of such mutations, which enabled them to estimate migration rates and local effective population sizes.

The Genotyping Revolution

During the last decades, revolutionary advances in genotyping have occurred. New technologies dramatically changed the data available for demographic inference. This progress has spurred the development of completely novel inference methods. To provide the right context for the following sections, I first briefly review this genotyping revolution.

In the 1960s, it became possible to directly measure genetic variation at the molecular level. The first methods analyzed allozyme variation by means of studying electrophoretic variants. In the following decades, similar techniques such as RFLP (restriction fragment length polymorphism) analysis were developed. Typically, DNA fragments of various length are separated through a process known as agarose gel electrophoresis. For instance, one can ascertain whether or not a restriction enzyme cut at restriction sites, and therefore infer sequence variance at these sites. While such techniques are often slow and cumbersome, they were the first methods inexpensive enough to see widespread application. Analysis is typically limited to variation of a handful of specific markers. But these methods enabled eager researchers to yield a first glimpse into existing genetic variation (Hubby and Lewontin, 1966), and they provided a solid basis for further development.

In the following decades, more sophisticated methods to study genetic variation were developed. During the 1990s the analysis of microsatellite markers was popularized. These genetic markers are tracts of tandemly repeated (i.e. adjacent), several nucleotide long DNA motifs. The mutation rate of microsatellite markers can be markedly higher than the nucleotide mutation rate, and therefore microsatellite markers are often found to be highly variable in a population. Since PCR (polymerase chain reaction) became ubiquitous in laboratories in the early 1990s, researchers were able to design primers and amplify sets of microsatellites at low cost. They quickly became an important signal for demographic inference (Manel et al., 2003), and they have remained an important tool up until today.

In addition to microsatellite markers, it has become feasible to genotype single nucleotide polymorphisms (SNPs), which are single site variations in DNA. Several applications have been developed that interrogate SNPs, typically by hybridizing complementary DNA probes to the SNP site. SNPs have a much lower mutation rate (Scally (2016) estimate a mean rate of $\approx 0.5 \cdot 10^{-9}$ per base pair and generation in humans) than most microsatellites (Brinkmann et al. (1998) estimate a mutation rate of up to $\approx 1.0 \cdot 10^{-3}$ per generation for some human microsatellites). Therefore existing SNP variation typically traces back to much older mutations. Modern SNP arrays can in-

terrogate hundreds of thousands of SNPs simultaneously. This approach is often considerably cheaper than sequencing whole genomes once suitable arrays have been designed, while still providing similar amounts of information (Gabriel et al., 2002).

Nowadays we are rapidly approaching an era of population genomics. The advent of fluorescence sequencing technologies in the 2000s has made it possible to determine most of the DNA sequence of a sample at high speed and at reasonable cost. New sequencing technologies continue to bring down the cost of sequencing even further. For instance, modern high-throughput sequencing allows to sequence a human genome for \$10000, which is a dramatic decrease in cost since sequencing the first human genome for approximately \$100 million in 2001 (Wetterstrand, 2013). One can speculate that in the foreseeable future full-genome sequencing of many individuals will become widely affordable, even for non-model organisms.

Method Development

Quickly following these dramatic advances in genotyping, novel methods to infer population structure from observed genetic variation have been developed. While the underlying ideas often trace back a long time, the methods to fit the data and their underlying signal have been driven by the amount of available genetic data, and also by increasing computational power. Here I give a brief overview of such methods. I focus on signals that have been utilized for inference, and summarize approaches to fit the observed genetic data.

The first allozyme studies almost immediately triggered methods that fit demographic models to observed geographic genetic variation. A lot of necessary theory had been already developed, but had not been applied to empirical data yet. Right from the beginning, a common underlying theme of many methods has been that geographic structure of a population will cause spatial partitioning of genetic variation. The important idea is that studying this variation can be used to infer gene flow between subpopulations together with the rate of random drift within them.

Tracing back to the famous population geneticist Wright, the perhaps most popular method to do so is based on the fixation index F_{ST} , a measure of genetic differentiation, to learn about gene flow between different subpopulations. In Wright's original definition, F_{ST} measures the correlation between genotypes chosen randomly from within the same subpopulation relative to the entire population (Wright, 1949). Since then, a whole zoo of similar measures has been developed, many of them formally equivalent to F_{ST} . In particular, it has become increasingly popular to describe F_{ST} in terms of diversities within and between subpopulations (Nei, 1973), since these quantities can

be estimated from empirical data. A modern population genetics approach is to trace back ancestral lineages and describe coalescence times (i.e. the timing of most recent common ancestors). In this framework, defining F_{ST} in terms of coalescence times has proven to be very useful (Slatkin, 1991). Up until today, F_{ST} has remained one of the most frequently used statistics in population genetics.

The great value of F_{ST} for demographic inference derives from a simple formula (Wright, 1949) for the so called infinite island model. According to Wright's famous formula, if a large number of subpopulations are equally likely to exchange migrants and if mutation is rare, then at equilibrium:

$$F_{ST} \approx \frac{1}{4N_e m' + 1},$$

where m' is the fraction of migrants each population receives and N_e is the effective population size of local populations. This formula has been widely used to estimate the product $N_e m'$ from sample data. Typically, first some method is used to estimate F_{ST} from genetic data, which is then translated into an estimate for $N_e m'$ via Wright's formula. Several robust methods to estimate this summary statistic from genotype data have been developed. In particular, method-of-moments estimates have been widely used (Weir and Cockerham, 1984). Utilizing F_{ST} is a source of information that is relatively robust to specific model deviations, such as selection and variation in population model (Slatkin and Barton, 1989). However, as I will review below, recently there has been much concern about this method (Whitlock and McCauley, 1999).

There have been many refinements and extensions to methods for inference of population structure based on signals of genetic differentiation. One can directly fit $N_e m'$ via maximum likelihood approaches (Barton et al., 1983; Slatkin and Barton, 1989), based on the result that in the infinite island model the allele frequency distribution follows a beta distribution (Wright, 1938). Methods for microsatellite markers have been developed, taking their high mutation rate and their specific mutation pattern into account (Slatkin, 1995). Also, the spatial distribution of rare variants has been utilized as the signal for inference (Slatkin, 1985). In all cases, the underlying idea remains similar: Population structure can be inferred via its effect on spatial genetic variation across different subpopulations.

Another popular approach based on patterns of genetic differentiation is to utilize classical isolation by distance. This concept traces back to Wright (1943) as well. The underlying idea is that in populations with local migration, nearby individuals are genetically more similar on average than more distant ones, as they have a higher chance to be related. Based on classical models of Malécot (1948) and Kimura and Weiss (1964),

Rousset (1997) shows that the rate of decay of a measure of genetic differentiation such as F_{ST} with pairwise geographic distance is primarily informative about a model parameter termed neighborhood size N_{bh} , which is roughly the number of individuals in the area from which parents of an individual can be drawn. Therefore, by fitting isolation by distance models to observed genetic data, one can estimate this demographic parameter. This is typically done by regressing a measure of pairwise diversity against pairwise geographic distance, either for pairs of demes (Rousset, 1997) or for pairs of individuals (Rousset, 2000).

A different kind of signal used since relatively early on is the non-independence of different markers. Such independences can arise because some markers are mostly co-inherited together, in particular for markers that are genetically close (i.e. linked markers). In some scenarios, these shared patterns across loci contain information that can be utilized for demographic inference. For instance, linkage patterns for pairs of markers are informative about dispersal patterns in admixture zones (Barton and Gale, 1993), in which there is a balance between linkage disequilibrium built up by dispersal and recombination which breaks down these associations. Even in spatially unstructured populations, random drift can build up random associations between markers, and this signal can be informative about the recent size of a population (Hill, 1981; Waples and Do, 2008).

The advent of cost-efficient, large-scale genotyping techniques has induced a flood of genetic data. Simultaneously, rising computational power has enabled researchers to fit more complex models. Together, these two advances triggered the development of a next generation of inference schemes. In the early years of this millennium, methods that use Markov Chain Monte Carlo algorithms to fit parameters for models for population structure were introduced (Beerli and Felsenstein, 2001). Ever since then, they have been in wide use to estimate the joint posterior distribution of these parameters, without the need to explore the whole, often very complex, parameter space. The underlying ideas of inference have remained mostly the same, but these methods directly fit models to discrete genotype data, without the intermediate step of summary statistics. On the other hand, approximate Bayesian Computation methods that fit summary statistics of the data to estimate parameters of sometimes complex models of recent demography have been widely used (Beaumont et al., 2002).

The wealth of data, in particular the possibility to genotype many markers for a large number of individuals, has also enabled the development of completely novel approaches. One can learn about population structure by clustering samples into population units based on genetic similarity, or by directly inferring immigrants based on their genotype (Pritchard et al., 2000; Guillot et al., 2005). Nowadays genetic data can

be even utilized to directly identify close relatives based on their genotypes (Blouin, 2003). However, identifying relatives or immigrants is conceptually closely related to direct observations, and therefore demographic inference based on these signals suffer from similar shortcomings. They typically only yield a snapshot in time, and rare, but important, events could be missed.

Landscape Genetics

In parallel to these advances in classical population genetics, a whole new discipline that tries to infer complex demographic patterns has been founded (Manel et al., 2003). This field named landscape genetics is concerned with quantifying the effect of landscape composition and configuration on gene flow and spatial variation (Storfer et al., 2007). Typically, questions about connectivity and barriers to gene flow are addressed by correlating some measure of genetic differentiation - often F_{ST} - against environmental variables. Triggered by the increasing availability of genetic data, the ability of high resolution landscape data and the relative ease with which such methods can be applied, the number of landscape genetic publications has been rising rapidly recently (Storfer et al., 2010). Landscape genetic methods are usually not based on explicit population genetic models, but are rather statistical methods that often originate from ecology. These approaches can be broadly grouped into two approaches.

First, one can try to infer geographic genetic structure without first including information about landscape. For instance, one can group individuals into population clusters based on their genotypes (Pritchard et al., 2000; Guillot et al., 2005). This is typically done by simultaneously inferring the allele frequencies within clusters and population memberships via minimizing both linkage disequilibrium and Hardy-Weinberg disequilibrium (Falush et al., 2003). Another method is to directly infer areas of sharp genetic change (Womble, 1951; Manni et al., 2004; Cercueil et al., 2007). After applying these methods, one can ask whether the inferred delimitations or cluster boundaries agree with known landscape features. For example, such approaches have been used to detect barriers to gene flow (Safner et al., 2011).

Second, one can directly model the effect of known landscape features. For instance the Mantel and partial Mantel test (Smouse et al., 1986), statistical tests of the correlation between two or more distance matrices, have been widely used to test whether some measure of environmental distance has an effect on genetic differentiation (Storfer et al., 2010). More recently, resistance methods that draw upon electrical circuit and random walk theory have become popular (McRae et al., 2008). These approaches utilize a resistance model for connectivity in heterogeneous landscapes to calculate

pairwise coalescence times. Although they only approximate exact population genetics theory, these approaches can be used to calculate expected genetic differentiation such as pairwise F_{ST} (McRae, 2006). Using this intermediate step, circuit methods have been applied to fit the effects of complex migration landscapes. They are of approximate nature, but they offer an often necessary numerical speedup compared to exact coalescence methods (Petkova et al., 2015).

The Beginning of a New Era: Haplotypes

It has become feasible to genotype markers spaced densely throughout the genome. Nowadays methods can therefore make use of a wealth of linkage information and haplotype structure for demographic inference (Hellenthal et al., 2014; Goldberg et al., 2014; Sedghifar et al., 2015). One signal is particularly promising for demographic inference: In humans and some model systems, it has become possible to directly detect the traces of recent relatedness, so called identity by descent (IBD) blocks. Such blocks of shared genome are co-inherited from a recent co-ancestor, and therefore carry exceptionally few distinguishing mutations. They are the direct genetic signal of recent co-ancestry, and are not affected by older patterns. This advantage makes IBD blocks an ideal source for inferring the recent structure of a population. Moreover, the length of such blocks is informative about age of co-ancestry, and this helps to avoid several fundamental limitations of existing methods (Barton et al., 2013). This approach opens up completely novel avenues for the inference of recent demography (Novembre and Peter, 2016). First methods that make use of this signal have been already developed (Ralph and Coop, 2013; Palamara and Pe'er, 2013; Hellenthal et al., 2014; Ringbauer et al., 2017a). These inference schemes have already yielded novel insight into the fine-scale population structure of humans, and most likely this field has a promising future.

Common Caveats and Concerns

Demographic inference based on genetic data is an attractive substitute for direct observations. However, there are also severe pitfalls and caveats. Fitting simple models to observed genetic structure can be misleading. The real demographic history can be very complex, but most methods require that stringent model assumptions are met. For instance, methods that translate F_{ST} into estimates of gene flow assume that a set of panmictic demes symmetrically exchange migrants, with a low mutation rate and enough time to equilibrium, and without any further substructure. It seems plausible that one or more of these assumptions are frequently violated - and therefore the un-

derlying model has been termed the “fantasy island model” (Whitlock and McCauley, 1999).

Moreover, it is often difficult to validate inference methods against simulated data. The true model is usually unknown and the parameter space can be huge. Therefore it is a complex and computationally expensive task to test such a large space of models and study the effects of model misspecification. In addition, it is often challenging to interpret the outcome of inference schemes, and link it back to biologically relevant parameters. In particular, most methods do not use a population genetics model to fit genetic data, but rather use statistical models to empirically fit heuristic patterns. Therefore, most inference methods do not model the process that generates spatial genetic structure, but merely fit its effects. However, without knowledge about the underlying process, it is difficult to interpret the fitted parameters, and link them back to biologically relevant models.

The last problem seems to be most pronounced for inference in populations with continuous spatial structure. For such populations, estimates are often only based on heuristic statistical models. This disconnect is somewhat surprising, as spatial population genetics theory is relatively well developed (e.g. (Wright, 1943; Malécot, 1948; Nagylaki, 1978; Barton et al., 2002, 2013). The use of a geographic diffusion approximation for the spread of ancestry combined with modern coalescence theory (Kingman, 1982) has delivered much theoretical insight (Barton et al., 2002; Wilkins, 2004), and first ideas along these lines trace back to Wright (1943). While ideas for inference schemes for continuous populations have been laid out by Barton et al. (2013), very few existing methods draw upon explicit spatial population genetic theory (e.g. Rousset (1997)).

In particular, many landscape genetic methods suffer from these shortcomings. While genetic data is often limited and the true underlying, perhaps very complex, scenario is usually unknown, models with many parameters are fit to often very limited genetic data. Observed genetic patterns are usually assumed to be caused exclusively by present landscape structure. The genetic structure due to deeper patterns of phylogeography are often not accounted for.

For instance, many popular circuit resistance methods (McRae, 2006) assume that genetic differentiation is solely due to current migration patterns, but in practice they are often applied to geographical scales of whole species. On such large scales, deeper phylogenetic patterns can easily invalidate the model assumptions and lead to spurious inferences, especially when old genetic polymorphisms such as SNPs are used. Some other landscape genetic methods such as using the Mantel statistic do not test the effect of landscape against any population genetic null model at all. This problem makes their

use even more problematic, as they can be fundamentally flawed in the presence of spatial-autocorrelation (Guillot and Rousset, 2013).

Together, these caveats have caused much skepticism in the population genetics community about the general validity of most landscape genetic methods. Overall, mismatches between population structure inferred from genetic structure and direct observations have caused a widespread belief that demographic inference from genetic data is problematic in general, and that genetic data is often over-interpreted (Bossart and Prowell, 1998). Further studies that compare estimates based on genetic data to direct observations of population structure are urgently needed to settle this issue.

1.3 Contributions of this Thesis

The overarching theme of this thesis is to develop novel inference methods for spatially extended populations. My aim has been to overcome some of the above mentioned shortcomings. In particular, the goal has been to develop inference schemes that link demographic inference to population genetic theory, and are based on signals that are not confounded by deep, ancestral structure. In Chapter 2, I do so by basing inference on a promising novel signal: Identity by Descent blocks.

In all cases, the ultimate goal has been to develop a full-fledged inference scheme, that can be applied to genetic data by other researchers to address central questions about recent demography. To build such a scheme, I typically partitioned the work into four steps:

1. Develop theory

To infer parameters that can be directly interpreted in terms of population genetics, I directly fit population genetic models. Models with a low number of parameters can obviously only be an approximation to the complex, real demography of natural populations. Therefore, the goal was not to develop a mathematically exact model, which in many cases is formally problematic anyways, especially in two spatial dimensions (Felsenstein, 1975), but one that robustly fits observed patterns based on important summary parameters. I tried to make use of patterns in the data that are robust to many confounding factors, such as ancestral population structure.

Usually, there exists a plethora of population genetic theory to draw from, in particular for unlinked markers. In Chapter 2, I make use of a promising novel signal, identity by descent (IBD) blocks. This young area of population genetics

that deals with the sharing of such blocks is still relatively unexplored. Therefore, while drawing upon previous work, I also developed novel theory. In Chapter 3 I modify and expand existing theory to obtain a robust numerical framework for the analysis of barriers to gene flow.

2. Develop methods to fit theory to data

To do inference, one has to connect theory with observed data. One has to apply methods to fit the model to observed genetic data. There are usually many possibilities to do so. Different methods have different properties with respect to estimator bias and variation, and a priori it is often not clear which method performs best. Therefore, it is important to test and compare these methods.

3. Test methods on simulated data

In order to learn about the power and the limitations of an inference scheme, it is essential to test it on datasets where the actual population structure is known. A straightforward way to produce such data is to analyze synthetic data simulated with known ground truth. Throughout this work, I implemented my own simulation engines. In most cases, the goal was not to compete with existing methods, but to have simulations that can be easily tweaked to simulate different scenarios, and produce data that fit seamlessly into the inference pipeline.


4. Apply methods to real data

The ultimate test of an inference is to apply the developed inference methods to empirical datasets. Such an application showcases the ability of the inference scheme to fit genetic patterns. During this process, shortcomings of a method and hidden caveats can be uncovered. One is also able to also explore the fit of the best model, which can yield important insights into deviations from the fitted model.

In Chapter 2, I develop a method based on the spatial distribution of pairwise shared long blocks of genome, so called IBD (identity by descent) blocks. Such blocks are a novel type of signal, which can be detected from dense genotype or sequence data from individuals. The detection of shared blocks is already feasible for human data, and will likely become practical for many other species in the near future (Browning and Browning, 2012). Long IBD blocks are genetic traces of recent coalescence events, and their lengths hold information about the age of recent co-ancestry. This signal holds much promise for demographic inference, as several drawbacks of existing methods can be avoided. I develop a method that fits sharing of IBD blocks in a two-dimensional population with limited migration. I use the signal that block sharing decays with increasing sample distance. As the study of IBD blocks is rather novel subfield in popula-

tion genetics, I develop new theory and a way to fit this decay. I show that the decay of IBD block sharing with distance mostly depends on block map length and on the dispersal rate σ^2 . Importantly, this method can be used to learn about the dispersal rate σ^2 separately, which is not possible when using traditional methods that are based on similarities of single markers. This chapter is based on Ringbauer et al. (2017a).

In Chapter 3, I introduce a method to infer barriers to gene flow in two-dimensional populations. This scheme uses geographically local fluctuations of allele frequencies as a signal. I extend existing theory that describes the effect of a barrier, utilizing a model in which movement of lineages back in time is approximated by a partially reflected random walk. The method fits the strength of a linear barrier by fitting data to this explicit population genetic model. The fitted parameters can be directly linked to existing population genetic theory (Nagylaki, 1988; Barton, 2008). This bridges a gap, as most existing methods to infer barriers are not based on any population genetics model, and only heuristically fit the effects of a barrier. This chapter is based on Ringbauer et al. (2017b).



2. Inferring recent demography from isolation by distance of long shared sequence blocks

Abstract. Recently it has become feasible to detect long blocks of nearly identical sequence shared between pairs of genomes. These IBD blocks are direct traces of recent coalescence events and, as such, contain ample signal to infer recent demography. Here, we examine sharing of such blocks in two-dimensional populations with local migration. Using a diffusion approximation to trace genetic ancestry, we derive analytical formulae for patterns of isolation by distance of IBD blocks, which can also incorporate recent population density changes. We introduce an inference scheme that uses a composite likelihood approach to fit these formulae. We then extensively evaluate our theory and inference method on a range of scenarios using simulated data. We first validate the diffusion approximation by showing that the theoretical results closely match the simulated block sharing patterns. We then demonstrate that our inference scheme can accurately and robustly infer dispersal rate and effective density, as well as bounds on recent dynamics of population density. To demonstrate an application, we use our estimation scheme to explore the fit of a diffusion model to Eastern European samples in the POPRES data set. We show that ancestry diffusing with a rate of $\sigma \approx 50\text{--}100 \text{ km}/\sqrt{\text{gen}}$ during the last centuries, combined with accelerating population growth, can explain the observed exponential decay of block sharing with increasing pairwise sample distance.

THERE has been a longstanding interest in estimating demography, as migration and population density are key parameters for studying evolution and ecology. Demographic models are essential for disentangling the effects of neutral evolution from selection, and are crucial to understanding lo-

cal adaptation. Moreover, the inference of demographic parameters is important for conservation and breeding management. Given the intensive nature of obtaining such parameters by direct observations, which are moreover necessarily limited to short time scales, the increasing availability of genetic markers has spurred efforts to develop inference methods based on genetic data.

This work focuses on a method to estimate dispersal rate and population density in two-dimensional habitats by analyzing the geographic distribution of so called identity by descent (IBD) blocks, which are commonly defined as coinherited segments delimited by recombination events (Fig. 2.1). It has now become feasible to detect long regions of exceptional pairwise similarities from dense SNP or whole genome sequences (Gusev et al., 2009; Browning and Browning, 2011). For regions longer than a few cM, the bulk mostly consists of a single IBD block unbroken by recombination, at least when inbreeding is rare (Chiang et al., 2016). This yields novel opportunities for inferring recent demography, as one can study the direct traces of coancestry.

Moreover, the length of shared blocks contains information about their age. That is, the longer the time to the most recent common ancestor, the shorter the expected IBD block length, as recombination has more chances to break up ancestral genetic material. The probability that no recombination occurs in a block of a given map length decays exponentially going back in time. Hence, long IBD blocks originate mostly from very recent coancestry and provide insight into the recent history of a population. Shared long blocks between pairs of populations can be used to infer the distribution of recent coalescence times (Ralph and Coop, 2013), and fitting deme and island models can yield information on recent population sizes (Palamara et al., 2012; Browning and Browning, 2015) and migration patterns (Palamara and Pe'er, 2013). These works are complementary to the analysis of short identical segments, which are informative about deeper times scales (Li and Durbin, 2011; Harris and Nielsen, 2013), and they showcase the utility of long IBD blocks for inferring recent demography.

Here, we focus on a pattern of isolation by distance of IBD blocks within populations extended in two dimensions with local migration. For such populations, the classical Wright-Malecot formula describes an increase of mean pairwise genetic diversity with increasing geographic separation (Wright, 1943; Malécot, 1948). Several inference methods utilize such classical isolation by distance patterns as signals to infer the parameters of recent demography. For example, fitting increasing pairwise genetic diversity with geographic distance is widely used (Rousset, 1997, 2000; Vekemans and Hardy, 2004), and ABC methods have been applied (Joseph et al., 2016). Similarly, the extent of geographic clustering of rare alleles can be used as a signal for inference (Novembre and Slatkin, 2009). While the signal of locally decreased pairwise genetic

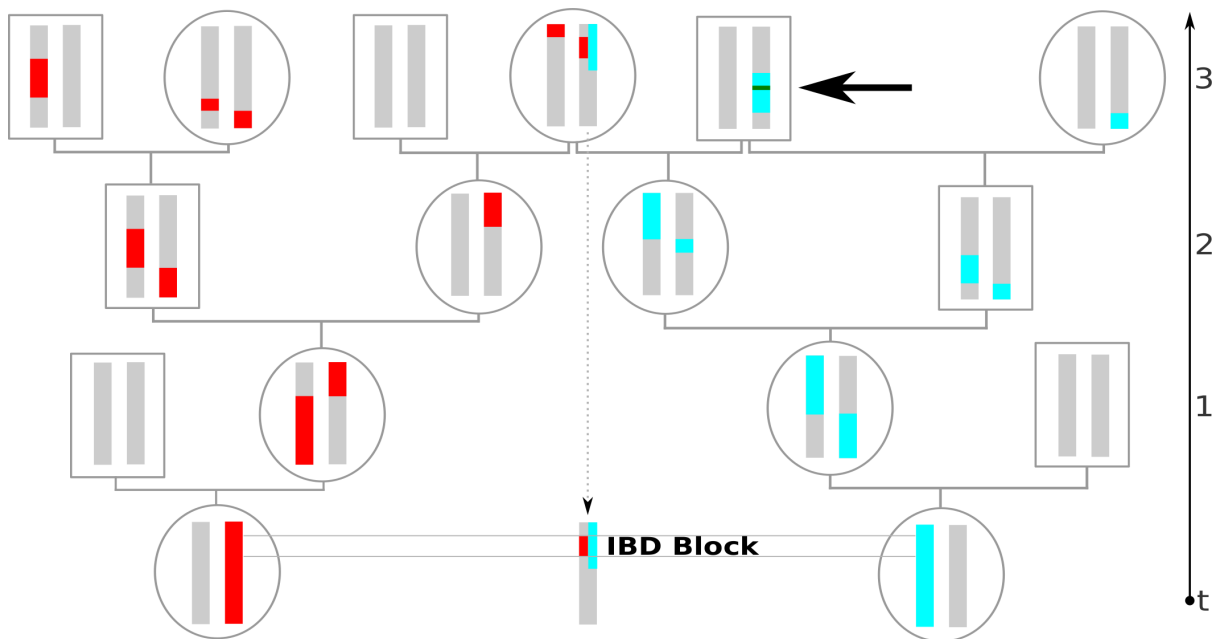


Figure 2.1: Example of an IBD block coinherited from a common ancestor three generations back. Going back in time, recombination splits up genetic ancestry (colored red and blue here) into blocks distributed among ancestors. If, as depicted here, such ancestral blocks overlap in a recent common ancestor, the intersecting stretch of the genome will be shared and both individuals will carry few distinguishing mutations. Here, we define IBD blocks to be delimited by any recombination events on the genealogical path to the most recent common ancestor. Thus, the recombination events that are fused again quickly by inbreeding loops, as depicted by the blue chromosome (thick arrow), also delimit IBD blocks. However, this recombination is not detectable in practice, and the two adjacent IBD blocks would be identified as one long IBD segment.

diversity mostly stems from recent times (Leblois et al., 2004), such patterns can be severely confounded by deeper, often unknown ancestral patterns (Meirmans, 2012). Moreover, such methods can usually only infer the neighborhood size $4\pi D_e \sigma^2$, which is proportional to the product of dispersal rate σ^2 with effective density D_e . Usually, these important parameters cannot be estimated separately, as the underlying signal is mostly based on a short-term equilibrium between local drift and dispersal. An exception is quickly mutating organisms such as viruses, for which phylogeographic diffusion approaches yield separate estimates of σ (Lemey et al., 2010). However, the mutation rates are usually too low to provide significant additional information on recent demography (Barton et al., 2013). In summary, inference schemes based on pairwise genetic diversity suffer from several fundamental limitations.

To overcome these problems, this work builds upon the ideas of Barton et al. (2013), who observed that the analysis of long shared IBD blocks would, in principle, allow one to estimate dispersal and population density separately. They argued that such an inference scheme would be robust to confounding by ancestral structure, since long IBD blocks mostly originate from not long ago. Here, we introduce a practical inference scheme based on this idea. We first expand the theoretical results of Barton et al. (2013). We utilize a model of spatial diffusion of ancestry, which yields analytical formulae for block sharing patterns. We then fit these results using a composite likelihood framework, similar to Ralph and Coop (2013). This approach allows one to readily include error estimates for block detection, such as limited detection power or wrongly inferred block lengths which are problems that usually arise when IBD blocks are called from genotype data (Browning and Browning, 2012; Ralph and Coop, 2013). Recently, Baharian et al. (2016) have independently derived similar equations for block sharing under the diffusion approximation and used them for demographic inference by fitting binned data. We extend this work in several ways. We additionally deal with growing and declining populations, and our composite likelihood method offers several significant advantages over fitting binned data. Importantly, as a major part of this paper, we extensively evaluate our estimation scheme against simulated data. We test its power to recover demographic parameters for several geographic models and we investigate how model deviations, such as nearby habitat boundaries, affect inference. This yields valuable novel insight into the validity of the underlying idealized diffusion model and examines the scope of the inference scheme.

Currently, large IBD block data sets are available mainly for humans. To showcase a practical application of our inference scheme, we use it on a subset of the POPRES dataset, which Ralph and Coop (2013) previously analyzed for long IBD blocks. Although human demography is without doubt very complex, the diffusion model pro-

vides a good fit to the data, which allows us to draw conclusions about the extent of human ancestry spread in continental Europe during the last centuries. We also infer a rapidly increasing population density, which stresses the importance of accounting for rapid population growth when analyzing human IBD block sharing.

2.1 Materials and Methods

2.1.1 The Model

To describe block sharing in two spatial dimensions with local migration, we use two basic model assumptions to approximate a wide range of scenarios. Obviously, the true demographic history of a population is more complex than any such simple model. Thus, the aim is not to have a mathematically rigorous model, which is often formally problematic (Felsenstein, 1975; Nagylaki, 1978) and only holds exactly in specific settings, but to have an accurate approximation that captures general patterns that can be used for robust inference of basic demographic parameters. In the following, we outline these two central modeling assumptions.

Poisson recombination

We approximate recombination as a homogeneous Poisson process, i.e. crossover events are assumed to occur at a uniform rate along a chromosome. Throughout this work, the unit of genetic distance will therefore be the Morgan, which is defined as the distance over which the expected average number of intervening chromosomal crossovers in a single generation is one. Small scale processes, such as gene conversion, are not captured by the Poisson approximation, but for the large genomic scales of typically several cM considered here this can be neglected (Lynch et al., 2014). Similarly, we ignore the effect of interference, which is reasonable when describing the effects of recombination over several generations. Since the female and male recombination rate can be markedly different, for our purposes we use the sex-averaged rate $r = \frac{r_m + r_f}{2}$. In every generation, loci on autosomes have an equal chance to trace back to a female or male ancestor. Thus, the female and male Poisson processes together are described by a single Poisson process with the averaged recombination rate. Generalizing this line of thought, any individual differences in map length can be modeled by a single Poisson process with the population-averaged rate.

Diffusion approximation

Following a long tradition of modeling individual movement in space by diffusion (Fisher, 1937; Wright, 1943; Malécot, 1948; Nagylaki, 1978), we approximate the spatial movement of genetic material back in time using a diffusion process. The position of ancestral material at some time t in the past is the sum of the migration events until then, which are often correlated only on small timescales. Therefore, using the central limit theorem, the probability density for the displacement of a lineage can be approximated using a Gaussian distribution with axial variance of $\sigma^2 t$ (Fig. 2.2). This approximation does not depend on details of the single-generation dispersal kernel, provided its variance is finite. It seems plausible that diffusion of ancestry is often an accurate approximation on recent to intermediate timescales (Barton et al., 2002), which are important for the sharing of long IBD blocks.

If consecutive single-generation dispersal events are uncorrelated, σ^2 is the average squared axial parent-offspring distance (Rousset, 1997). Even with small-scale spatial or temporal correlations between dispersal events, one can model the spread of ancestry using the diffusion approximation (Robledo-Arnuncio and Rousset, 2010). In this case, σ^2 has to be interpreted as a parameter that describes the rate of the spread of ancestry back in time (Barton et al., 2002), which can differ markedly from the single generation squared axial parent-offspring distance.

Here, we will need to describe the chance that pairs of lineages of homologous loci come into close proximity. For this, we assume that the two lineages diffuse independently. In this case, the sum of their movements can be described using a two-dimensional Gaussian distribution with twice the variance of a single lineage. The probability density that two lineages that were initially separated by (x_0, y_0) have a pairwise distance of 0 along each axis at time t , or equivalently, that the sum of the movements is $(-x_0, -y_0)$, is therefore:

$$\frac{1}{4\pi t\sigma^2} \exp\left(-\frac{x_0^2 + y_0^2}{4t\sigma^2}\right) = \frac{1}{4\pi t\sigma^2} \exp\left(-\frac{r^2}{4t\sigma^2}\right), \quad (2.1)$$

where $r = \sqrt{x_0^2 + y_0^2}$ is the initial Euclidean distance between the two lineages.

This ignores the fact that once coalesced, lineages remain at a pairwise distance of 0. Wilkins (2004) gave recursions and approximate formulae (Form. A15, A17), that account for this interference of lineages in two spatial dimensions. They show that complex interference terms can be neglected as long as previous coalescence is sufficiently rare. Thus, for describing the chance of pairwise coalescence in the relatively recent past, Eq. 2.1 usually represents an accurate approximation, particularly for well-separated samples. Other causes of correlations of movements are often of local geo-

graphic nature, as in the cases of density fluctuations or local barriers. Such small-scale heterogeneities often average out when viewed on larger scales, and the approximation that lineages move independently remains accurate on these scales (Barton et al., 2002).

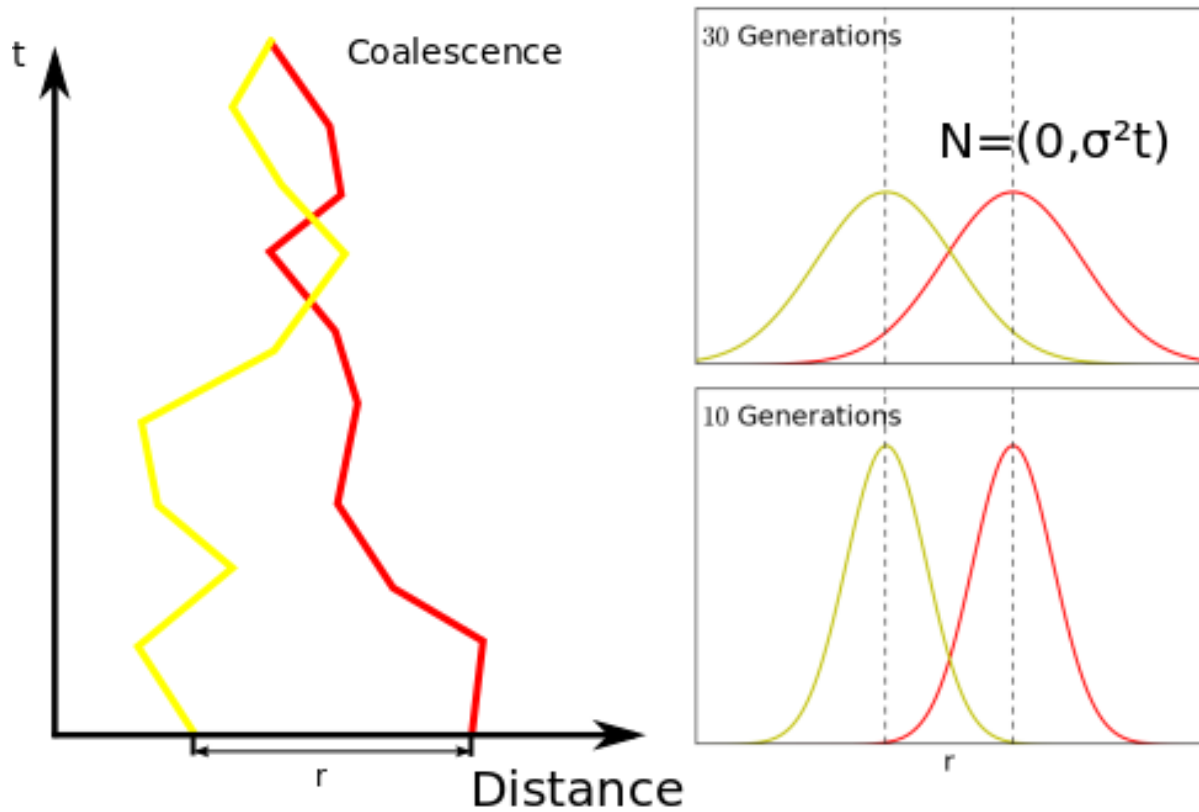


Figure 2.2: Diffusion model visualized in one spatial dimension. Our model is in fact two-dimensional, but is qualitatively similar. Left: One realization of the movement of ancestry of two homologous loci initially separated by distance r . In our model, there is a chance that they coalesce every time they come close, which is indirectly proportional to the local effective density parameter D_e (Appendix A). Right: In our model, the probability density function of having moved distance Δx at time t generations back spreads out as a Gaussian distribution $N(0, \sigma^2 t)$ with linearly increasing variance of $\sigma^2 t$.

2.1.2 IBD Sharing in the Model

Using similar assumptions, Barton et al. (2013) calculated the probability that two individuals a certain distance apart share an IBD block longer than a minimum length starting from a specified locus. For this specific purpose they could directly apply the Wright-Malecot formula by replacing mutation with recombination. For practical inference from IBD blocks, more general formulae describing the total number of shared

blocks of a specific length L are advantageous. In this section, we derive such equations.

IBD blocks of age t

Following Ralph and Coop (2013), we first partition N_L , the number of shared blocks of map length L for a given pair of samples, into N_L^t , the number of blocks coalescing at time t :

$$\mathbb{E}[N_L] = \int_0^\infty \mathbb{E}[N_L^t] dt. \quad (2.2)$$

Throughout this paper, such terms are always understood as a density with respect to block length and time. Following the ancestry of two chromosomes back in time, a change of genealogy only occurs when there is a recombination event somewhere along the lineage. Between these discrete jumps, genetic material can be traced as a single locus. This allows us to further split $\mathbb{E}[N_L^t]$ into the product of the expected number of blocks of length L obtained by splitting the two chromosomes according to the Poisson recombination over time t with the probability that a single locus coalesces time t ago. We denote the first factor by $\mathbb{E}[K_L^t]$, and the second factor, commonly known as the coalescence time distribution, by $\psi(t)$:

$$\mathbb{E}[N_L^t] = \mathbb{E}[K_L^t]\psi(t). \quad (2.3)$$

Number of candidate blocks

Under our model assumptions, the position of all recombination events on two independent chromosomes traced back until time t is given by a Poisson process with rate $2t$. The expected number of all block pairs overlapping at an intersection length L can then be calculated as follows. A recombination event occurs in a small region of map length ΔL with a probability of $2t\Delta L$, and the probability that a region of length L does not recombine follows the exponential distribution $\exp(-2Lt)$. For chromosomes of map length G , summing the possible start sites yields the expected total number of blocks of length L :

$$\mathbb{E}[K_L^t] = 2 \cdot 2t \exp(-2Lt) + (G - L)4t^2 \exp(-2Lt), \quad (2.4)$$

where the first term describes the blocks starting at either edge and the second term the fully interior blocks, which require two delimiting recombinations. Neglecting the effects of chromosome edges ($G \gg L$), this is approximated by:

$$\mathbb{E}[K_L^t] \approx G4t^2 \exp(-2Lt). \quad (2.5)$$

We will use Eq. 2.5 to derive an approximate formula for capturing the qualitative behavior of mean IBD sharing. The slightly more complex result, including edge effects is derived analogously (Appendix B) and is used for inference.

Single locus coalescence probabilities

The probability $\psi(t)$ that two homologous loci have their last common ancestor time t ago depends on their pairwise sample distance r and the parameters of the demographic model. We can follow Barton et al. (2002) and approximate the probability of a recent coalescence as the product of the probability of the pairwise sample distance being 0 (Eq. 2.1) and a rate of local coalescence that, following Barton et al. (2013), we shall denote by $1/(2D_e)$. In Appendix A, we justify this approximation and give a formal definition of this so-called effective density D_e . In order to describe a globally growing or declining population, which is particularly important for the human case studied in this chapter, we let D_e depend on time t . Together, this yields:

$$\psi(t) = \frac{1}{2D_e(t)} \frac{1}{4\pi t\sigma^2} \exp\left(-\frac{r^2}{4t\sigma^2}\right). \quad (2.6)$$

Full formula

Substituting Eq. 2.5 and Eq. 2.6 into Eq. 2.3 gives:

$$\mathbb{E}[N_L^t] = \frac{Gt}{2D_e(t)\pi\sigma^2} \exp\left(-\frac{r^2}{4t\sigma^2} - 2Lt\right). \quad (2.7)$$

To determine the total number of expected shared blocks, we have to integrate all possible coalescence times t . For the class of power density functions, where

$$D_e(t) = Dt^{-\beta} \quad D > 0, \beta \in \mathbb{R}, \quad (2.8)$$

the integral yields explicit formulae. The important case of $\beta = 0$ models a constant population density, while $\beta > 0$ and $\beta < 0$ describe populations with a growing or declining density, respectively. With $\beta > 0$, the density approaches infinity for $t = 0$, which corresponds to a negligible chance of coalescence at the present. However, since we effectively fit block sharing on intermediate timescales (Fig. 2.9), this obvious problem of the model is not very limiting in practice. This class of functions has been used to

fit human demographic growth (Von Foerster et al., 1960). Importantly, linear combinations of such terms can be used to build more complex density functions, including polynomials for the special case $\beta \in \mathbb{N}$, which then also yield analytical formulae.

Performing the integral of Eq. 2.2 gives the main result:

$$\mathbb{E}[N_L] = 2^{\frac{-3\beta}{2}-3} \frac{G}{\pi D \sigma^2} \left(\frac{r}{\sqrt{L}\sigma} \right)^{2+\beta} K_{2+\beta} \left(\sqrt{2L} \frac{r}{\sigma} \right). \quad (2.9)$$

Integrating this formula with respect to the block length gives the expected number of shared blocks longer than the threshold length L_0 :

$$\mathbb{E}[N_{>L_0}] = \int_{L_0}^{\infty} \mathbb{E}[N_L] \, dL = 2^{\frac{-5-3\beta}{2}} \frac{G}{\pi D \sigma^2} \left(\frac{r}{\sqrt{L_0}\sigma} \right)^{1+\beta} K_{1+\beta} \left(\sqrt{2L_0} \frac{r}{\sigma} \right), \quad (2.10)$$

where K_γ is the modified Bessel function of the second kind of degree γ (Abramowitz and Stegun, 1964). We analyze Eq. 2.9 and 2.10 qualitatively in the discussion section, and Fig. 2.3 depicts their accuracy on simulated data.

For a widely used functional form of population density change, an exponential growth with rate β , the integral converges only for blocks of length $2L > \beta$. Otherwise, the exponential rate at which the long blocks are broken up is slower than the exponentially increasing chance of local coalescence, and the expected number of blocks does not vanish for large t . However, we can approximate exponential growth on intermediate timescales by using its standard Taylor expansion up to a certain term, and then applying our results for the power density functions. Again, this effectively fits a population density up to the intermediate timescales, where the truncated Taylor approximation is accurate, while circumventing the pathological behavior of the distant past.

2.1.3 Fitting the Model to Data

To learn about recent demography, we fit the observed block-sharing between a set of samples to Eq. 2.9. Here, we use a likelihood method, in which we approximate the likelihood function $f : \theta \rightarrow \Pr(x | \theta)$ of the observed data x for a given set of parameters θ (σ, D, β) with a composite likelihood $\tilde{f}(\theta)$. This allows us to estimate the approximate standard deviations and confidence intervals from the empirical Fisher information matrix. One can utilize standard numerical optimization techniques to find the maximum likelihood estimates $\hat{\theta}_{\text{MLE}}$. In our analysis we use the Nelder-Mead method, as implemented in the class `GenericLikelihoodModel` of the Python package `statsmodels`, which proved to be numerically robust and quick.

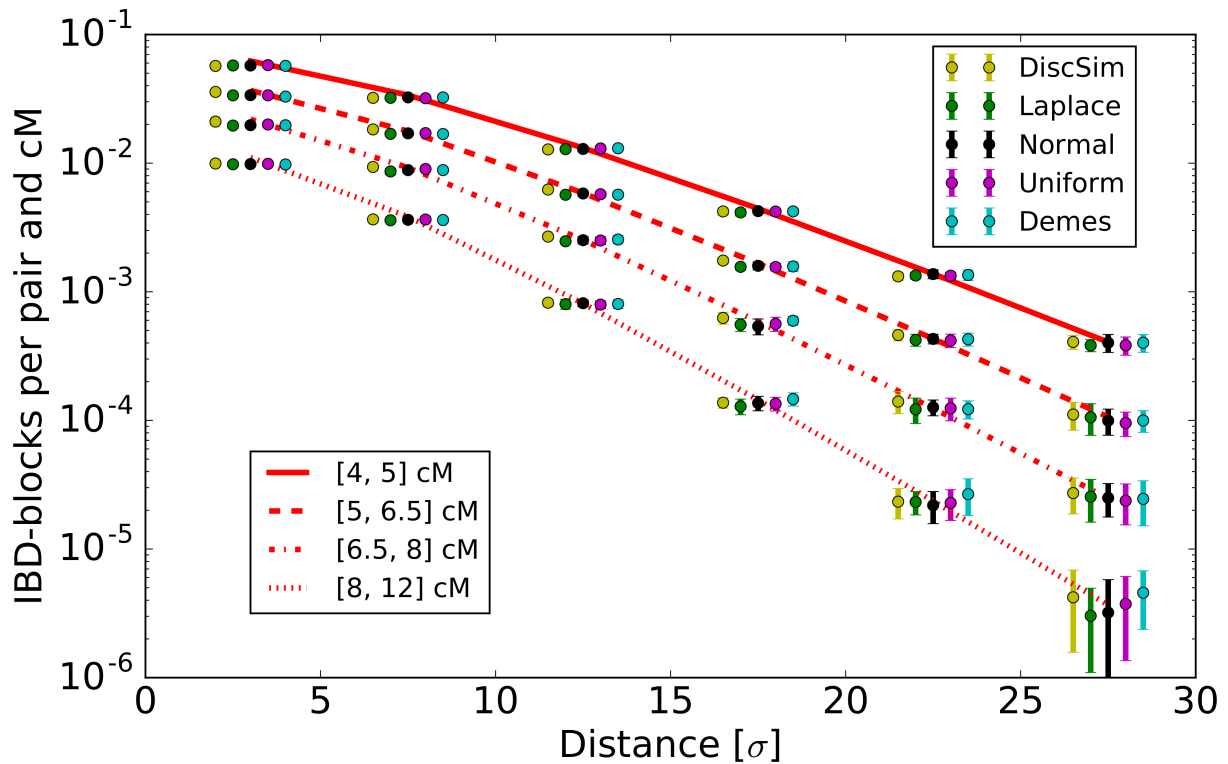


Figure 2.3: Simulated IBD block sharing compared with theoretical expectations. We show values normalized to give rates per pair and cM. Theoretical expectations are calculated for each length bin using Eq. 2.10. For the five models described in the Methods section, we kept the population density constant at $D_e = 1$, with a dispersal rate of $\sigma = 2$ on a torus of size 180, and simulated IBD block sharing between 150 cM chromosomes spread out on a sub-grid with nodes 2 distance units apart (for a full set of specific simulation parameters, see 2.5.1). For every model, we ran 20 replicate simulations. Distances are measured in dispersal units (so that $\sigma = 1$) and error bars depict the estimated standard deviations for each bin among the 20 runs to visualize the uncertainty of the estimates. Dots are spread out for better visualization around their original positions (middle dot).

Poisson model

We can construct an approximate likelihood of observed block sharing by using an approach that follows Ralph and Coop (2013). First, for every pair of samples, we bin block sharing with respect to shared block length into small length bins. Then, we model the number of shared blocks within each of these bins as independent Poisson distributions around expected rates λ_i , which, for a small enough bin $[L_i, L_i + \Delta L]$ can be approximated using Eq. 2.9:

$$\lambda_i(r, \theta) = \mathbb{E}[N_{L_i}(r, \theta)]\Delta L. \quad (2.11)$$

Using this equation, we can calculate a composite likelihood of the observed data given the demographic parameters θ (Appendix C).

Block detection errors

The detection of IBD blocks from genetic data is not a trivial task. In practice, one often has to deal with erroneous detection (Browning and Browning, 2012). Blocks might be called in the absence of true IBD blocks (false positives), and only a fraction of true IBD blocks of a given length are detected (limited power), and there is a probability of assigning them the wrong length (error). Following Ralph and Coop (2013) we can include these errors into our likelihood framework. Careful analysis allows one to estimate block detection errors (Ralph and Coop, 2013), and the expected rates per bin can be updated accordingly (Appendix D).

Assumption of Independence

This Poisson approximation assumes that all shared blocks are the outcomes of independent processes. This is obviously an over-simplification. Block-sharing can be correlated along chromosomes and among different sample pairs because of the initially shared movement of genetic material. Taking all these correlations into account would go beyond the simple pairwise diffusion model. However, maximizing the likelihood of actually correlated observations (composite likelihood) is a widely used practice in inference from genetic data (e.g., Fearnhead and Donnelly (2002)). It still gives consistent and asymptotically normal estimates, although the errors calculated from the curvature of the maximum-likelihood surface at its maximum (Fisher-Information matrix) will be too tight when the observations are actually correlated (Lindsay, 1988; Coffman et al., 2016). Moreover, in many cases, correlations among blocks can be expected to remain fairly weak since initial correlations in spatial movement are broken up quickly

by recombination. When analyzing well-separated samples, sharing of long blocks is a rare event and, thus, most of the observed block-sharing will originate from independent coalescent events.

Adjacent IBD blocks

The theory and inference scheme introduced here are based on IBD blocks that have been defined to be ended by any recombination event on the path to the common ancestor. However, multiple, consecutive IBD blocks of recent coancestry, for unphased data from all four possible pairings of two sets of diploid chromosomes, produce an unbroken segment of exceptionally high similarity that is detected as a single long IBD block in practice. This can significantly inflate the number of observed IBD blocks of a given length beyond the true value, especially for shorter IBD blocks (Chiang et al., 2016).

If adjacent IBD blocks happen to be neighbors, the error estimation model by Ralph and Coop (2013), which is based on introducing artificial IBD blocks of known length partly accounts for this effect. However, this does not estimate the effect of short inbreeding loops where ancestral genetic material that was broken up by a recombination fuses together again quickly (Fig. 2.1), rendering the IBD block ending recombination event ineffective (Barton et al., 2013).

Intuitively, for a large neighborhood size (a parameter proportional to the product of σ^2 and effective density $4\pi\sigma^2 D_e$) short inbreeding loops that significantly extend a recent IBD block by quick re-coalescence are rare, and our approach remains valid. However, for a population with small neighborhood size, ineffective recombination events can potentially confound observed block-sharing patterns and the estimates based on them. This effect is driven mostly by coalescence within a few generations, before the blocks migrate away from each other. Hence, local dispersal and breeding patterns are important; however, the diffusion of ancestry usually only becomes accurate on intermediate timescales. Therefore a generally applicable theoretical treatment of this issue is not feasible. In this chapter, we study the effects of this model inaccuracy using simulations, and show that the inference scheme is not greatly affected in the simulated scenarios.

2.1.4 Simulations

To test our equations and inference scheme, we simulated the sharing of IBD blocks in a set of samples by tracing the ancestry of the chromosomes back in time for a vari-

ety of spatial population models. Simulations were mostly done on a two-dimensional torus that was large enough that IBD block sharing over more than half of the torus was very unlikely; thus we effectively simulated a two-dimensional population without boundary effects. Since sharing of long IBD blocks is very unlikely to originate far back in time, we ran the simulations up to a maximal time t_{max} . If not otherwise stated, we analyzed the sharing of true IBD blocks, in which every recombination event was assumed to be effective.

Grid models

In our grid models, the nodes of a rectangular grid were occupied by a prespecified number of pairs of homologous chromosomes to mimic diploid individuals. Similar to the classic Wright-Fisher model of panmictic populations, for every chromosome, a parent was chosen independently for every discrete generation back, with the probabilities described by a prespecified dispersal kernel. Poisson recombination events along the chromosome induced a switch between the two parental chromosomes. Whenever the ancestral material of the two distinct initial chromosomes fell on the same chromosome and overlaps for longer than a given threshold chromosome length, we stored the resulting IBD block. We simulated the dispersal following discretized uniform, Gaussian, and Laplace probability densities along each axis to have representatives of dispersal kernels with low, intermediate, and high kurtosis. To analyze the effects of a growing or declining population density, we simulated a varying number of multiple pairs of homologous chromosomes per node. A chromosome then first picks an ancestral node as before, and subsequently a random diploid ancestor from this node. The grid model was also easily modified to simulate a classic nearest neighbor stepping stone model (Kimura and Weiss, 1964). Nodes were grouped into demes, and each chromosome either chose its parent uniformly from within its own or one of the neighboring demes.

Continuous model: Spatial Lambda-Fleming-Viot Process

We additionally simulated a model in which each individual occupied a position in continuous space. For this, we utilized DISCSIM, a fast implementation (Kelleher et al., 2014) of the recently introduced spatial Lambda-Fleming-Viot process. Summarizing briefly, this model introduced by Barton et al. (2010) follows lineages backwards in time and events are dropped randomly with a certain rate parameter and uniform spatial density. In each such event, every lineage within radius R is affected with the probability u by this event. A prespecified number of parents, here two, are dropped uniformly

within the disc, and every affected lineage jumps to them, switching parents according to the recombination rate. Given an initial set of loci, DISCSIM generates their coalescence tree up to a specified time. The output contains a list of all coalescent nodes, which we further analyzed to detect IBD block sharing.

2.1.5 Application to Eastern European Data

Currently, population genomic datasets which allow one to analyze long IBD blocks are available mainly for humans. To test the inference scheme, we applied a dataset of blocks shared between Europeans, which was generated previously by Ralph and Coop (2013), and includes detailed error estimates for IBD block detection. They reported significant differences in patterns of block sharing between Eastern and Western European populations. Therefore, we concentrated our analysis on block sharing in the Eastern European subset, as diffusion should be a better approximation for modeling the spread of ancestry in continental regions. Moreover, Eastern European countries are on average geographically more compact and, thus, the position data at the country level is expected to be more accurate.

The data

The detection method and the error analysis of the IBD block data were described in detail by Ralph and Coop (2013). Summarizing briefly, IBD blocks were called for a subsample of the POPRES dataset (Nelson et al., 2008) and genotyped at $\sim 500,000$ SNPs using the fastIBD method, as implemented in Beagle v3.3 (Browning and Browning, 2011). Every sample used in the analysis was required to have all reported grandparents from the same country. We analyzed block sharing between 125 Eastern European samples (see also 2.5.1). We followed the geographic classification of Ralph and Coop (2013), but excluded the six Russian and one Ukrainian samples, as location data at the country level are likely very inaccurate for these two geographically extended countries. We analyzed shared blocks longer than 4 cM. Within our subsample, 1,824 such blocks were reported (Fig. 2.9). We set the position of each country to its current demographic center, defined as the weighted mean location (2.5.1). In our analysis, we used sex average map lengths of autosomes given by the Decode map (Kong et al., 2002), consistent with Ralph and Coop (2013).

Data analysis

Throughout the analysis, we worked with block length bins ranging from 0 to 30 cM with a bin width of $\Delta L = 0.1$ cM, and applied the error function estimates reported by Ralph and Coop (2013). For maximizing the likelihood, we calculated the likelihood of block sharing in the bins from 4 cM to 20 cM, which is informative about the last few centuries (Fig. 2.9). We excluded the longer shared blocks from our analysis since these blocks have a considerable chance of originating in the last few generations, which is not expected to be accurately captured in the diffusion model. Longer shared blocks are also confounded by the sampling scheme that excluded individuals with reported grandparents from different countries.

We used our inference scheme to fit several specific models of past density D as follows:

- For a constant population: $D = C$.
- For a population growing at accelerating rate: $D = C/t$.
- For a growth model where the growth rate is fitted as well: $D = Ct^{-\beta}$.

In each case, t measured time back in generations. To learn about the certainty of estimates, in addition to using the curvature of the likelihood surface (Fisher information matrix), we bootstrapped the data. Since we suspected strong correlations and systematic deviations from the model, we resampled different units. We bootstrapped on the level of blocks by redrawing each block a number of times following a Poisson distribution of mean 1, and similarly over country pairs, since we suspected systematic correlations on this level.

Furthermore, we analyzed the deviation of pairwise block sharing between pairs of countries from the expected value predicted by the best fit model. For this, we assumed that the observed block sharing was Poisson distributed around the predicted block sharing. Transforming the block count data $x \rightarrow 2\sqrt{x}$ converts these Poisson distributions into approximately Gaussian distributions with standard deviation 1, which helped visual inspection of the statistical significance of residuals.

Data Availability

We implemented the described methods to simulate and analyze IBD block sharing data in Python. The source code was uploaded to the freely available Github repository <https://git.ist.ac.at/harald.ringbauer/IBD-Analysis>. The preprocessed human IBD block

sharing data, including the detection error estimates used here, were the result of the analysis of Ralph and Coop (2013), and can be freely accessed at <http://www.github.com/petrelharp/euroibd>.

2.2 Results

2.2.1 Block-sharing in Simulated Data

We compared simulated block sharing patterns with the theoretical expectations. For each bin, we depicted rates per pair and normalized for a rate per cM.

Constant population density

For a constant population density the theoretical expectation (Fig. 2.3) is given by Eq. 2.9:

$$\mathbb{E}[N_L] = \frac{G}{8\pi D\sigma^2} \left(\frac{r}{\sqrt{l}\sigma} \right)^2 K_2 \left(\sqrt{2l} \frac{r}{\sigma} \right),$$

where K_2 is the modified Bessel function of the second kind of degree 2 (Abramowitz and Stegun, 1964). This formula predicts that block-sharing approaches exponential decay with distance, as Bessel functions $K_\gamma(x)$ converge to $\sqrt{\frac{\pi}{2r}} \exp(-r)$ for $r \gg 1$ (Abramowitz and Stegun, 1964). This decay then dominates the polynomial terms in front of the Bessel functions, and the slope of this exponential decay (on a log scale) converges to $\frac{\sqrt{2L}}{\sigma}$ as $\sqrt{2L} \frac{r}{\sigma} > 1$. For the long blocks considered here, this quick decay is approached for pairwise sample distances of a few σ . In all simulations, block-sharing patterns were very similar among the five different simulated models, and closely followed the theoretical expectation (Fig. 2.3).

Growing and declining populations

We simulated block sharing for three scenarios of a growing, declining, and constant population with growth parameters $\beta = 1, 0, -1$. Fig. 2.4 shows that the results are again in good agreement with theory. We depicted the result for the simulated Laplace dispersal. The other dispersal kernels yielded almost identical results. In all scenarios, the decay of block sharing with distance approached exponential decay with rate $\sqrt{2l}/\sigma$, where the specific density scenario determined the speed of convergence.

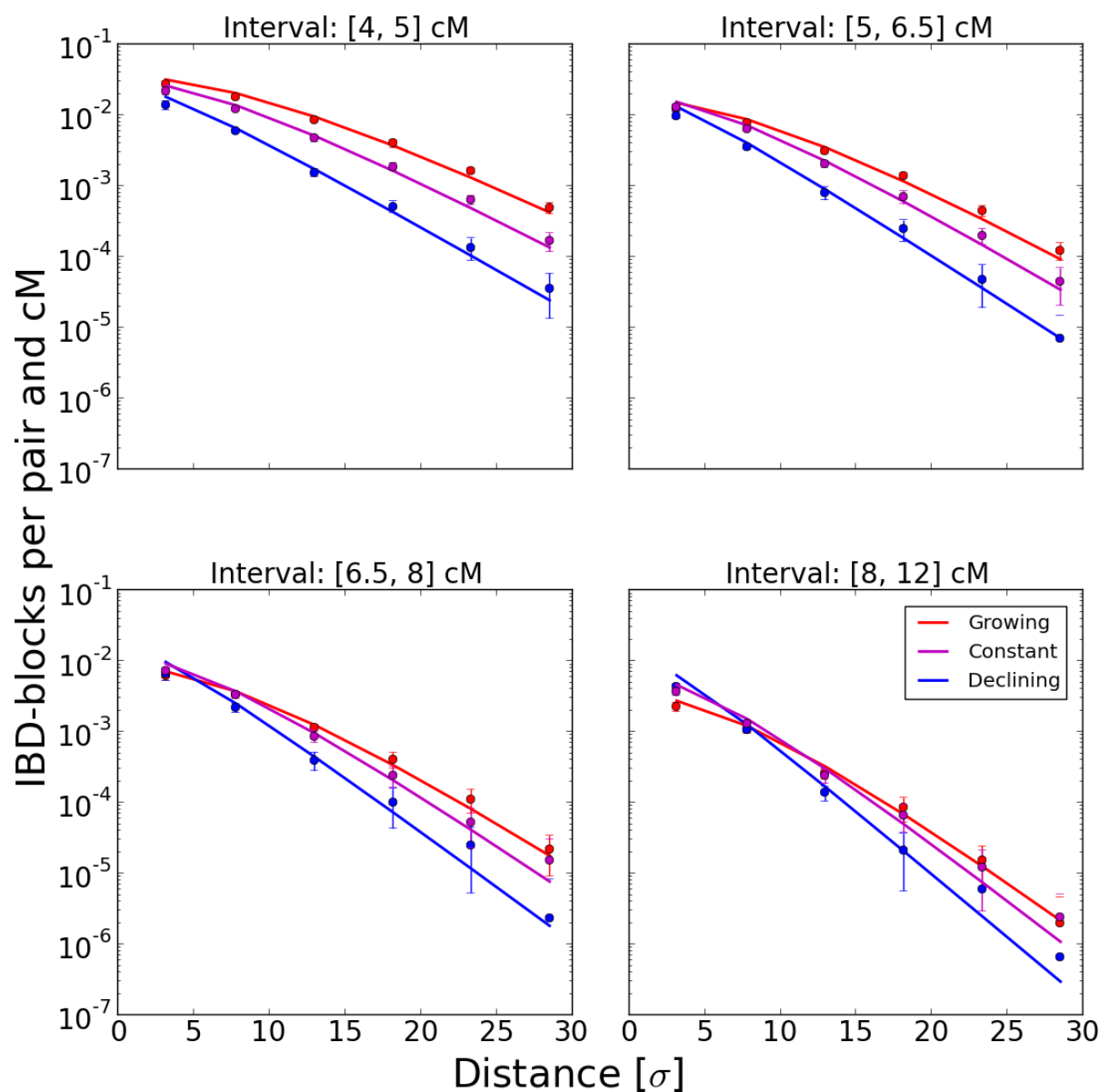


Figure 2.4: Various population density scenarios. Simulated IBD block sharing per pair and cM in various density scenarios was compared to theoretical expectations based on Eq. 2.9. The block sharing of a subset of 150 cM chromosomes 4 distance units apart placed on an initial grid was analyzed. Along each axis, dispersal was modeled by a Laplace distribution with $\sigma = 1$, and the number of diploid individuals per node n either remained constant at $n = 10$, grew as $n(t) = t$, or declined as $n(t) = 200/t$; in all cases, t denotes the time back measured in generations, and at every step, $n(t)$ was rounded to the nearest integer value. For each scenario, 20 replicate runs were done. Dots depict the mean and error bars the standard deviation for every bin. The solid lines show the theoretical prediction based on Eq. 2.9.

2.2.2 Inference in Simulated Data

We tested our parametric inference scheme and analyzed its ability to recover the underlying demographic parameters from simulated block-sharing data. For every simulated block data set, we numerically computed the maximum likelihood estimates θ_{MLE} for shared blocks between 4cM and 20 cM, put into bins of width 0.1 cM.

Constant population density

Results for the parameter inference for a population of constant density are depicted in Fig. 2.5. We simulated a varying number of samples on a grid, consisting of one chromosome each, to test the behavior of the inference scheme with respect to limited sample size. Naturally, the variance of estimates increased with decreasing sample size, but the bias remained small. Moreover, the estimated standard errors captured the true estimator variance relatively well (2.5.2). This confirms that most of the shared blocks were the result of uncorrelated coalescence events, as heuristically argued above. The typical log-likelihood surface for a single simulated IBD block sharing data set was found to be smooth (Fig. S2), and in all cases, numerical maximization did not result in spurious maxima, even for initial estimates orders of magnitudes off. Moreover, estimates of density and dispersal rate were only slightly correlated in the scenario considered here (2.5.2).

Varying population density

We also tested the ability of the inference scheme to detect recent changes in population densities. For this, we simulated three scenarios of a growing, declining, and constant population with growth parameters $\beta = 1, 0, -1$. Results are depicted in Fig. 2.6. The estimates of the demographic parameters allowed us to robustly distinguish these three scenarios. Interestingly, accurate estimates of the dispersal rate were feasible in all these demographic scenarios; even when fitting a model with constant population size to the other two scenarios of a recently quickly changing population size (Fig. S1). This can be explained by the fact that the eventual rate of decay, the main signal for estimating σ from fitting Eq. 2.9, remains the same, independent of the specific population density scenario. The speed of convergence varies, but in all cases, the eventual rate is approached relatively quickly within several dispersal distances (Fig. 2.4).

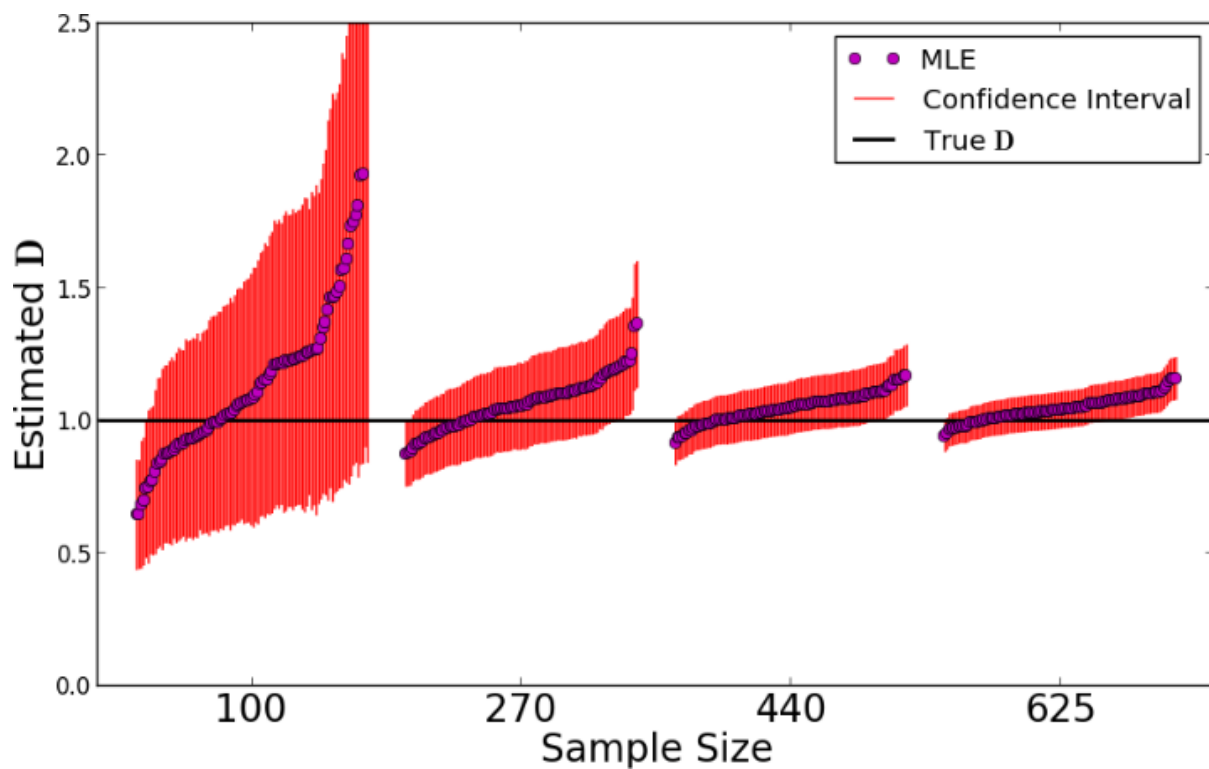


Figure 2.5: Maximum likelihood estimates. We simulated a Laplace model on a grid of nodes of size 180×180 for $t = 200$ generations back. We set the dispersal rate at $\sigma = 2$ and the number of individuals per node to $D = 1$. In every run, a random subset of 100, 270, 440 or 625 chromosomes of map length 150 cM was picked from an initial sample grid spaced two nodes apart. For each sample size, 100 simulations and subsequent parameter estimates were run. Every dot depicts the maximum likelihood parameter estimate of a single run. The 95% confidence intervals were calculated from the Fisher information matrix.

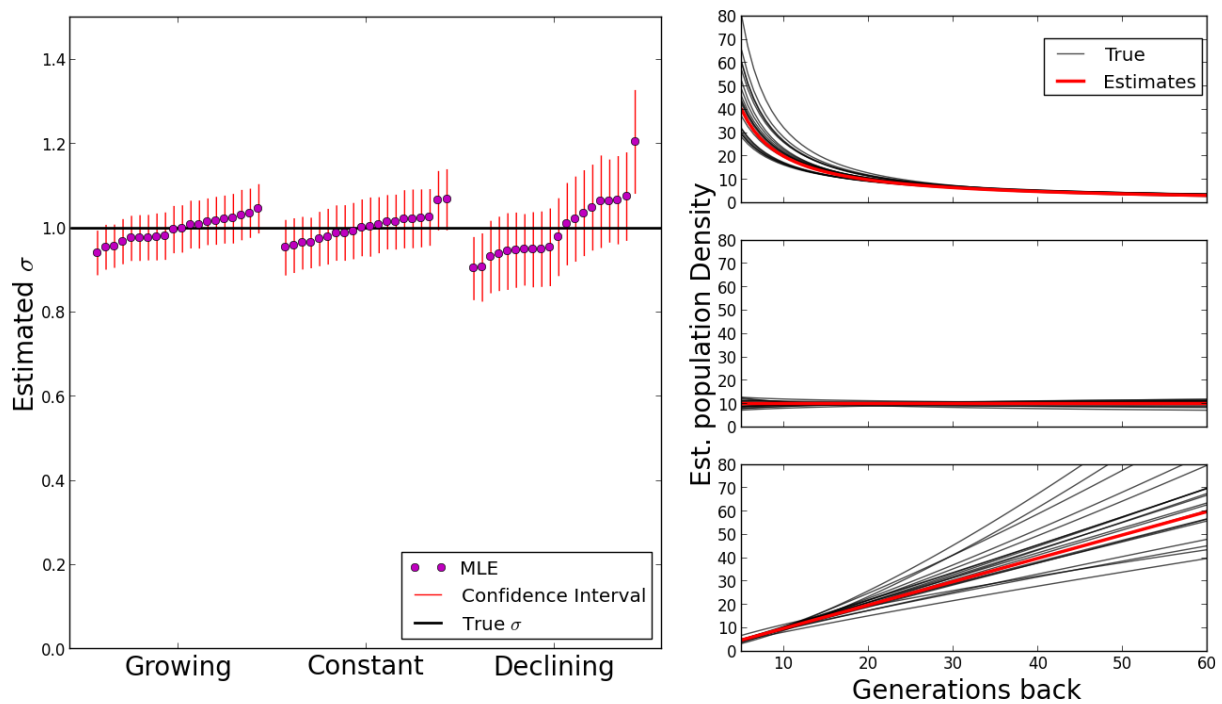


Figure 2.6: Likelihood estimates for various population density scenarios. The same scenarios used in Fig. 2.4 were simulated. For 20 runs each, 625 chromosomes of length 150 cM were randomly picked from a sample grid and traced back using a Laplace dispersal kernel with $\sigma = 1$; and the maximum likelihood fits and 95% confidence intervals were calculated from their block sharing. For the estimated population density, the true value of the simulations and the MLE estimate for every run are shown.

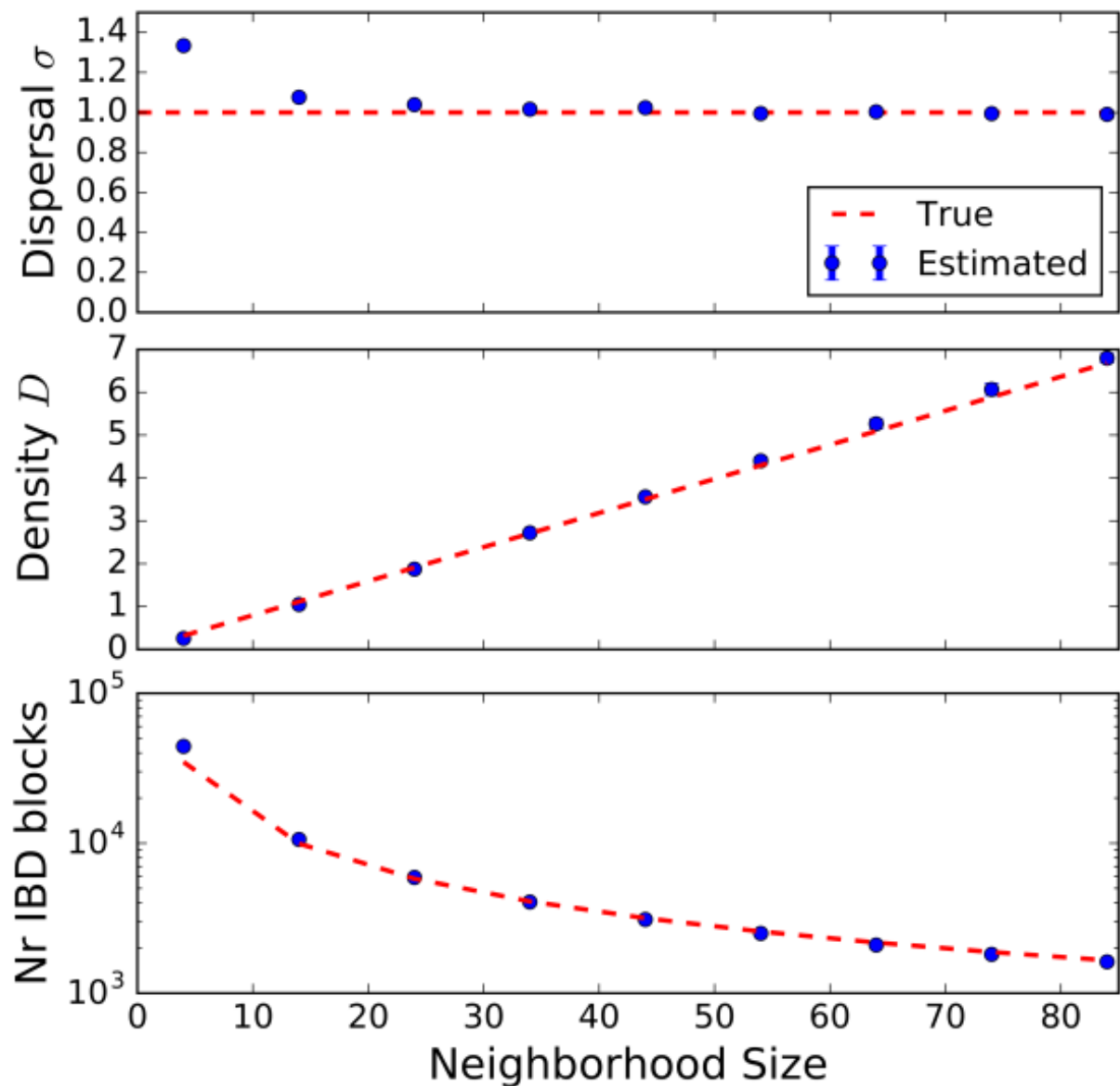


Figure 2.7: Observable IBD blocks compared with theoretical predictions for true IBD blocks. Simulations were run with DISCSIM for an initial grid of 150 cM chromosomes that were 3 distance units apart on a torus with axial size 90. Dispersal rate was set to 1. IBD blocks were detected as consecutive runs of coalescence times < 1000 generations, and then used to estimate demographic parameters. For various densities corresponding to neighborhood sizes 4 – 86, 10 DISC-SIM runs were simulated. The mean of these runs was compared with the theoretical prediction using Eq. 2.9 that assumes that every recombination event is effective.

True versus detectable IBD blocks

In Fig. 2.7, the effect of undetected recombination events on estimates of demographic parameters and overall IBD block number is depicted. This was investigated with simulations in the DISCSIM model, as it allowed easy and continuous tuning of the neighborhood size $4\pi\sigma^2 D_e$ through the parameter describing the probability that an event hits an individual within its range (Barton et al., 2013). Pairwise coalescence times for all pairs of loci along the chromosome were extracted, but now, only the effective recombination events were counted, which were defined as jumps of coalescence times between adjacent pairs of loci with at least one coalescence time older than a preset time threshold of 1000 generations back, the time at which the backward simulations were run. This would capture most short recombination-coalescence loops, while still being well below the bulk of ancestral coalescence times. The effect of the non-detectable recombination events became significant only for very low neighborhood sizes (< 15) when the detected number of IBD blocks of a certain length was inflated by wrongly inferring multiple shorter blocks as a single longer block. While estimates for density remained almost unbiased, the inferred dispersal rates increased significantly, likely due to an excess of block sharing for distant samples. However, even for very low neighborhood sizes, when the effective density of individuals measured in dispersal units was about one (for neighborhood size $4\pi D_e \sigma^2 \approx 12.6$), the upward bias remained less than 50%.

2.2.3 Sampling Guidelines

Edge effects

In practice, populations do not extend infinitely beyond the sampling area, but have range boundaries. This forces lineages to deviate from the simple diffusion model, as they cannot wander out of the species range (Wilkins and Wakeley, 2002; Wilkins, 2004). This might be a common violation of our model assumptions. We assessed how much our inference method was affected using simulated data from habitats of limited size. In these simulations, we assumed that the lineages were reflected once they reached a range boundary.

Our results indicated that, in cases when the boundaries were close to the samples, such that the distance to the nearest samples was on the same order of magnitude as σ , the estimates for the dispersal rate σ and density D become biased downward (2.5.3), an effect also observed for the inference method of Novembre and Slatkin (2009) that is based on the sharing of rare alleles. Similarly, we observed that the estimates for D and σ become biased downward for habitats of width $\approx 10 \sigma$. Therefore, we recommend

to always check whether most of the samples are collected far from the habitat edges ($> \sigma$) and whether the habitat is sufficiently large (diameter $> 10 \sigma$).

For the special case of rectangular habitats with reflecting boundaries, the method of images described by Wilkins (2004) gives a simple way of calculating the coalescence probabilities for two spatially diffusing lineages (Eq. 2.6). In principle, it is straightforward to update our formulae for expected block sharing accordingly. One simply has to add terms describing the expected block sharing with ghost samples reflected at the edges. However, we did not implement this correction, as this approach cannot be extended to more irregularly shaped habitats and boundary edges, as usually encountered in reality.

Clumped sample distribution

In practice, samples are not always evenly distributed, but are often clumped due to sampling constraints. To investigate how such clustered sampling affects our inference scheme, we compared the results of various scenarios of clumping (2.5.3). The estimates and their inferred uncertainty were not affected substantially, only in the cases of very asymmetric clumping we observed a small upward bias of dispersal estimates. This overall robustness is not surprising, as the distribution of pairwise sampling distances is not changed much as long as the clumping is not overly pathogenic (i.e., a very low number of sample clusters).

2.2.4 Analysis of POPRES Data

Best fit models

When fitting our models to the Eastern European subset of the POPRES IBD block data, the model of quick population growth with a population density $D_e(t) = 1/t$ fit markedly better than a model of constant population size, which underestimated sharing of short blocks (Fig. 2.8) at the maximum likelihood parameters. In the more complex model, $D_e(t) = t^{-\beta}$, the growth rate parameter β was estimated to be close to 1. The increase of log-likelihood was small ($\Delta L = 1.1$), especially when considering that there are correlations in the data that make the difference of true likelihood even smaller (Coffman et al., 2016). Similarly, fitting several more complex density functions as sums of power terms did not significantly increase the likelihood. In all three models, the estimates for dispersal σ were about 60–70 km/ $\sqrt{\text{gen}}$, even under the likely misspecified constant population size model (Table 2.1), and bootstrapping on the country pair level yielded 95% confidence intervals that ranged from 45–80 km/ $\sqrt{\text{gen}}$.

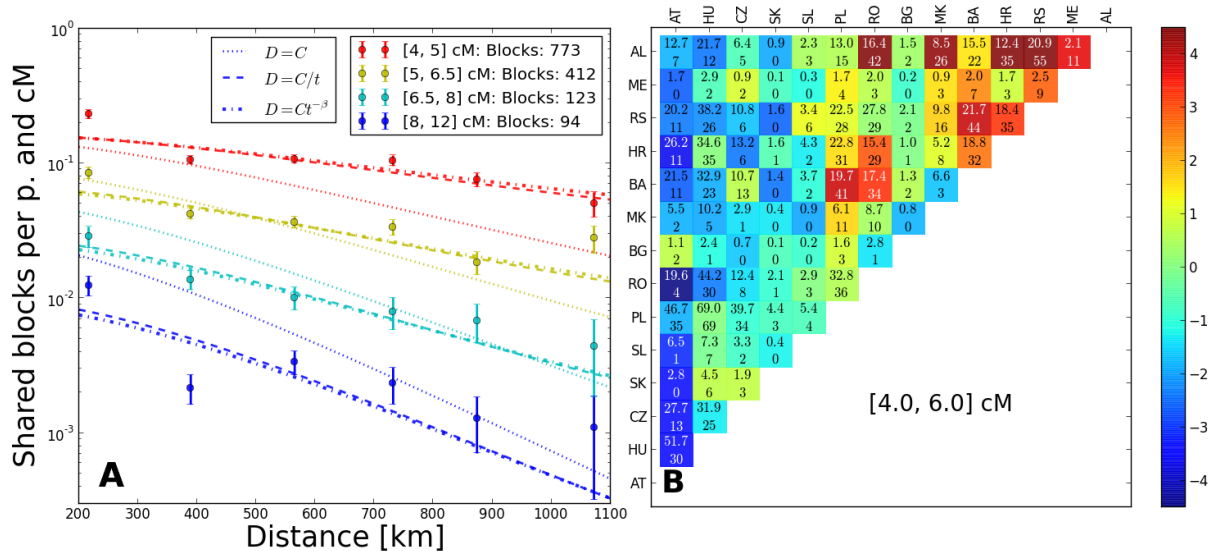


Figure 2.8: Fit of models to Eastern European block sharing data. (A) To better visualize the data, observed block sharing was binned into distance and block length bins. The dots depict the average block sharing within each bin and the lines are predictions from the best fit models. The error bars represent standard deviations under the assumption of Poisson counts in every bin; some are clearly too tight and there are outliers, which hints at more systematic deviations at the country-pair level (see also Fig. 2.5).

(B) Residuals for pairs of countries for blocks of length 4–6 cM: Upper line in every field: Total number of IBD blocks predicted by the best fit model. Lower line: Observed number of IBD blocks. Color of every field is determined by statistical significance (z -Value when transformed $x \rightarrow 2\sqrt{2x}$). Abbr.: AT: Austria, HU: Hungary, CZ: Czech Republic, SK: Slovakia, SL: Slovenia, PL: Poland, RO: Romania, BG: Bulgaria, MK: Macedonia, BA: Bosnia, HR: Croatia, RS: Serbia, ME: Montenegro, AL: Albania.

Table 2.1: Maximum Likelihood Estimates for Eastern European IBD data.

Density Model	Parameter	MLE-Estimate	95% CI	Emp. 95% CI
$D_e = D$	D	0.047	0.043–0.051	0.038–0.065
	σ	67.8	62.9–72.8	53.03–81.50
$D_e = D/t$	D	1.71	1.48–1.94	1.22–2.87
	σ	62.6	56.1–69.0	42.2–82.6
$D_e = Dt^{-\beta}$	D	2.13	1.39–2.86	1.16–5.83
	σ	63.0	56.2–69.8	44.2–82.4
	β	1.05	0.98–1.13	0.90–1.25

All units so that distances are measured in km and time in generations. CIs are based on Fisher Information matrix and empirical CIs are based on 100 estimates bootstrapped over country pairs.

The estimated parameter uncertainty when bootstrapping over single blocks was only slightly larger than was estimated from the curvature of the likelihood, but bootstrapping over country pairs gave markedly increased confidence intervals (Fig. S3), which implies that there are systematic correlations at this level in the data. This was further confirmed by the analysis of the residuals for the country pairs, which yielded a gradient toward the Balkans for more block sharing than predicted by the best fit models. The deviations were statistically most significant for short blocks because of the increased power due to the higher number of shared blocks (Fig. 2.5); however, the overall pattern also held for longer blocks (Fig. S4).

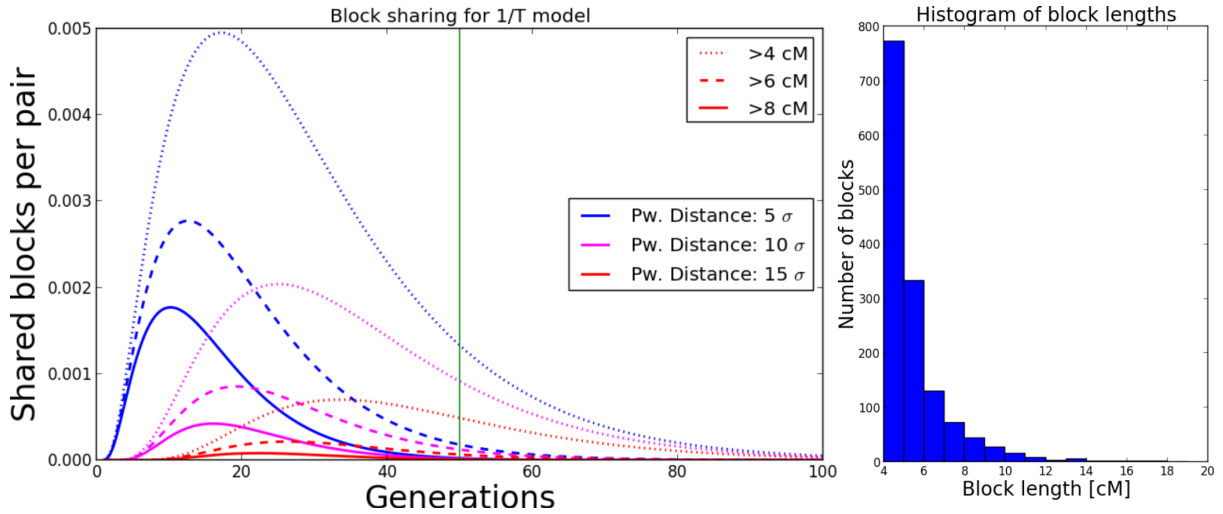


Figure 2.9: Left: Age of shared IBD blocks. Density of blocks of certain length originating t generations ago, as calculated from the $1/T$ population density growth model with best fit parameters. Most of the signal is predicted to have arisen within the last 50 generations (green line). Block sharing would have been more recent assuming a constant population density. Right: Distribution of block lengths used in our analysis of empirical human data.

2.3 Discussion

The main goal of this article was to develop a robust inference scheme for populations extended in two spatial dimensions that utilizes pairwise shared long IBD blocks to reliably estimate the dispersal rate σ and the effective population density D_e separately. For this, we derived analytical formulae for block sharing under an model of diffusion of ancestry that extended the previous work of Barton et al. (2013), and fit these results by maximizing a composite likelihood similar to that used by Ralph and Coop (2013). Using extensive tests on data simulated under a variety of scenarios, we demonstrated

that our method could robustly perform this task.

Baharian et al. (2016) recently independently arrived at similar formulae for block sharing under a model of spatial diffusion, which they fit by regressing block sharing binned according to pairwise geographic distance and block length. Our work is conceptually similar, but provides several important extensions. We additionally described the effect of recent population density changes, which seems especially relevant for human populations. For the special case of a constant population size, our results matched an equivalent result of Baharian et al. (2016), and our approach allowed us to additionally incorporate chromosomal edge effects. Moreover, our likelihood framework offers several advantages over regressing binned data because it makes use of the information contained in the lengths of shared blocks. It can also be used to quantify the uncertainty of the parameter estimates, and can readily include error estimates for block detection. Another major contribution is our extensive testing on simulated data. These simulations yielded insights into the general accuracy of the underlying idealized model assumptions. They also helped us to investigate how several deviations from the model that might occur in practice, including ineffective recombination events, wrongly specified growth models, irregular sample distribution and nearby habitat boundaries, affect inference.

Exponential decay of IBD block sharing with distance and block length

The derived formulae for sharing long IBD blocks under diffusion of ancestry are structurally similar to the Wright-Malecot formula (Barton et al., 2013) that describes allelic identity by state using similar approximations. A polynomial factor is multiplied with a Bessel function of the second kind, $K_\gamma(\sqrt{2\lambda}\frac{r}{\sigma})$; for allelic correlation $\lambda = \mu$, the mutation rate, while here $\lambda = L$, the IBD block map length. For long blocks, L is much larger than typical mutation rates μ . This allows us to probe the tail of the Bessel function ($\sqrt{2\lambda}\frac{r}{\sigma} > 1$) where it approaches exponential decay that dominates the polynomial factor. This exponential decay occurs with both an increasing block length \sqrt{L} and an increasing geographic distance r . The theory predicts that for long blocks this decay can be over orders of magnitude when pairwise geographic sample distance increases multiple dispersal distances (Fig. 2.3), which is observed in the human data (Fig. 2.8). This pattern persists even in the case of recent population density changes (Fig. 2.4). When global density changes can be modeled as the sums of power terms of the form Eq. 2.8, the result for expected block sharing will be given by the sums of the corresponding Bessel functions (Eq. 2.9). Each of those approaches exponential decay with rate $\sqrt{2L}/\sigma$; thus, also their sum does. Therefore, estimates of the dispersal rate σ that

use the decay rate in the exponential regime can be expected to be robust with respect to recent demographic history.

2.3.1 Implications for Demographic Inference

The fast rate at which long blocks are broken up and the ability to probe the exponential regime of decay offer several significant advantages for demographic inference, which our inference scheme can utilize. First, long blocks typically stem from very recent times (Fig. 2.9). This is clearly advantageous for populations that have been in equilibrium for only a relatively short time, as is likely often the case. Inference methods that rely on allelic correlations probe recent timescales as well (Barton et al., 2013). They similarly pick up locally increased identity by state by recent coancestry. However, this is often only a small signal on top of a majority of identity by state stemming from ancient times. Thus, these methods are much more susceptible to confounding by ancestral structure (Meirmans, 2012), and have stringent, often unrealistic, equilibrium time requirements (Leblois et al., 2003), which our method can avoid. For instance, in the human case, the best fit model predicts that most long blocks stem from within the last 50 generations (Fig 2.9). Second, quick exponential decay, both with sampling distance and block length, offers a very robust signal for demographic inference. As demonstrated, the expected number of blocks that are multiple cM long decays by orders of magnitude over a geographical scale of several dispersal distance units. This pattern should be relatively robust with respect to small-scale heterogeneities of habitat or dispersal. Such quick decay also aids robust inference, as shown by the accuracy of the inference method on simulated data. This is in contrast to inference that is based on classic measures of pairwise genetic similarity. Such measures usually only decay with the logarithm of distance (Barton et al., 2002), which causes low and often problematic signal to noise ratios (Watts et al., 2007). Third, utilizing the logarithmic regime of the Wright-Malecot formula only allows one to infer the neighborhood size proportional to the product of density and dispersal. Naturally, however, their separate values are of interest. As demonstrated, inference based on long IBD blocks allows one to obtain robust separate estimates of these two important demographic parameters.

2.3.2 Analysis of Human Data

The analysis of human data nicely demonstrates our inference scheme. The true demographic scenario is doubtless more complex, including heterogeneous, time-dependent migration rates, and large-scale migrations. However, qualitatively the patterns of IBD

block sharing appear to fit well with our diffusion model. Despite several significant deviations, the best fit model explains the overall broad trends in the empirical data (Fig. 2.8), such as the decay of the number of shared blocks with both increasing geographic distance and block length. Using our inferred model, we predicted most of the shared blocks we used (> 4 cM) and hence, our signal originates within the last 50 generations (Fig. 2.9), which corresponds to the past 1450 years (assuming 29 years per generation (Fenner, 2005)). This mostly postdates the period of large-scale migrations in Europe (“Völkerwanderung” (Davies, 2014)). Our inferred demographic parameters seem to be plausible. There is a clear signal for rapidly accelerating recent population growth, which is in agreement with historical estimates (McEvedy et al., 1978) and previous genetic studies based on the allele frequency spectrum (Keinan and Clark, 2012; Gao and Keinan, 2016). Historical dispersal estimates infer values of typical migration distances per generation ranging from a few to several dozen kilometers (Wijsman and Cavalli-Sforza, 1984; Pooley and Turnbull, 2005). While agreeing on orders of magnitude, these are somewhat lower than our estimates ($\sigma \approx 50 - 100\text{km}/\sqrt{\text{gen}}$). However, there is also evidence that preindustrial individual human migrations over large distances are rare, but occur at a significant rate (Pooley and Turnbull, 2005).

We detected a systematic, large-scale deviation from a simple diffusion model with uniform population density, as there is a clear gradient for higher block sharing in the direction of the Balkan countries (Fig. 2.5). This was already observed by Ralph and Coop (2013). They hypothesized that this could be due to the historic Slavic expansion, a hypothesis supported by admixture analysis (Hellenthal et al., 2014). However, the pattern of increased block sharing also holds for longer, typically younger blocks, which could hint additionally at a consistently lower population density in these regions. Such systematic regional deviations from the diffusion model also imply that care should be taken when estimating parameters and their uncertainty ranges.

2.3.3 Outlook

Our inference scheme based on long IBD blocks requires large amounts of data, as it needs dense genotype data from a few dozen individuals, spatial information of the samples, and a linkage map. However, the novel opportunities and advantages for inference of recent demography should justify the effort. The possibility to accurately estimate dispersal distances and past effective population densities could yield interesting novel insights for a whole range of organisms. The necessary datasets are within reach for several systems, and they will become even more accessible in the near future with increasing genotyping capacities.

A salient extension of our model would be to address complications such as anisotropy (Jay et al., 2013) and large-scale heterogeneities in migration patterns or population densities across the landscape. For classic measures of genetic similarity, elaborate computational techniques have been recently applied for inference within such complex demographic scenarios (Duforet-Frebourg and Blum, 2014; Petkova et al., 2015). As argued above, analysis of IBD blocks would be even more suited to this task, as the length of shared blocks gives additional information. However, analytical solutions, which hugely facilitate inference, are likely no longer feasible. Inference will have to be based on numerical predictions, although utilizing block sharing of different lengths will be even more computationally intensive than extracting information from a single genetic similarity matrix. Consequently, this challenge is beyond the scope of this paper. We hope that future work will help to fully utilize the potential of shared IBD blocks, and that our inference scheme marks only one step in a new era of demographic inference.

2.4 Appendix

2.4.1 Appendix A: Effective Density

We use diffusion to model the separation of two lineages backward in time. Let $r(\mathbf{x}, t)$ denote the probability density of the vector \mathbf{x} of pairwise distances along each axis at time t back. In our model two lineages coalesce instantaneously with a coalescence rate $\nu(\mathbf{x})$ that depends on \mathbf{x} . For the probability of coalescing at time t ago, we get:

$$\psi(t) = \int_{\mathbb{R}^2} r(\mathbf{x}, t) \nu(\mathbf{x}) \, d\mathbf{x}.$$

In cases where only discrete sample distances \mathbf{x} are possible, such as the stepping stone model, the integral has to be replaced with a sum. The key observation is that $\nu(\mathbf{x})$ is usually negligible outside a small area around the origin, since in most models only very close samples ($|\mathbf{x}| \approx \sigma$) have an appreciable chance to coalesce. Within such small areas around the origin, for $t \gg 1$ we approximate $r(\mathbf{x}, t)$ with $\approx r(0, t)$ and get:

$$\psi(t) \approx r(0, t) \int_{\mathbb{R}^2} \nu(\mathbf{x}) \, d\mathbf{x} = r(0, t) \frac{1}{2D_e}, \quad (2.12)$$

where we have defined $1/(2D_e) := \int_{\mathbb{R}^2} \nu(\mathbf{x}) \, d\mathbf{x}$. It can be shown that stepping stone models asymptotically converge to this model when rescaling appropriately (Barton et al., 2002, 2013). With demes separated by one distance unit D_e corresponds to the number of diploid individuals per deme, which motivates the name effective density.

Here we give this more general definition of D_e to allow one to directly calculate its value in various scenarios we simulated above (2.5.2).

2.4.2 Appendix B: Chromosomal Edge Effects

Here, we give the full result for block sharing that includes chromosomal edge effects, which we use for inference. We shall denote the formula Eq. 2.9 with fixed $G = 1M$ with $n_L(\beta)$, where the dependencies other than β are suppressed for ease of notation. Then, integrating Eq. 2.4 yields:

$$\mathbb{E}[N_L] = (G - L)n_L(\beta) + n_L(\beta - 1),$$

the formula for one chromosome of length G . For multiple chromosomes of different lengths one has to sum this formula over all chromosomes. For pairs of diploid individuals, the resulting formula has to be also multiplied by a factor of four, since for every pair of individuals four pairs of chromosomes are compared.

2.4.3 Appendix C: Likelihood

Using the Poisson approximation (Eq. 2.11), the likelihood of a pair of samples (j) at distance r sharing blocks of length $\vec{L} = L_1, \dots, L_n$ that fall into a set of length bins i_1, \dots, i_n is given by:

$$\tilde{f}_j = \Pr(\vec{L} | r, \theta) = C \lambda_{i_1} \lambda_{i_2} \cdots \lambda_{i_n} \exp(-\sum_i \lambda_i), \quad (2.13)$$

where C absorbs all constants that do not depend on the model parameters θ . This constant can be dropped when doing likelihood based analysis. Continuing to assume independence, we take the product over all pairwise likelihoods \tilde{f}_j to get the total composite likelihood:

$$\tilde{f}(\vec{L}, \theta) = \prod_{\text{Pairs } j} \tilde{f}_j(\vec{L}^j, r_j, \theta),$$

where \vec{L}^j denotes the shared blocks of the j th pair.

The number of pairs $\frac{n(n-1)}{2}$ increases quadratically with sample size n . This scaling is advantageous for an inference scheme, but implies that the runtime also grows with the square of sample size. However, algorithms to maximize functions with a low number of parameters are very efficient, so even sample sizes of hundreds of individuals can be easily handled. Calculation can be also sped up by grouping pairs with the same pairwise distance. For instance, when analyzing multiple individuals from

a population with the same spatial coordinates the λ_i do not have to be calculated repeatedly for every individual pair. Denoting the length bins of blocks shared over all pairs by i_1, \dots, i_n and the number of pairwise comparisons by k yields:

$$Pr(\vec{L} | r, \theta, k) = Ck^n \lambda_{i_1} \lambda_{i_2} \cdots \lambda_{i_n} \exp(-k \sum_i \lambda_i),$$

where the factor Ck^n does not depend on the model parameters and can be dropped when maximizing the likelihood.

Appendix D: Block Detection Errors

The probability density $\tilde{\lambda}(y)$ of actually observing a pairwise shared block of length y can be calculated from the probability $\lambda(x)$ of sharing true blocks of length x :

$$\tilde{\lambda}(y) = f(y) + \int_0^G R(y, x) c(x) \lambda(x) dx, \quad (2.14)$$

where $f(y)$ describes the false discovery rate function depending on block length y , $c(x)$ the power to detect a block of length x and $R(y, x)$ the probability of detecting a block of true length x as block of length y . Doing a careful analysis using techniques such as manually inserting shared blocks and rerunning the IBD block detection allows one to estimate these error functions (Ralph and Coop, 2013).

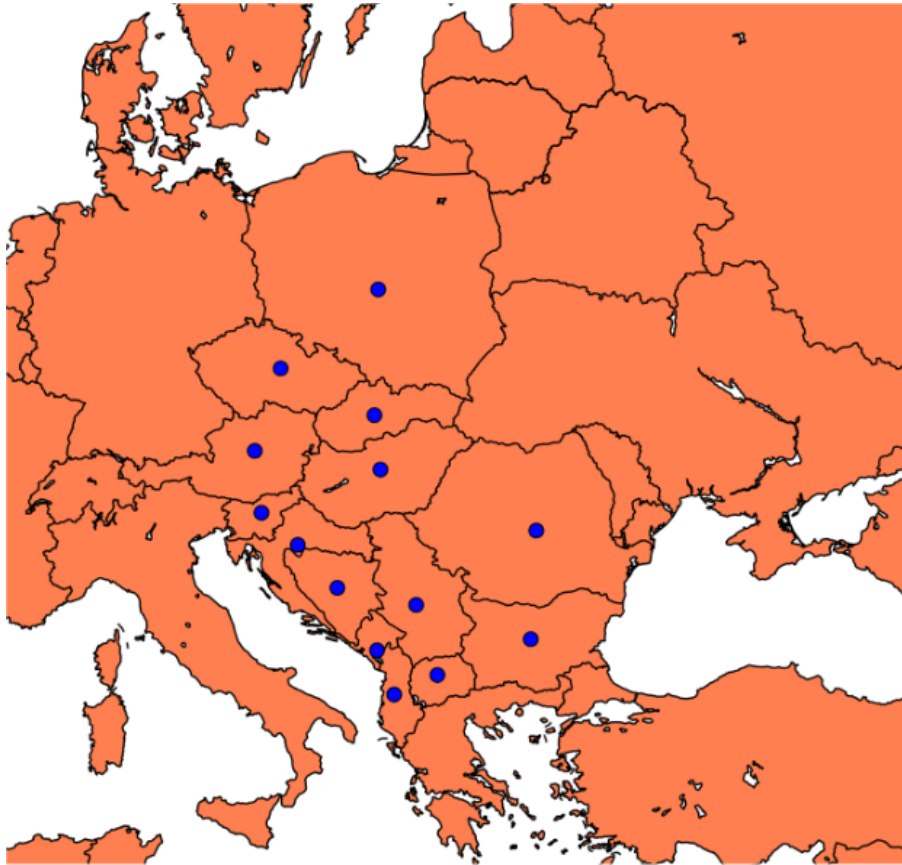
This error model is straightforwardly included into the framework of working with small length bins. First, for every block length bin of a pair of samples first the predicted true sharing λ is calculated for a set of demographic parameters θ . Second, the predictions are updated according to Eq. 2.14 with the detection error estimates to get the predicted rates $\tilde{\lambda}$ when accounting for errors. Third, the likelihood of observed block sharing are computed as before.

2.5 Supplementary Material

2.5.1 Supplementary Information 1: POPRES analysis

Location data for every country was downloaded from http://cs.baylor.edu/~hamerly/software/europe_population_weighted_centers.txt. Montenegro was not listed, so coordinates were set to its capital Podgorica. This should not be problematic as only one sample from Montenegro entered the analysis.

Table 2.2: Location and number of samples used for inference (compare (Ralph and Coop, 2013)).



Country	Nr individuals
Austria	14
Hungary	19
Czech Republic	9
Slovakia	1
Slovenia	2
Poland	22
Romania	14
Bulgaria	1
Macedonia	4
Bosnia	9
Croatia	9
Serbia	11
Montenegro	1
Albania	9
Total	125

2.5.2 Supplementary Information 2: Simulation Details

Parameters values

Given the formula $1/(2D_e) := \int_{\mathbb{R}^2} \nu(\mathbf{x}) \, d\mathbf{x}$ (Appendix A) for the effective density, it is straightforward to calculate the effective density D_e for the models used here to simulate IBD block sharing.

First, in the case of the described grid model, where every node occupied by a single diploid individual is separated by one distance unit, $\nu(\mathbf{x})$ is simply the chance of changing pairwise distance by $-x$ multiplied with a factor $1/2$. Summing $2 \sum \nu(\mathbf{x})$ over all pairwise distances therefore is simply the sum over all probabilities of moving distance x . This trivially sums to 1, independent of the exact shape of the dispersal kernel. Therefore, the effective density is given by $D_e = 1$. When modeling multiple individual per node, such as in a stepping stone model, an analogous calculation shows that D_e is given by the number of diploid individuals per node.

For the DISCSIM model, two lineages coalesce when they are hit by the same event and both lineages jump to the same ancestral chromosome. Therefore, $\nu(\mathbf{x})$ is the integrated rate of all those events affecting two individuals at distance x . Exchange the order of integrals in the full integral then yields $\nu(\mathbf{x}) = \frac{1}{2}u^2R^4\pi^2$. Normalizing with the rate by which an individual gets hit by an event gives the rate in generation times: $\frac{1}{2}R^2\pi u$ and therefore $D_e = \frac{1}{R^2\pi u}$.

The dispersal rate σ^2 in the deme and grid model is simply given by the axial variance of the single generation dispersal kernel. In the DISCSIM model, in every migration event a lineage jumps from a random point within a circle of radius R to another random point, the average squared distance of such jumps can be calculated to be R^2 . This is two times the axial variance, and thus $\sigma^2 = \frac{R^2}{2}$.

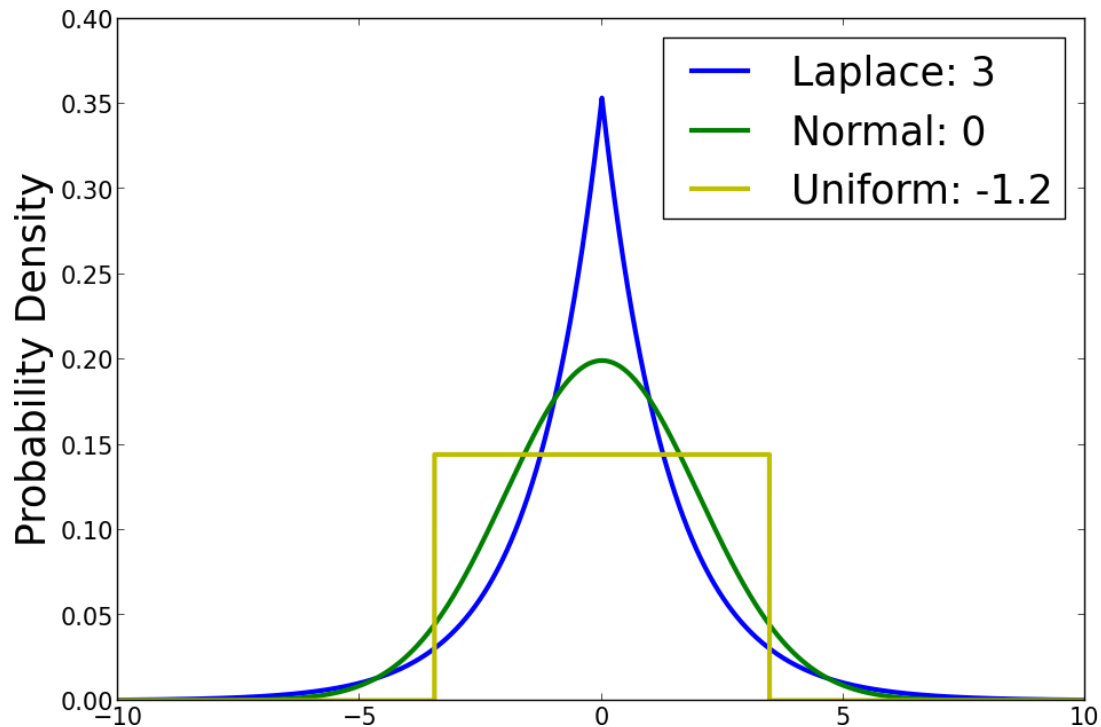
Simulation parameters

Here we give the simulation parameters used for generating the pictures above. Since we used discretized dispersal kernels for the simulations on grids, throughout the paper we always adjusted the parameters to draw from the continuous dispersal kernels (Laplace, Normal, Uniform) such that the empirical single generation axial σ^2 matched the target value on a simulated very large number of discretized random draws.

For the simulations used to generate Fig. 3, we used the following parameters:

Parameter	Value
Torus Size (Axis)	180
Sample Distance (Axis)	2
Chromosome length	150 cM
Locus Nr. (DISCSIM)	1500
σ (axial, per generation)	2
u (DISCSIM)	$\frac{1}{8\pi}$
R (DISCSIM)	$\sqrt{2} \cdot 2$
Deme Size (Axis, deme model)	5
t_{Max} (in generations)	200

Visualisation of axial dispersal kernels with $\sigma^2 = 2$, value on right top gives excess kurtosis:



Comparison Confidence Intervals

Here we compare the empirical 95% confidence interval from Fig. 5 to the mean length of the estimated confidence intervals, which shows that the estimates from the curvature of the likelihood surface capture parameter uncertainties rather well. This indicates that IBD blocks originate from mostly independent events in the scenarios simulated here. We also give the correlation of σ and D based on 100 replicate runs for each sample size.

Sample Size	σ		D_e		Correlation (MLE-estimates)
	Emp. CI	Est. CI	Emp. CI	Est. CI	
100	0.967	1.12	1.24	1.07	0.1
270	0.304	0.390	0.350	0.332	-0.16
440	0.258	0.239	0.219	0.200	- 0.10
625	0.209	0.166	0.186	0.140	-0.20

2.5.3 Supplementary Information 3: Simulations

Clumping of individuals

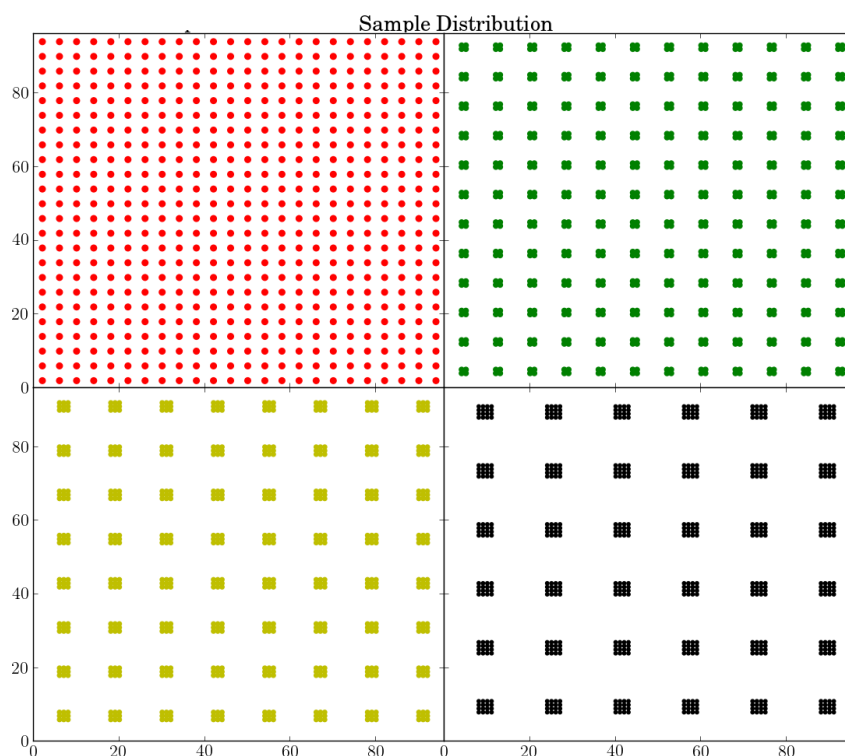
To assess how much our inference method is affected by clumping, we simulated four scenarios of two different kinds of clumping. Simulations were always done on a torus grid with axial length 96, with Laplace dispersal along each axis, dispersal distance set such that $\sigma = 1$ and assuming a constant population size.

Regular Clumping

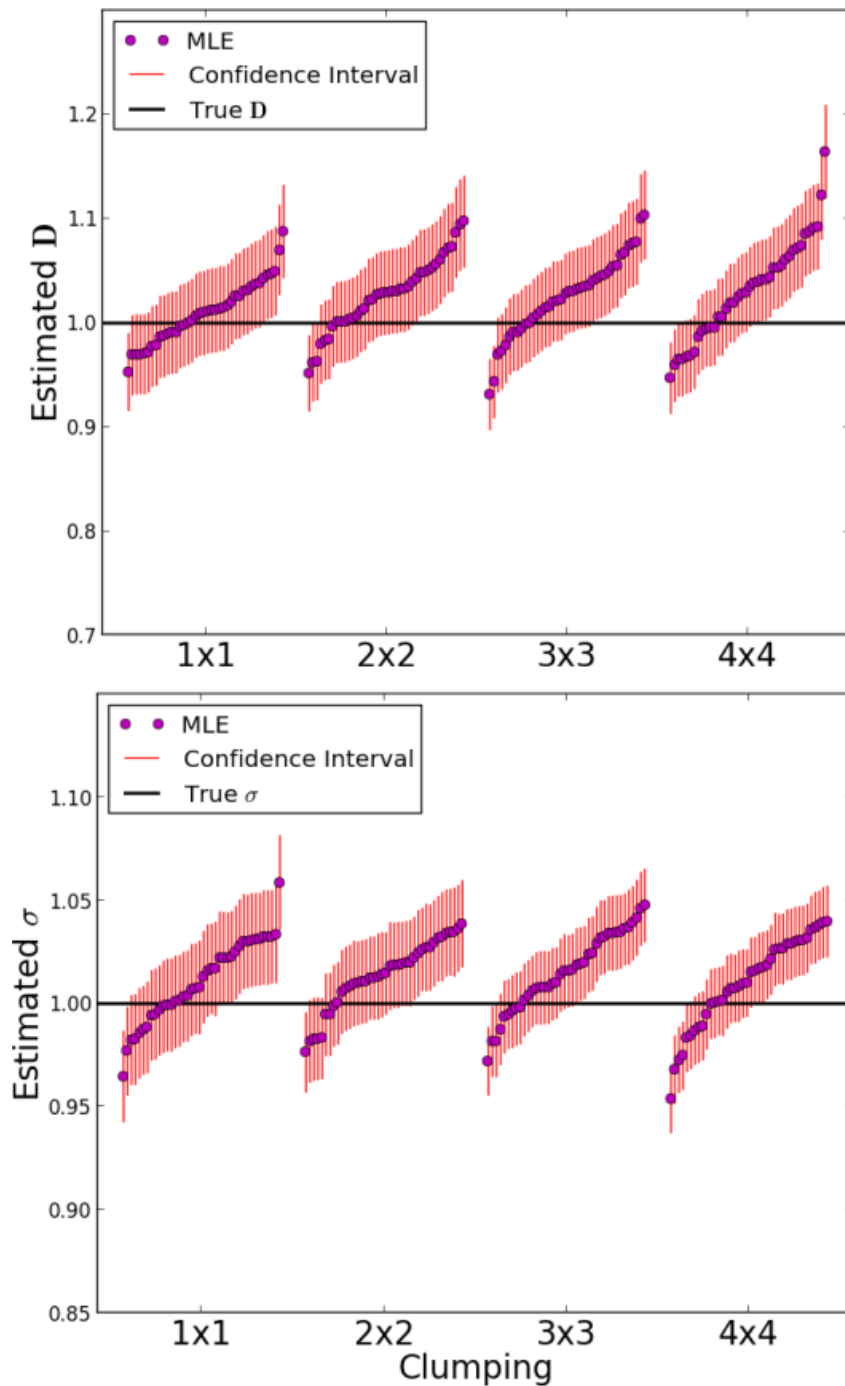
First, in four scenarios, a total number of 576 samples (150 cM chromosomes) were clustered on regular grids:

- Scenario 1: Samples are evenly spaced 4 distance units apart.
- Scenario 2: Samples are clustered into 2×2 clusters 8 distance units apart (along each axis).
- Scenario 3: Samples are clustered into 3×3 clusters 12 distance units apart (along each axis).
- Scenario 4: Samples are clustered into 4×4 clusters 16 distance units apart (along each axis).

This sampling scheme is visualized in the following picture:



We simulated block sharing until $t = 200$ generations back. Our inference scheme applied to blocks 4–20 cM simulated under 20 replicates of each of these scenarios yielded the following parameter estimates and 95% confidence intervals:



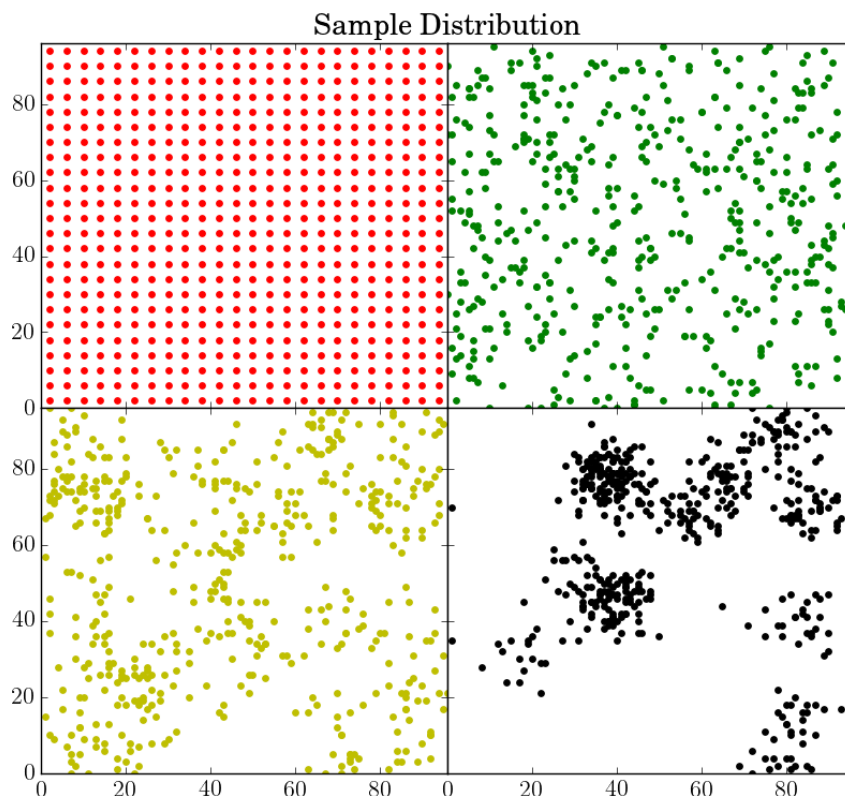
The inference method was very robust with respect to clumping, neither the bias nor the uncertainty of estimates differed substantially between the sampling scenarios considered here.

Irregular Clumping

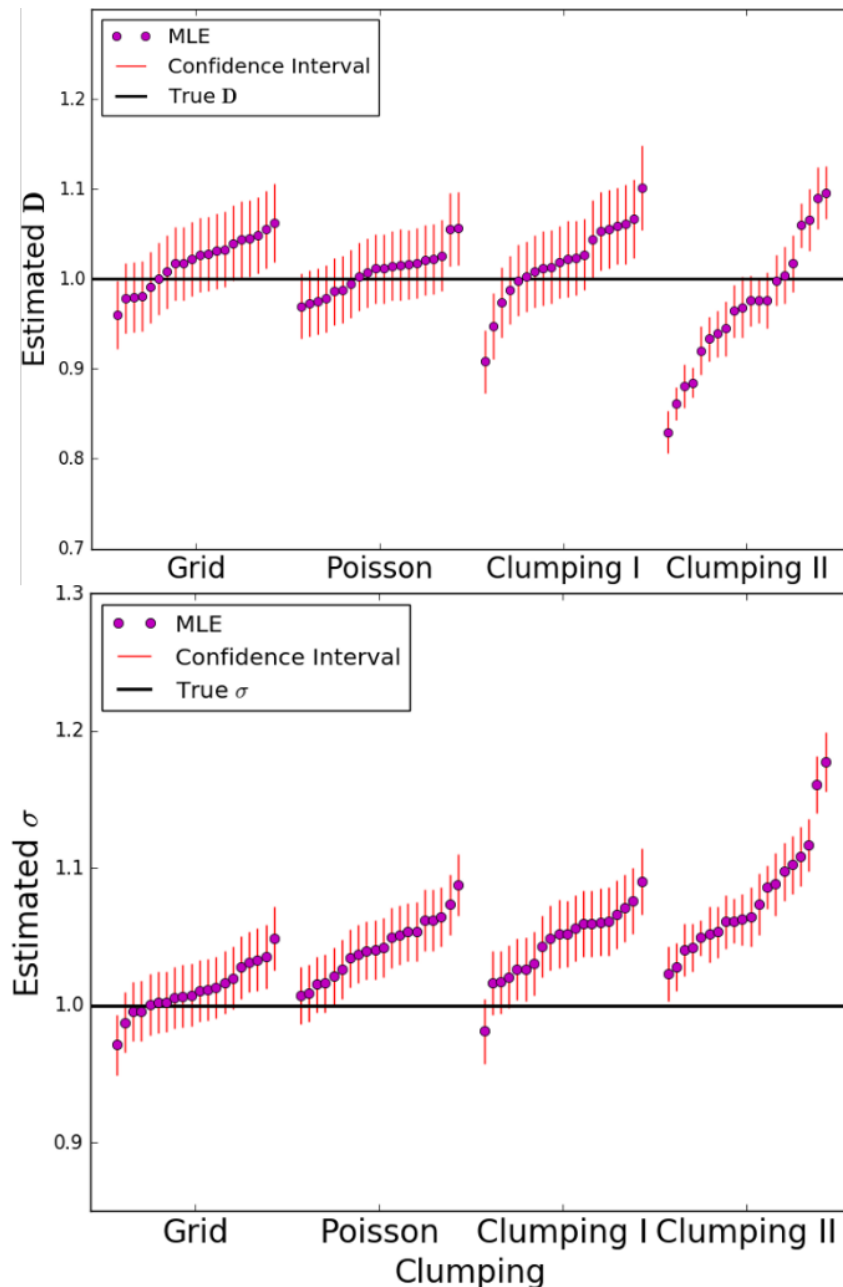
We then also assessed the effect of more irregular and asymmetric clumping. In four scenarios, a total number of 576 samples (150 cM chromosomes) were clustered to different degrees:

- Scenario 1: Samples are evenly spaced 4 distance units apart on a regular grid.
- Scenario 2: Every sample is independently drawn from all grid positions.
- Scenario 3: Randomly clustered samples are simulated in three steps. First, centers of clusters are drawn at random. Then, for every center a random number is drawn from a geometric distribution with mean 5 to determine the number of samples per cluster. Last, for each individual sample, the offset by the mean is determined by a discretized Gaussian with standard deviation $\sigma = 5$ along each axis. Individual samples are drawn until the total number of 576 samples is reached.
- Scenario 4: Same as in Scenario 3, but now the number of samples in every cluster is geometrically distributed with mean 50.

Single realizations of the random clustering schemes described above are visualized in the following picture:



For each scenario, we did 20 replicates. In Scenario 2 – 4, the sample positions were generated independently for every replicate, and in each case we simulated block sharing until $t = 200$ generations back. Our inference scheme applied to blocks of length 4–20 cM yielded the following parameter estimates and 95% confidence intervals:



Overall, irregular clumping did not severely affect estimates. However, dispersal estimates became slightly biased upwards with increasing degree of clumping. This weak effect is likely due to clusters of individuals on very small spatial scales; since for geographically close samples the diffusion approximation is not expected to accurately predict leptokurtic single generation dispersal kernel. This hypothesis is supported by the fact that the bias vanishes if spatial size of the sample clusters is increased (data not

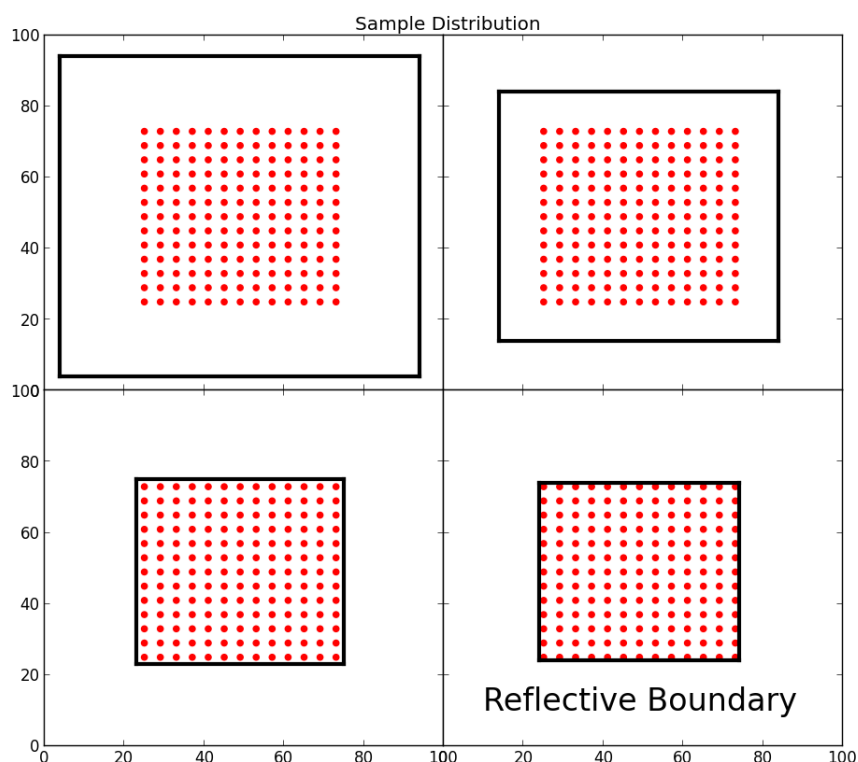
shown).

Edge effects

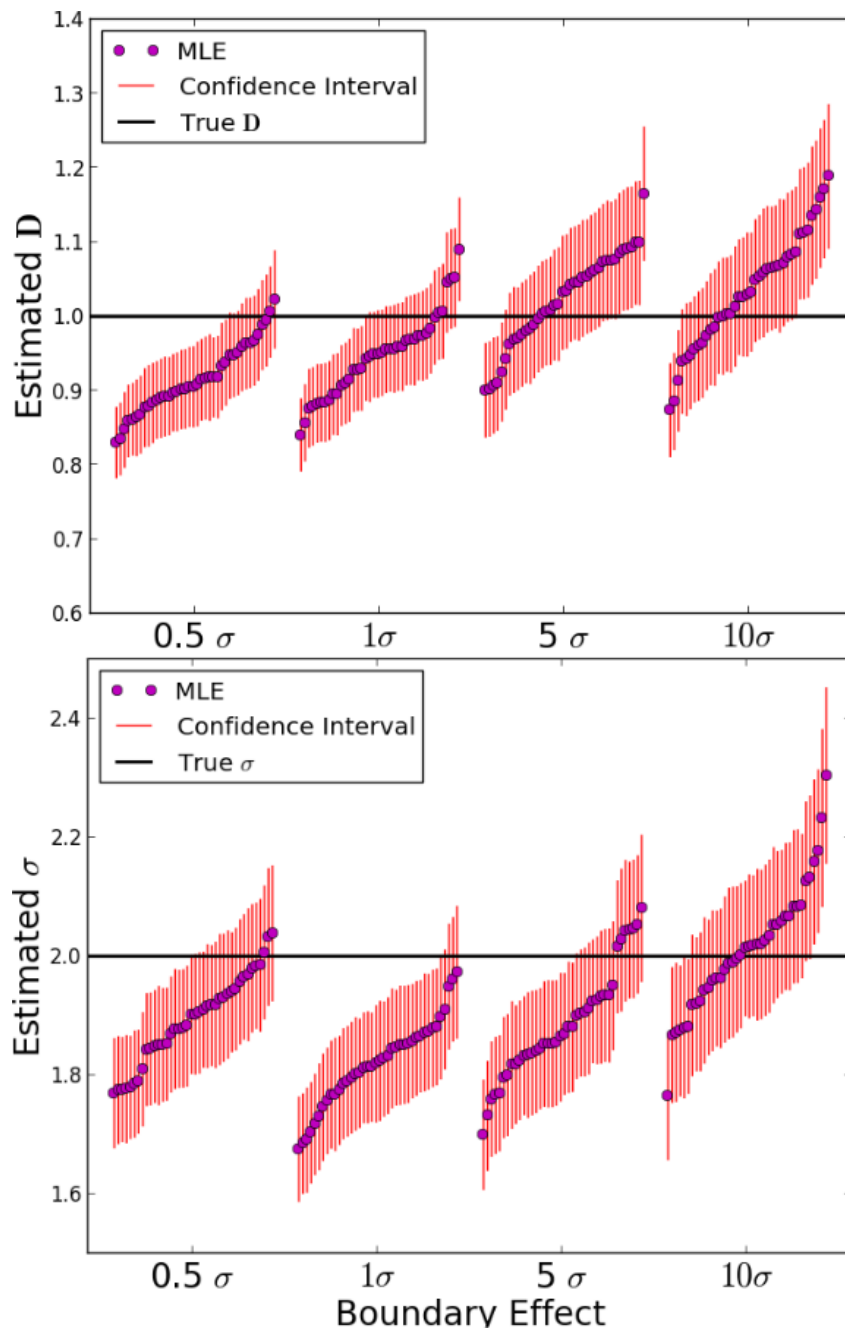
To assess the effect of nearby habitat edges on our inference method, we simulated finite rectangular habitats. For simulations, we utilized our node model, as above with axial Laplace dispersal such that $\sigma = 2$. We assumed that lineages get reflected each time they would trace back beyond an edge. That is, if in a single generation back a lineage would have migrated a certain distance beyond the habitat boundary, that lineage moves to a location an equally far distance from the boundary, but inside the habitat. We simulated four scenarios, where reflective boundaries were surrounding a 12×12 array of samples (150 cM chromosomes) spaced two σ apart.

1. The reflective boundary was at a distance 10σ .
2. The reflective boundary was at a distance 5σ .
3. The reflective boundary was at a distance 1σ .
4. The reflective boundary was at a distance 0.5σ .

This sampling scheme is illustrated in the following figure:



We simulated block sharing until $t = 200$ generations back. Our inference scheme applied to blocks 4–20 cM simulated under 20 replicates of each of these scenarios yielded the following parameter estimates and 95% confidence intervals:



Both dispersal and density estimates got noticeably biased downwards as soon as range boundaries got closer, but remained right on orders of magnitude.

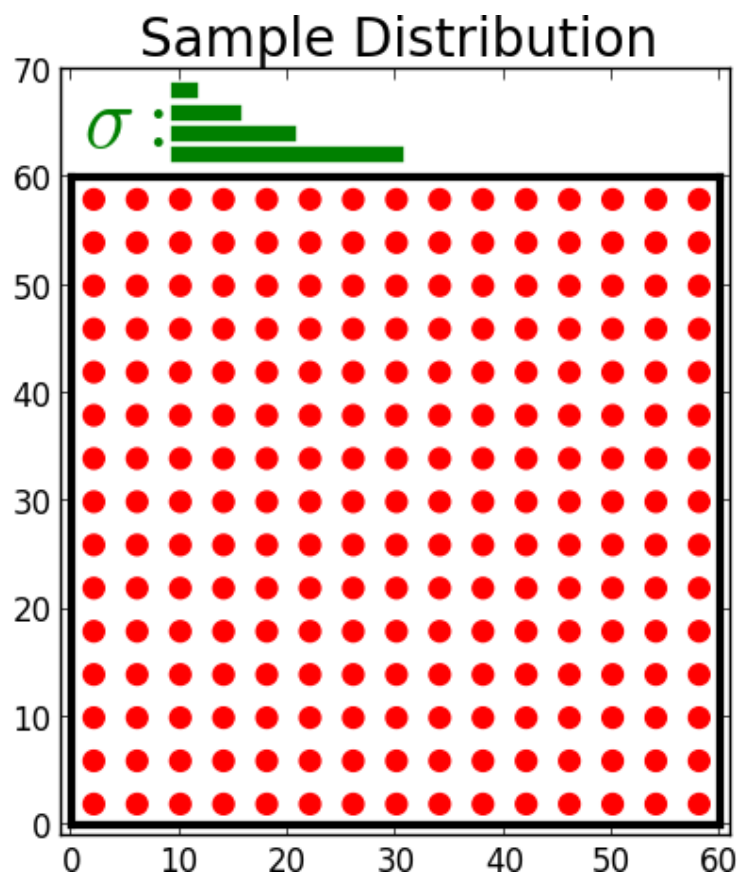
Limited habitat size

To assess the effect of limited habitat size compared to dispersal width σ , we simulated a finite rectangular habitat. Similarly, we utilized our grid model with reflective

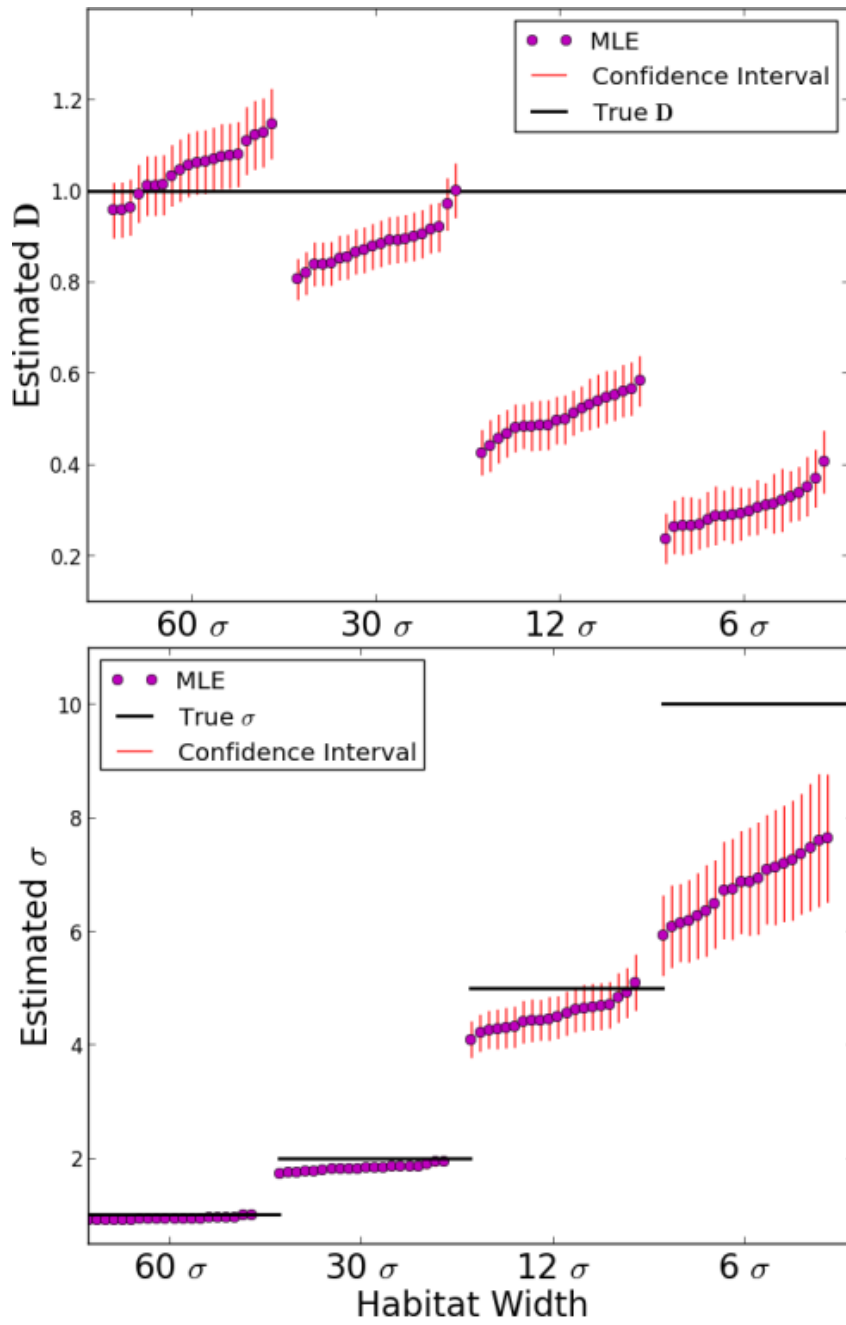
boundaries, and a Laplace dispersal kernel along each axis. We simulated four scenarios, where the samples were distributed on a fixed 15×15 grid two distance units apart from each other within a total habitat of axial width 60. Dispersal was set to various values:

- Scenario 1: $\sigma = 1$ (habitat width is 60σ).
- Scenario 2: $\sigma = 2$ (habitat width is 30σ).
- Scenario 3: $\sigma = 5$ (habitat width is 12σ).
- Scenario 4: $\sigma = 10$ (habitat width is 6σ).


These sampling schemes and width of σ are illustrated in the following figure:



We simulated block sharing until $t = 200$ generations back. Our inference scheme applied to blocks 4–20 cM simulated under 20 replicates of each of these scenarios yielded the following parameter estimates and 95% confidence intervals:



For habitats of width $\approx 10\sigma$ and smaller the inference method gave markedly downward biased estimates for σ and D_e , with particularly large bias for D_e .



3. Estimating barriers to gene flow from distorted isolation by distance patterns

Abstract. In continuous populations with local migration, nearby pairs of individuals have on average more similar genotypes than geographically well separated pairs. A barrier to gene flow distorts this classical pattern of isolation by distance. Genetic similarity is decreased for sample pairs on different sides of the barrier and increased for pairs on the same side near the barrier. Here, we introduce an inference scheme that utilizes this signal to detect and estimate the strength of a linear barrier to gene flow in two dimensions. We use a diffusion approximation to model the effects of a barrier on the geographical spread of ancestry backwards in time. This approach allows us to calculate the chance of recent coalescence and probability of identity by descent. We introduce an inference scheme that fits these theoretical results to the geographical covariance structure of biallelic genetic markers. It can estimate the strength of the barrier as well as several demographic parameters. We investigate the power of our inference scheme to detect barriers by applying it to a wide range of simulated data. We also showcase an example application to a *Antirrhinum majus* (snapdragon) flower color hybrid zone, where we do not detect any signal of a strong genome wide barrier to gene flow.

MANY populations are distributed across geographically extended habitats that are sometimes interrupted by barriers to gene flow. They can arise due to physical obstacles that reduce migration, but can also be caused by genetic incompatibilities, which reduce gene flow across a hybrid zone (Barton, 1979). Barriers can prevent locally adapted populations from being swamped by dispersal, and they can facilitate divergence, ultimately leading to

speciation. Therefore, they play a central role not only in conservation, but also in evolutionary biology and ecology. As direct observations of individual movement and reproduction are time-consuming and expensive, and moreover can only give a snapshot in time, there is much interest in indirect methods that infer such barriers from observed geographic genetic structure.

Such methods to detect barriers from genetic data can be grouped into two distinct approaches (Guillot et al., 2009): Clustering methods, that detect geographic genetic discontinuities between populations by grouping individuals into population units based on genetic similarity (Falush et al., 2003; Guillot et al., 2005; Dupanloup et al., 2002), and edge detection methods, that identify areas of sharp genetic change (Womble, 1951; Cercueil et al., 2007; Manni et al., 2004). None of these approaches is directly linked to any spatial population genetic model. They can therefore infer the existence of a barrier, but cannot give meaningful and biologically interpretable estimates of its strength. In addition, these approaches are often confounded by isolation by distance patterns (Safner et al., 2011; Meirmans, 2012), whereby individuals nearby are more similar than distant individuals (Wright, 1943) due to recent co-ancestry. While the description of this effect has a long history in theoretical population genetics of homogeneous populations (Malécot, 1948; Slatkin, 1993; Rousset, 1997; Hardy and Veke-mans, 1999; Barton et al., 2002), it has not been included into a practically applicable method to estimate the strength of a barrier to gene flow.

Here, we fill this gap, and introduce a method that infers the strength of a barrier in a two-dimensional population by fitting a population genetic model. Our method utilizes the fact that a barrier to gene flow distorts classical isolation by distance patterns (Fig. 3.1). Based on theoretical work of (Nagylaki, 1988), Barton (2008) constructed a theoretical framework. He showed that in two spatial dimensions, where fluctuations of allele frequencies are more localized than in one dimension, these effects of a barrier on allele frequency fluctuations can be significant already for intermediate barrier strengths. This signal therefore holds big potential for demographic inference. The derivation of Barton (2008) also shows that the effect of a barrier depends primarily on short-lived, localized fluctuations. In general, isolation by distance patterns equilibrate relatively quickly and depend mostly on recent demography (Barton et al., 2013; Aguillon et al., 2017). Therefore, an inference scheme based on distorted isolation by distance patterns infers contemporary barriers to gene flow, and should be robust to confounding effects of ancestral structure.

Here, we first expand previous theoretical results that describe the effect of a barrier on classical isolation by distance patterns (Nagylaki, 1988; Barton, 2008). We introduce a model where ancestry diffuses backwards in time and is partially reflected by a barrier.

This allows us to numerically calculate the probability of recent co-ancestry, which can then be fitted to genetic data. As single nucleotide polymorphism (SNP) datasets are currently widely used, we develop and implement ways to fit such biallelic genetic markers.

We test our inference scheme on synthetic data simulated under an explicit population genetics model and investigate how it is affected by confounding factors, for instance in a scenario of secondary contact. We also show a practical application, in which we apply our inference scheme to a hybrid zone population of *Antirrhinum majus*, in which a sharp transition in flower color and a causal flower color gene occurs (Whibley et al., 2006). We apply our method to test whether there is also a genome wide barrier to contemporary gene flow. To this end, we analyze a dataset of 12389 individuals and 60 suitable SNP markers.

3.1 Materials and Methods

We first outline the model underlying our inference scheme and discuss its assumptions, and then describe how our method fits this model to observed genotype data. In brief, we use a diffusion approximation for the spread of ancestry to calculate the probability of recent identity by descent between pairs of samples. We then fit our model to data by finding the demographic parameters that maximize the fit of observed homozygosity between all sample pairs.

3.1.1 Model

No model can capture all complexities of the real demographic history of a population. Therefore, the aim is not to have a mathematically exact model, but one that robustly captures general patterns of spatial fluctuations of allele frequencies. We use a model of a two-dimensional continuous habitat that is interrupted by a barrier and assume that the demographic parameters are the same on both sides. In short, we calculate the chance of pairwise coalescence before a long distance migration or mutation event. We use a diffusion model to trace lineages backwards in time, and assume that rare long distance migration events, which counteract the build up of local allele frequency fluctuations, occur at a constant rate. To calculate the equilibrium identity by descent pattern, i.e. the probability that two lineages coalesce before a rare long distance migration or mutation event happen, we first derive the coalescence probability at specific times t in the past and then integrate over t .

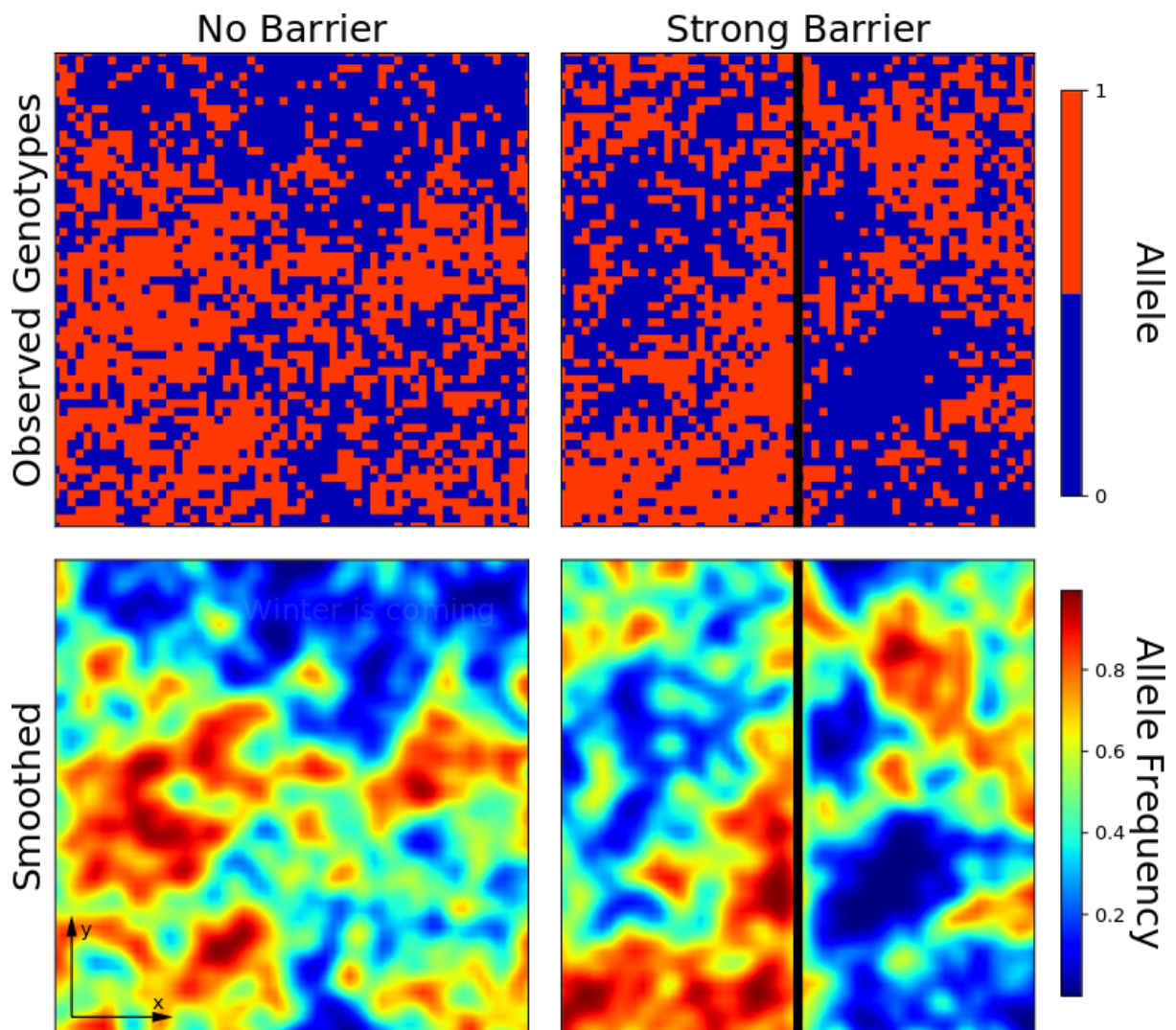


Figure 3.1: Geographic fluctuations of allele frequencies. We simulated individual biallelic genotype data for one locus in a two-dimensional habitat in absence of a barrier to gene flow (left) and in presence of a strong barrier (right). We used the backward simulation scheme on a grid outlined below with a axial variance of dispersal $\sigma^2 = 1$, one genotype per deme, a long distance migration rate $\mu = 0.001$ and a complete barrier $\kappa = 0$. Top: Discrete genotype data. The two colors in each grid position code for the two possible allelic states. Bottom: Allele frequencies are smoothed with a Gaussian Kernel for better visualization. This figure demonstrates the underlying idea of our method: A barrier distorts random geographic fluctuations of allele frequencies. The average strength of these local fluctuations increases next to the barrier, and there is less correlation across the barrier. This signal can be used to infer the strength of the barrier.

Our model is closely related to previous theoretical treatments of allelic identity by state in presence of a barrier to gene flow (Nagylaki, 1988; Barton, 2008). For a population occupying a linear habitat, Nagylaki (1988) derived continuous equations for identity by state by taking the limit of a model of linear demes that exchange migrants (the so called stepping stone model). Barton (2008) expanded this by solving an analogous equation for two-dimensional populations. His formulas are given as a numerical Fourier transform that diverges for nearby individuals. These equations for two-dimensional populations are formally problematic, as they were not obtained by rescaling (which is impossible in two spatial dimensions), but Barton (2008) demonstrated that the solution is in close agreement with the solutions from a discrete stepping stone model for all but very close distances.

Here, we base our inference scheme on a different approach. We use a diffusion approximation to describe the spread of ancestry backwards in time. While the results are formally equivalent, we found our model to be computationally more robust, and, most importantly, more efficient. This approach allows us to apply our method to sample sizes of hundreds to thousands of individuals, and dozens to hundreds of loci.

Diffusion approximation

We model the spread of ancestry using a geographic diffusion approximation, which has a long history in population genetics (Fisher, 1937; Wright, 1943; Malécot, 1948; Nagylaki, 1978). Tracing a lineage of one locus backwards in time, the total spatial movement is the sum of many independent migration events. If these events are sufficiently uncorrelated, the central limit theorem establishes that the total displacement tends towards a normal distribution. Therefore, the overall spread of ancestry can be approximated as a random walk process. This approximation is accurate as long as rare large scale events do not significantly influence the movement of ancestral lineages. The diffusion approximation is expected to be most accurate on recent to intermediate time scales, on which large scale events such as colonizations often play only a minor role.

In the absence of a barrier, the process is a free diffusion, and the probability density function (PDF) of finding an ancestor at position x at time t back along a given axis is given by a Gaussian probability density function around the current position x_0 and variance $\sigma^2 t$ that increases linearly backwards in time:

$$G_0(x, y, t) := \frac{1}{\sqrt{2\pi\sigma^2 t}} \exp\left(-\frac{(x - y)^2}{2\sigma^2 t}\right). \quad (3.1)$$

The dispersal rate σ describes the speed of the spread of ancestry. In case of a homogeneous density of individuals across the landscape, the backward dispersal probability

density equals the probability density of lineages moving forward in time. The dispersal rate σ^2 can be interpreted as the axial variance of the one generational dispersal kernel then, if time is measured in generations (Rousset, 1997). The diffusion approximation is fully determined by the equation:

$$\frac{\partial}{\partial t} G(x, y, t) = \frac{\sigma^2}{2} \frac{\partial^2}{\partial x^2} G(x, y, t) \quad (3.2)$$

with initial condition:

$$G(x, y; t = 0) = \delta(x - y). \quad (3.3)$$

Partial barrier to gene flow

We model a barrier as a partially permeable barrier to diffusion of ancestry. For a barrier at $x = 0$, the following interface boundary conditions have to be supplied (Grebenkov et al., 2014):

$$\begin{aligned} \lim_{\epsilon \rightarrow 0} \left(\frac{\partial}{\partial x} G(x, y, t) \right)_{x=+\epsilon} &= \lim_{\epsilon \rightarrow 0} \left(\frac{\partial}{\partial x} G(x, y, t) \right)_{x=-\epsilon} \\ &= \frac{2\kappa}{\sigma^2} \left(\lim_{\epsilon \rightarrow 0} G(\epsilon, y; t) - \lim_{\epsilon \rightarrow 0} G(-\epsilon, y; t) \right). \end{aligned} \quad (3.4)$$

The first line describes the constancy of the flux across the barrier. For $\kappa = 0$ there is no flux across the barrier, and the barrier is infinitely strong. On the other hand, the case $\kappa = \infty$ implies the continuity of the probability density across the barrier and the solution reduces to free diffusion. Comparing to differential equation (54) of (Nagylaki, 1988), which is derived by rescaling a stepping stone model, gives an intuitive interpretation of κ : This parameter corresponds to the fraction of successful migrants across a barrier if demes are spaced one dispersal unit apart. The quotient $\gamma := \frac{2\kappa}{\sigma^2}$ corresponds to an equivalent factor in formula (54) of Nagylaki (1978). It is also the inverse of the barrier strength parameter B defined by Barton and Bengtsson (1986), which has dimension of distance.

Equations 3.2, 3.3 and 3.4 allow for an analytic solution for the probability density with a barrier (Grebenkov et al., 2014):

$$\begin{aligned}
G(x, y, t) &= \frac{\exp\left(-\frac{(x-y)^2}{2\sigma^2 t}\right) + \exp\left(-\frac{(x+y)^2}{2\sigma^2 t}\right)}{\sqrt{2\pi\sigma^2 t}} \\
&\quad - \frac{2\kappa}{\sigma^2} \exp\left(\frac{4\kappa}{\sigma^2}(y-x+2kt)\right) \operatorname{erfc}\left(\frac{y-x+4\kappa t}{\sqrt{2\sigma^2 t}}\right), \quad x > 0 \\
G(x, y, t) &= \frac{2\kappa}{\sigma^2} \exp\left(\frac{4\kappa}{\sigma^2}(y-x+2kt)\right) \operatorname{erfc}\left(\frac{y-x+4\kappa t}{\sqrt{2\sigma^2 t}}\right), \quad x < 0,
\end{aligned} \tag{3.5}$$

where $\operatorname{erfc}(z)$ denotes the complementary error function. These expressions are valid for $y > 0$, and their extension to $y < 0$ is straightforward by the symmetry $x, y \rightarrow -x, -y$. In Fig. 3.2, these formulas are compared to random walk simulations. The PDF converges to the Gaussian of free Brownian motion for $\kappa \rightarrow 0$. For a two-dimensional diffusion process with a linear barrier at $x = 0$, the full solution is given by multiplying the one-dimensional density functions Eq. 3.2 for movement parallel to the barrier and Eq. 3.5 for movement normal to the barrier.

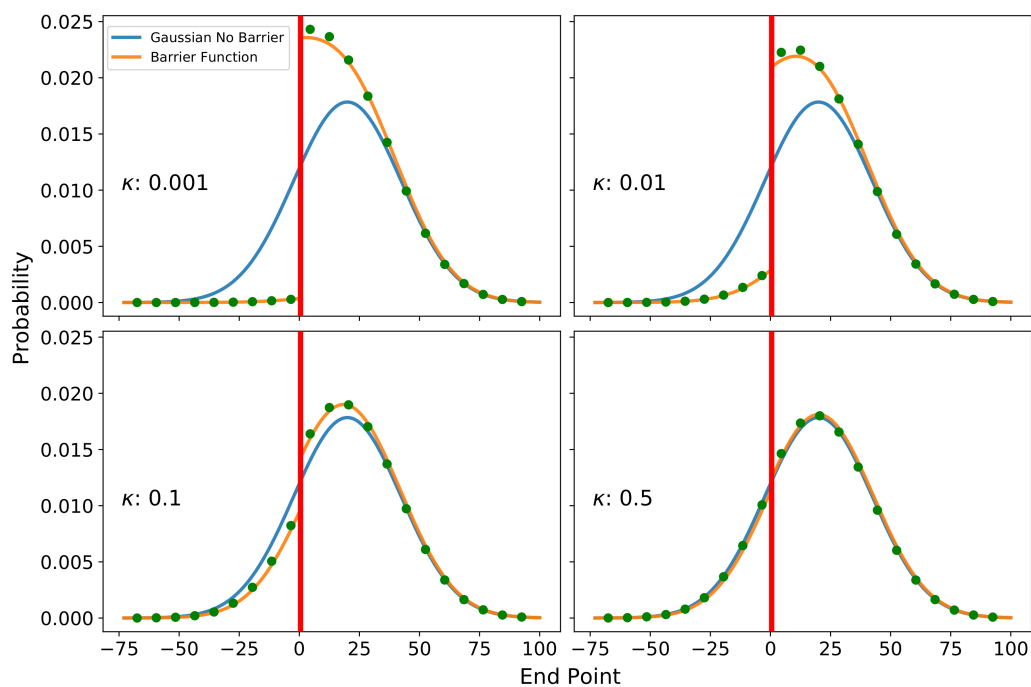


Figure 3.2: Comparison of analytical diffusion formulas (Eq. 3.5) with discrete random walk simulations for $t = 500$ in the past. We simulated a one-dimensional random walk on an array of linear discrete nodes at all integers and a barrier at $x = 0.5$. Every generation, a random step to one side is made (which implies that $\sigma = 1$). If a movement would be across the barrier, it is realized with probability κ (otherwise no step is made). For each of four barrier strengths, we simulated 10^6 replicates starting at $x = 20$. The blue line depicts the corresponding Gaussian probability density of free diffusion in absence of a barrier.

Pairwise coalescent probabilities

We use the diffusion approximation to model the distribution of coalescence times for pairs of individuals. We approximate the chance of co-ancestry originating in a small time interval dt around time t ago as the product of the probability of coming close and a rate of local coalescence $1/(2D_e)$ (Ringbauer et al., 2017a). The local density D_e describes a rate with which nearby lineages coalesce (Wright, 1943). In a stepping stone model, D_e corresponds to the number of diploid individual per deme (Barton et al., 2002). This approximation ignores that lineages do not move apart again once they have coalesced. An equivalent simplification is made by Barton (2008), and it has been shown to be accurate as long as coalescence is sufficiently rare (Wilkins, 2004). Since the dispersal process is symmetrical in time in our model of constant population density, the chance that two lineages at current position x and y are close at time t back equals the probability density that a single lineage moves from x to y in time $2t$. Formally, the probability density $\psi(x, y, t)$ of coalescence T_c at time t in the past is approximated by

$$\psi(x, y, t) = \frac{\Pr(T_c \in [t, t + dt])}{dt} \approx G(x, y, 2t) \frac{1}{2D_e}. \quad (3.6)$$

Identity by descent

We define identity by descent F of two samples at \vec{x} and \vec{y} as the chance that two lineages coalesce before a long distance migration or, equivalently (but unlikely for single nucleotide polymorphisms), a mutation event occurs along one of the lineages. This definition is closely related to the widely used fixation index F_{ST} , and both definitions agree in the limiting case of an infinite population (see Rousset (2002) for a review). If one assumes that mutation or long-distance migration occur at a constant rate μ , it is straightforward to calculate the probability that two lineages coalesce before a long distance event:

$$F(x, y) = \int_0^\infty \psi(x, y, t) \exp(-2\mu t) dt \quad (3.7)$$

In the absence of a barrier, the probability of identity by descent varies only with the Euclidean distance r between two individuals. In this special case, the integral in Eq. 3.7 has an analytical solution, the classical Wright-Malecot formula (Barton et al., 2002, 2013):

$$F(r) = \frac{1}{4\pi D_e \sigma^2} K_0 \left(\sqrt{2\mu} \frac{r}{\sigma} \right), \quad (3.8)$$

where K_0 is the modified Bessel function of the second kind of degree zero. A well known caveat of this analytical solution is that K_0 diverges logarithmically as $r \rightarrow 0$ (Barton et al., 2002). Similarly, the integral in Eq. 3.7 diverges for nearby individuals. As for the Wright-Malecot formula, this is caused by a behavior of the diffusion approximation for short timescales, as in this model the chance of two lineages being close diverges as $1/t$ for $t \rightarrow 0$. An obvious solution to circumvent this problem is to start integration at time $t_0 > 0$. Here, we choose one generation time, a biologically plausible value. We could not find an analytical solution, but Eq. 3.7 can be numerically integrated.

Our results show that if a barrier to gene flow is present, identity is decreased across the barrier, and increased for points on the same side of the barrier (Fig. 3.3). Interestingly, the increase of F for a pair of points \vec{x} and \vec{y} on the same side of the barrier equals the decrease of identity between points \vec{x} and \vec{y}' , where \vec{y}' is the point \vec{y} reflected across the barrier. This symmetry originates from a reflection principle of the underlying random walk model, as lineages that do not cross the barrier behave as if they were reflected. This symmetry already occurs in the barrier point density function (Eq. 3.5). It implies that for a complete barrier identity by descent can increase to at most twice the value in absence of a barrier, as observed in the equivalent case of a range boundary (Wilkins, 2004).

Rescaling

Not all parameters in Eq. 3.7 are independent. Consequently, they cannot be estimated separately, as in absence of a barrier (Rousset, 1997; Barton et al., 2013). Therefore, we replace the four demographic parameters $\vec{\theta} : D_e, \kappa, \mu, \sigma$ in equation Eq. 3.6 with three compound parameters $\bar{\theta}$: Neighborhood size $N_b := 4D_e\sigma^2\pi$ (a classical parameter that goes back to Wright (1943)), a scaled barrier parameter $\gamma := \frac{2\kappa}{\sigma^2}$ (corresponding to the inverse of Barton's B), and a scaled long-distance migration rate $m := \frac{2\mu}{\sigma^2}$ (Appendix).

3.1.2 Fitting the model to data

A typical dataset consists of diploid genotypes $g_1^i, \dots, g_n^i \in \{0, 0.5, 1\}$ for a marker i and individuals at geographic positions $\vec{p}_1, \dots, \vec{p}_n$. To infer the underlying demographic parameters $\vec{\theta}$ from observed data, we have to develop a way to fit our model to such data.

In principle, it is straightforward to transform the probability of identity by descent as calculated by our model into expected allele frequency covariances. Denoting the

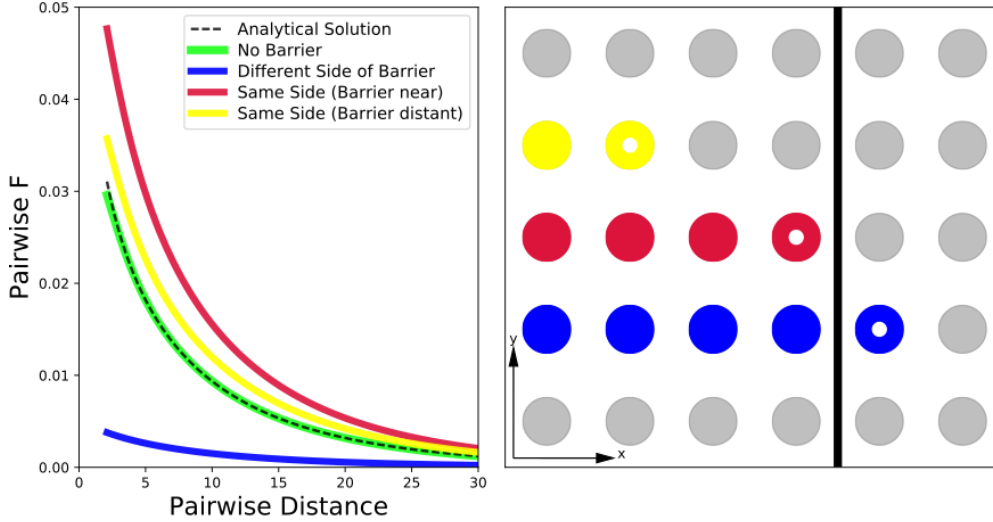


Figure 3.3: Decay of identity by descent F in presence of a strong barrier to gene flow ($\kappa = 0.01$), and moderate neighborhood size ($\sigma = 1$, $D_e = 5$, $\mu = 0.003$). We calculate F for an individual at $x = 1$ (blue illustration), an individual at $x = -1$ (red illustration) and an individual at $x = -5$ (yellow illustration) with, for all of them, other individuals to the left of the barrier at $x = 0$ that only differ in their x -coordinate (right panel). The analytical solution in the absence of a barrier is given by the Wright-Malecot formula (Eq. 3.8), the others are integrated numerically (Eq. 3.7).

probability of identity by descent between a pair of samples k and l by F_{kl} and the population mean allele frequency of a marker i by \bar{p}_i , the expected covariance between their genotypes is given by:

$$\text{Cov}(g_k^i, g_l^i) = F_{kl}\bar{p}_i(1 - \bar{p}_i)$$

However, often mean allele frequencies have to be estimated from the data, and estimating the means of allele frequencies for many markers would lead to over-fitting. To circumvent this caveat, we developed and tested different methods to fit identity by descent to genotype data without directly estimating all mean allele frequencies. We included one approach that models individual genotypes as binomial draws from latent allele frequencies modeled by a Gaussian random field (Supp. Text 1).

Fitting pairwise homozygosity

Of all tested fitting methods a relatively simple approach that fits the fraction of pairwise homozygosity (defined as fraction of identical genotypes) has least bias and sampling variation (Supp. Text 2). Throughout this work, we use this method for data anal-

ysis. In the following we give a brief outline of how we calculate expected homozygosity and how we fit it to data.

The observed average homozygosity h_{kl} for a pair of individuals k and l with genotypes g_k^i and $g_l^i \in \{0, 0.5, 1\}$ at markers $i = 1, 2, \dots, n$ can be straightforwardly calculated from the data:

$$h_{kl} = \frac{1}{n} \sum_{i=1}^n \left(g_k^i g_l^i + (1 - g_k^i)(1 - g_l^i) \right). \quad (3.9)$$

Our model predicts the pairwise chance of identity by descent F , and these probabilities can be used to calculate the expected values of average homozygosities $\mathbb{E}(h_{kl})$. The expected pairwise homozygosity of a pair of samples k and l at a marker i is given by:

$$\mathbb{E}(h_{kl}^i) = F_{kl} + (1 - F_{kl})(\bar{p}_i^2 + (1 - \bar{p}_i)^2).$$

The first term gives the probability of having the same genotype due to identity of descent and the second describes the probability of having the same genotype by chance. To avoid estimating all mean allele frequencies, we average over all n markers to get the expected fraction of pairwise identical genotypes:

$$\mathbb{E}(h_{kl}) = F_{kl} + (1 - F_{kl}) \underbrace{\frac{1}{n} \sum_i (\bar{p}_i^2 + (1 - \bar{p}_i)^2)}_{:=s} \quad (3.10)$$

Instead of fitting all unknown allele frequencies, now only one additional compound parameter s has to be fit in addition to the demographic parameters θ . We tried fitting this formula to observed data with a composite likelihood approach. However, we found that minimizing the sum of all squared deviations between the expected and observed pairwise homozygosities

$$\bar{\theta} = \min_{\theta, s} \sum_{k < l} (\mathbb{E}(h_{kl}(\theta, s)) - h_{kl})^2 \quad (3.11)$$

gives almost identical results, while being additionally much faster (Supp. Text 2).

Fitting pairwise homozygosities can be easily extended to deme data, where nearby individuals have been binned, by plugging deme allele frequencies into Eq. 3.9

Estimation uncertainties

To learn about estimation uncertainty, we bootstrap over genetic markers. Unlinked markers contain almost independent information because their spatial movements are typically correlated only on very short timescales. Therefore, resampling loci at random is expected to yield accurate empirical confidence intervals.

3.1.3 Implementation

In brief, our inference scheme runs the following three computational steps for a given set of demographic parameters $\vec{\theta}$:

1. Calculate pairwise F for all pairs of samples with integral Eq. 3.7.
2. Use these pairwise F to calculate the expected pairwise homozygosity for all pairs of samples with Eq. 3.10.
3. Calculate the sum of squared differences between the expected and observed pairwise homozygosities (Eq. 3.11).

Our program then finds the parameters $\bar{\theta}$ that minimize this function by using the Levenberg-Marquardt algorithm, as implemented in the Python package *Scipy*.

We implemented the described simulation and inference methods mostly in Python. To speed up calculations we parallelized the calculations for pairwise F , so that they can be run simultaneously on different CPUs. The evaluation of the integrand Eq. 3.7 is a computational bottleneck. We implemented this calculation in C, to make use of the superior speed of a compiled language.

The inference scheme has to compute the expected identity by state for every pair of samples. It therefore scales quadratically with the number of individuals, as there are $\frac{n(n-1)}{2}$ such pairs. It has a runtime of several hours for a individual/deme number of 1000 when run on a single standard desktop CPU. In order to produce a sufficient number of replicates and bootstraps, we utilized a scientific computer cluster at IST Austria. To speed up runtime, individuals can be grouped into demes. If clustering is done on small scales for bins at most a few σ in diameter, it does not significantly affect the estimation scheme (Supp. Text 2). In our results, we capped γ to a maximum value of 1, as for $\gamma > 1$ the effect of a barrier becomes negligible (Fig. 3.2).

3.1.4 Simulations

We extensively tested our inference scheme on simulated data sets. We used a stepping stone model with D_e individuals per deme, and we traced ancestry backwards in time (Fig. 3.4). Every generation each individual picks an ancestor at random with probabilities given by a dispersal kernel. Here, we use a discretized Laplace distribution as axial dispersal kernel. Due to the rapid convergence to the continuous diffusion approximation (Fig. 3.2), the specific choice of dispersal kernel has no significant impact as long as its axial variance σ^2 remains finite. If two lineages happen to pick the

same ancestor, they coalesce into a single ancestral lineage. We simulate long distance migration events to occur at a constant rate. If they occur, the corresponding lineage picks an allele at random from the population mean allele frequency \bar{p} . To model the effects of a barrier, we follow Nagylaki (1988) and realize migration events across the barrier only with relative probability κ , the barrier strength parameter. For constant deme sizes, this backwards model is equivalent to a forward model in which a large number of gametes disperse with the same dispersal kernel (Nagylaki, 1988).

After a preset maximum number of generations, every lineage picks an allele at random according to the mean allele frequency \bar{p} . Different, unlinked loci were simulated as independent runs. We picked mean allele frequencies at random according to a predetermined distribution, usually Gaussian with standard deviation $\sigma(\bar{p})$ around an overall mean of 0.5. We also investigated how robust our inference scheme is to scenarios of secondary contact. We simulated them by assigning each ancestral lineage an allele with probability \bar{p}_l or \bar{p}_r , according to its location at time of secondary contact.

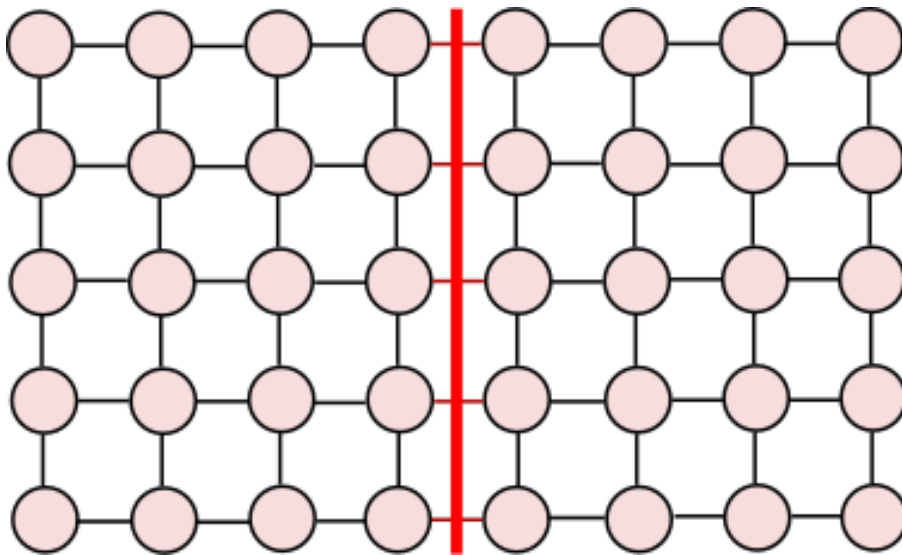


Figure 3.4: Model used to generate synthetic data sets. Ancestry is traced back on a two-dimensional grid of demes in discrete time steps. An ancestral deme is randomly picked according to a dispersal kernel, and then an ancestor is chosen at random from within this deme. If lineages fall on the same ancestor, they coalesce into a single lineage. For each unique lineage and every step, there is constant chance that a long distance migration event occurs. In this case, the lineage and all corresponding individuals pick the same allele randomly drawn from a mean allele frequency.

We stress that the data is simulated under a process very similar to our model. For real data, there could be other deviations which might further reduce the power of the inference scheme. Therefore, our results should be seen as limits for the inference

scheme in case of ideal data.

3.1.5 *Antirrhinum majus* Data

To show the practical utility of our inference scheme, we applied it to data from an *Antirrhinum majus* hybrid zone. This hybrid zone is located in a valley in the Eastern Pyrenees. It shows a geographically narrow transition between two flower color morphs, in which a range of hybrid flower color phenotypes occur. This transition is mainly determined by three major flower color genotypes that regulate the intensity and patterning of flower color (Whibley et al., 2006; Bradley et al., 2017). We applied our method to a dataset of 12389 plants collected between 2009 and 2014, which were genotyped for 112 SNP markers.

To satisfy the assumptions of our model as good as possible, we filtered markers based on four quality criteria: minor allele frequency, large scale geographic correlation, linkage disequilibrium and deviations from local Hardy Weinberg equilibrium. SNP design and filtering are explained in detail in Supp. Text 3. After this data cleaning step, we were left with 60 unlinked polymorphic SNPs, that were spaced throughout most of the genome (Supp. Text 3).

3.1.6 Data Availability

The source code for the implementation of our inference scheme is freely accessible at the Github repository <https://github.com/hringbauer/BarrierInferPublic.git>

The *Antirrhinum majus* dataset is a subset of samples collected from 2009 to 2014 with the long term goal to build a pedigree. The details of this dataset and data filtering are described in Supp. Text 3.

3.2 Results

3.2.1 Inference on Simulated Data

We investigated the overall capability of this method to estimate barrier strengths and the accuracy of empirical bootstrap uncertainty estimates. Our tests show that the inference scheme can reliably recover barrier strengths as well as demographic parameters (Fig. 3.5 and Fig. 3.6). Estimates of the neighborhood size are robust, but show a slight upward bias. These slight biases are likely due to the fact that a continuous model is

used to fit to discrete simulations. Estimates of the long distance migration rate m are more variable, but are not significantly biased. In all cases, the range of bootstrap estimates mostly overlaps with the true value to the expected degree. This result indicates that bootstrapping gives accurate uncertainty estimates (Fig. 3.6).

The stronger the barrier, the more strongly it affects allele frequency fluctuations. Our results indicate that the inference scheme has higher power to infer strong barriers ($\gamma < 0.1$), whereas weaker barriers cannot be inferred reliably (Fig. 3.5). The exact power of the method will depend on a combination of several factors, in particular the strength and the geographic extent of isolation by distance patterns, as well as the geographic sampling scheme. As a barrier mostly affects fluctuations near it, generally a high sampling density next to the putative barrier is preferable. Our power simulations for a specific, realistic scenario indicate that at least a few dozen markers and several hundred individuals are required for robust inference of strong barriers (Supp. Text 1).

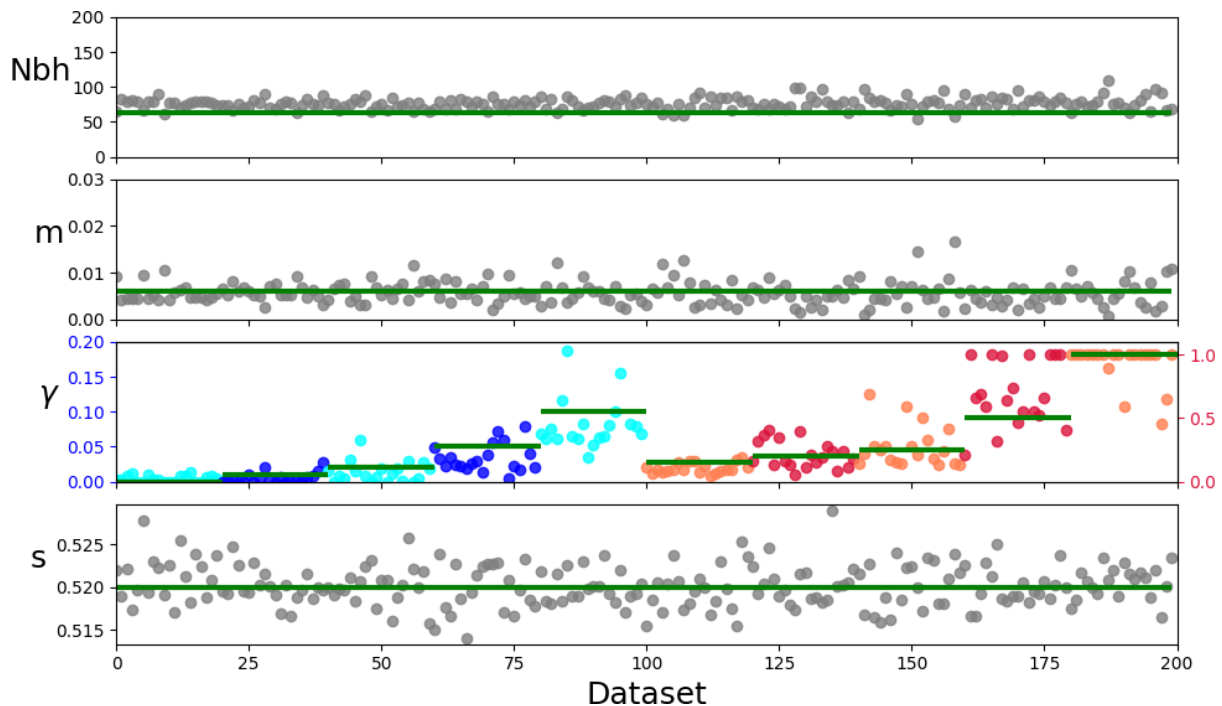


Figure 3.5: Parameter estimates on simulated data. For each replicate, we simulated a population of 60×40 individuals spaced one node apart, and then applied the inference scheme to fit neighborhood size N_{bh} , barrier strength γ , scaled long-distance migration rate m , as well as the allele frequency variance parameter s . To keep run-times manageable, we binned individuals into 2×2 demes. We simulated 10 different barrier strengths ($\gamma = 0, 0.01, 0.02, 0.05, 0.1, 0.15, 0.2, 0.25, 0.5, 1.0$), with 20 replicate runs each. Every dot represents an estimate for one such replicate. Horizontal lines depict the true parameters used in the simulations ($\sigma = 1, D_e = 5, \mu = 0.003, \sigma(\bar{p}) = 0.1$). We split up the barrier plot into two parts with different axes (blue and red) for better visibility.

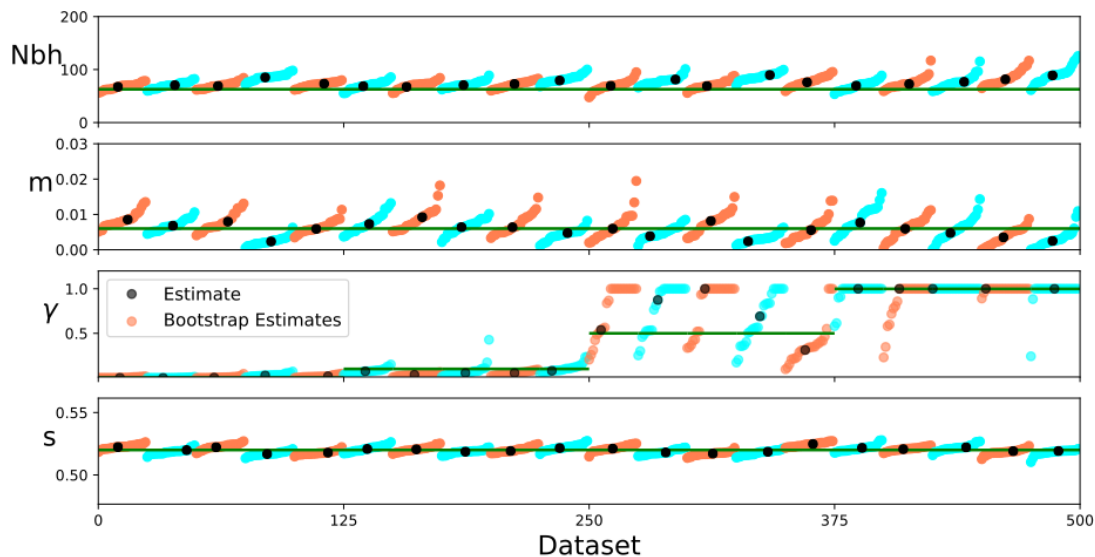


Figure 3.6: Bootstrap estimates on simulated data: We simulated 60×40 individuals on a grid spaced one distance unit apart. We simulated 5 replicates of 4 different barrier strengths ($\gamma = 0, 0.1, 0.5, 1.0$). For each of the 20 data-sets we inferred the parameters and additionally did 20 estimates when bootstrapping over loci. Horizontal lines depict the true parameters used in the simulations ($\sigma = 1, D_e = 5, \mu = 0.003, \sigma(\bar{p}) = 0.1$)

Secondary contact

Barriers to gene flow sometimes coincide with areas of secondary contact, for instance in secondary hybrid zones (Barton and Hewitt, 1985). Allele frequencies might have diverged before this contact, and present day allele frequency differences are not caused by the presence of a barrier. However, these clines resembles the effect of a barrier, and might be mistakenly inferred as such. One salient way to deal with this problem is to remove markers that show large scale geographic structure of allele frequency. One can base inference on a subset of markers that have similar mean allele frequency across the whole population range and only display fluctuations on small geographical scales, which equilibrate quickly (Barton et al., 2013). We tested this approach on simulated data. When applying the inference scheme to a simulated scenario of secondary contact with divergent allele frequencies, it wrongly infers a barrier in case of no filtering (Fig. 3.7). However, when using the subset of loci that show no large scale correlation with geography, the false positive signal decreases. Moreover, filtering out loci with large scale structure does not remove the signal in case of a true barrier since secondary contact (Fig. 3.7). However, if sampling is only done on small spatial scales, such filtering could become problematic, as one might remove signal from local fluctuations as well. Therefore we advise to always check that the sampling area is bigger than the spatial

scale of isolation by distance patterns before any markers are removed.

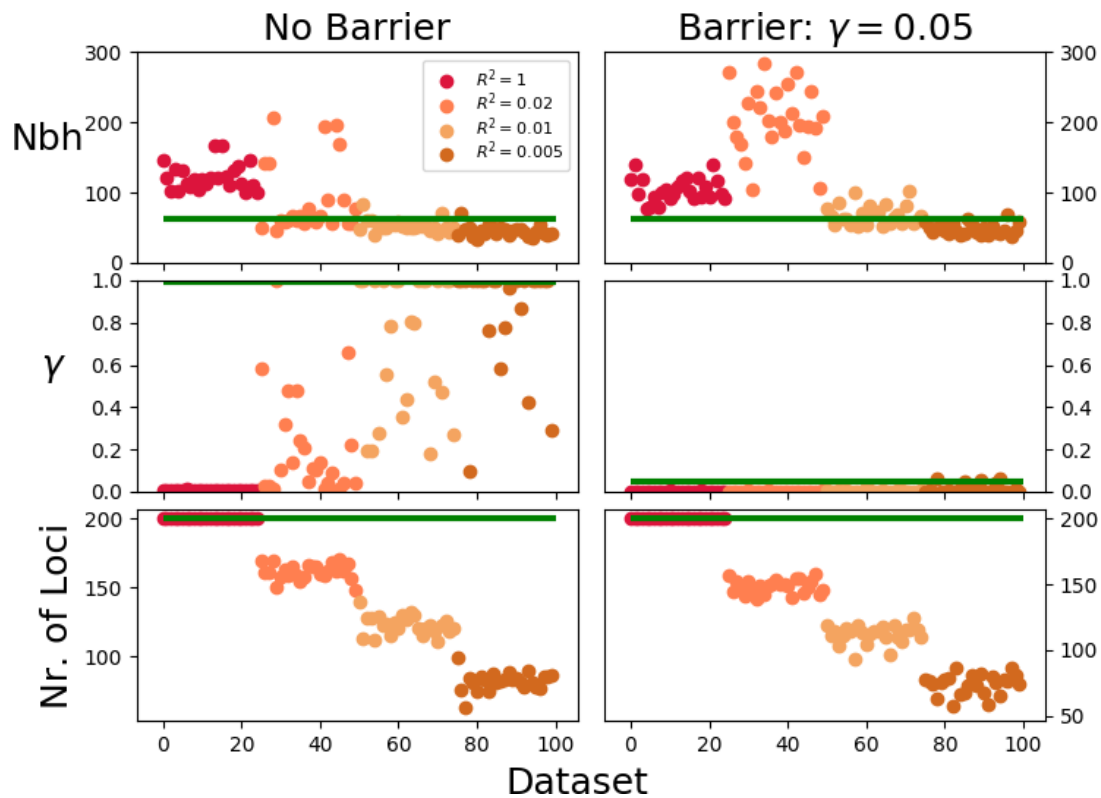


Figure 3.7: Parameter estimates in a simulation of secondary contact. The two ancestral populations allele frequencies were drawn independently from a Gaussian with standard deviation 0.1 around an overall mean of 0.5. We simulated a population of 50×20 individuals spaced two dispersal units apart, with a barrier in the middle of the x -axis, and secondary contact 100 generations ago. We simulated 25 replicates of two scenarios after contact. Left: No barrier to gene flow after contact. Right: A strong barrier to gene flow ($\gamma = 0.05$). We then filtered out loci that were correlated with either x - or y -axis coordinates more than four different R^2 values before doing inference for each of these replicates. The two bottom figures show the number of filtered loci that were used for inference.

Unknown barrier locations

Our inference scheme assumes that the location of the putative barrier is known a priori. In practice one might not always have this information, or one perhaps wants to test the hypothesis of barriers in different locations. In this case, one can repeatedly apply the inference scheme and fit different potential barrier positions. When testing this approach on simulated data, the inference scheme only inferred a strong barrier near the true position (Fig. 3.8). The estimate uncertainties on the habitat edges are inflated.

This effect is caused by limited power to infer barriers near sampling edges: One needs a sufficient number of samples on both sides of a barrier to fit the strength of a barrier.

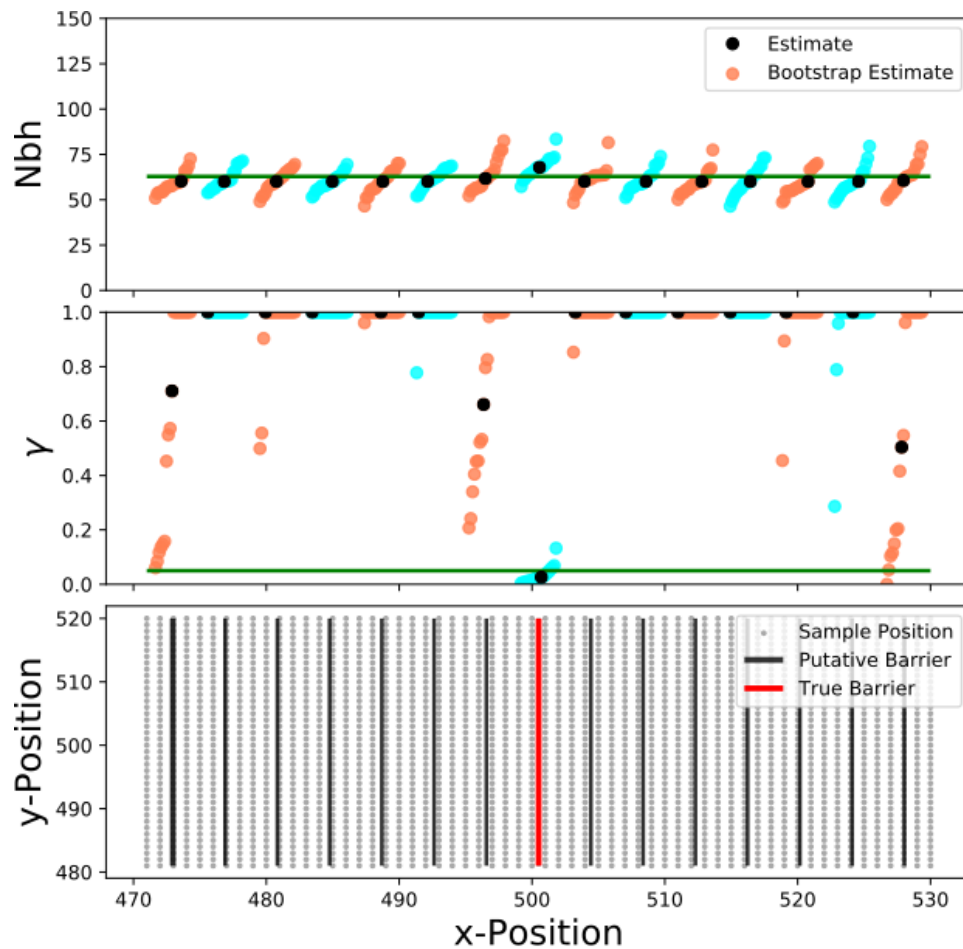


Figure 3.8: Testing various putative barrier positions on synthetic data. We simulated a single dataset as in Fig. 3.5 with a barrier strength of $\gamma = 0.05$. We applied our inference scheme to fit 15 putative barrier locations, and bootstrapped 20 times over loci for each putative barrier to visualize the uncertainty of the inferred parameters.

3.2.2 Hybrid Zone Analysis

We observed a clear isolation by distance pattern across the *Antirrhinum* hybrid zone (Fig. 3.9). On average, mean identity by state for nearby plants is elevated by about two percent above the background level, and falls away with increasing pairwise distance, most rapidly over the first 2000 meters. Our inference scheme fits this pattern well, with an estimated neighborhood size of 188 (95% bootstrap confidence interval: 120 – 240, Fig. 3.10).

Obviously there are demographic complications that are not captured by our model, such as heterogeneities in plant distributions and density. The population is not dis-

tributed uniformly in two dimensions, as plants are often found in patches of suitable habitat (Fig. 3.10). However, our analysis indicates that isolation by distance patterns are neither strongly influenced by the cardinal direction or relative plant positions, nor the geographic location within the hybrid zone (Supp. Text. 3). These observations imply that our model assumptions are not grossly violated.

For putative barriers in the center of the hybrid zone, our inference scheme estimates no barrier ($\gamma > 1$). Bootstrap estimates rarely fall below $\gamma = 0.5$, and all of them are above $\gamma = 0.1$. When testing for barriers towards the flank of the hybrid zone, estimates get more variable. Some estimates in both flanks indicate an intermediate barrier to gene flow, but in each case some of the corresponding bootstrap estimates also take the value of no barrier ($\gamma > 1$). This signal likely reflects the lower power to infer barriers in these regions, as there is a higher sampling density in the center of the hybrid zone (58.7% of the samples originate from within 500 meter of the flower color transition). Only in one case, for the leftmost tested barrier, the bootstrap estimates do not overlap with the value of the null hypothesis ($\gamma = 1.0$). This area also shows the strongest small scale isolation by distance pattern (Supp. Text 3), and the density of plants is very low in sampled patches (Personal Communication: Maria Melo). Therefore, a potential explanation for this significant deviation from the null hypothesis of constant demography and no barrier to gene flow is an exceptionally low plant density in this area.

Given the overall good fit of isolation by distance with our inference scheme (Fig. 3.9), our results indicate that there is no strong genome wide barrier to contemporary gene flow that coincides with the flower color transition. As such a strong barrier would require many barrier loci spaced densely throughout the genome (Barton and Bengtsson, 1986), this result comes as no surprise. At the moment, no other traits apart from flower color are known to be divergent across the hybrid zone, despite much work to detect them (Personal Communication: Maria Melo).

Previous results suggest the presence of a barrier to exchange of flower color alleles (Whibley et al., 2006; Bradley et al., 2017) and indicate that selection maintains differences in flower color (Ellis, 2016). Therefore, we applied our method to a subset of polymorphic markers in the genetic neighborhood of two genes known to affect flower color variation in the hybrid zone, *Rosea* and *Eluta*. However, bootstraps estimates varied widely for all tested barrier locations (results not shown), which indicates that there is not sufficient signal in the data. This lack of power is likely due to the low number of suitable SNP markers without steep allele frequency clines near this region in our dataset (< 10). Simulations confirmed that for this low number of markers there is not enough power to detect even strong barriers (Supp. Text 1).

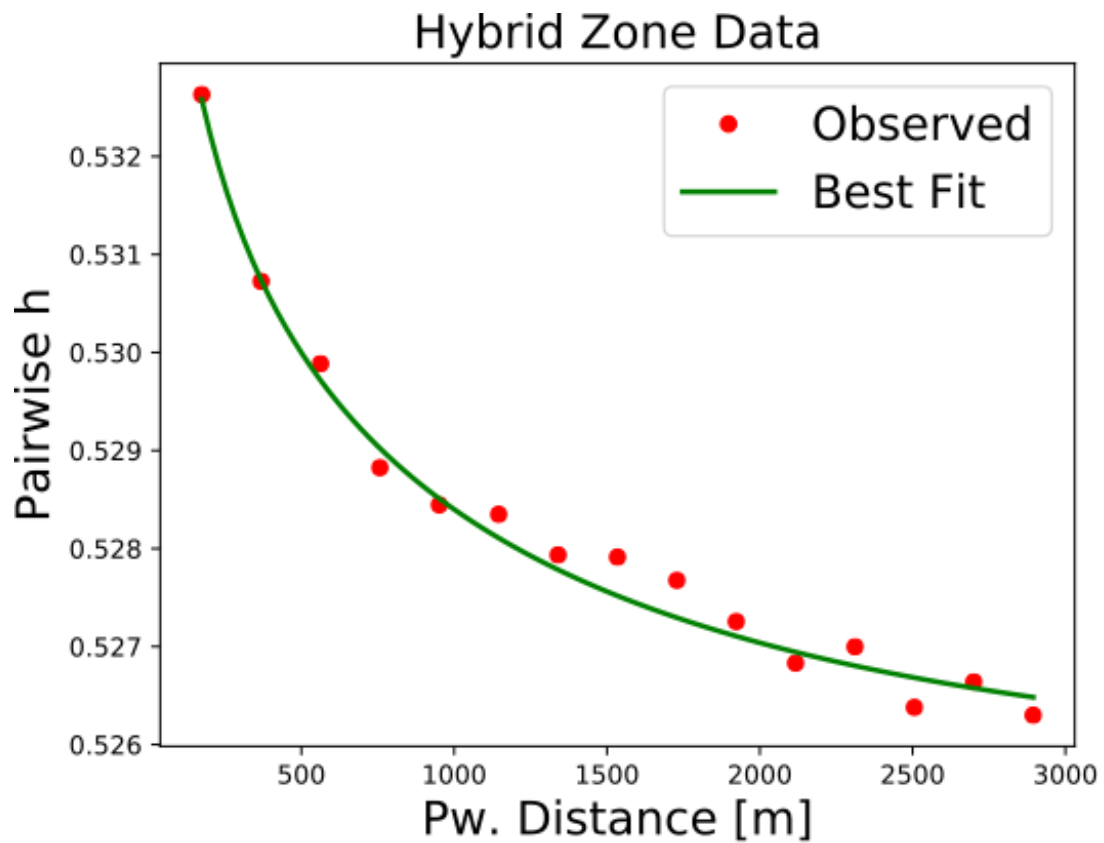


Figure 3.9: Decay of pairwise homozygosity with geographic distance for hybrid zone data. All pairwise homozygosities for the filtered dataset were binned according to pairwise distance. We also plot the best fit ($Nbh = 188, m = 4.4 \cdot 10^{-4}, s = 0.5247$). Formula 3.10 can be used to translate pairwise homozygosity into pairwise F .

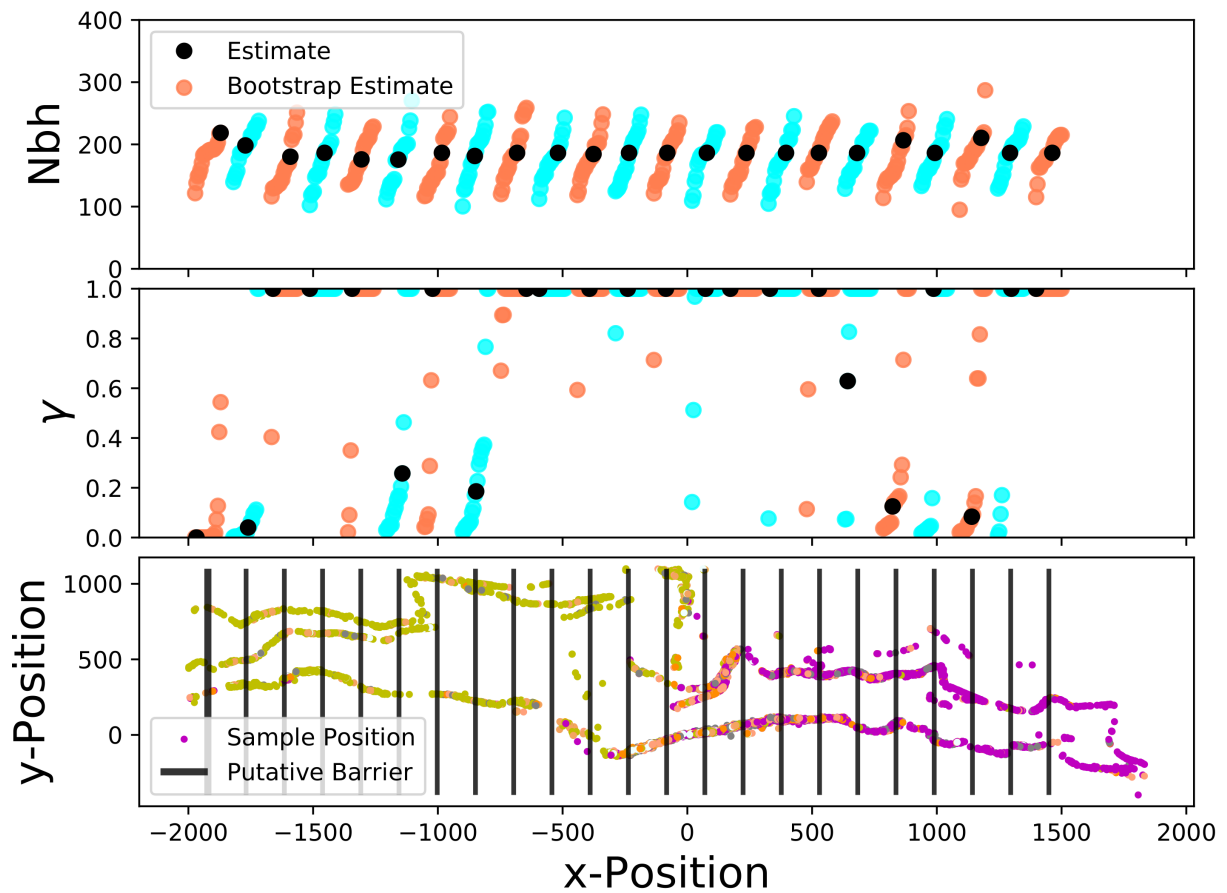


Figure 3.10: Inference of various putative barrier positions on data from a *Antirrhinum majus* population. We used our inference scheme to fit 15 barrier locations, and bootstrapped 20 times over loci for each putative barrier (bootstrap estimates are shown in orange and cyan). Bottom: Geographic distribution of 12389 samples. The color of each dot represents flower-color phenotype.

3.3 Discussion

To our knowledge, our scheme is the first method that infers the strength of the barrier based on an explicit spatial population genetic model. There are several similarities with the inference method BEDASSLE (Bradburd et al., 2013), which aims to disentangle the effects of geographical isolation by distance and differences in ecological variables. This scheme is not based on an explicit population genetic model, but rather fits the decay of genetic similarity with distance to a heuristic formula. In light of possibly very complex demographic structure this approach is not necessarily worse than fitting our spatial model. However, this approach does not take the increased covariances on the same side of a barrier into account, and does not make use of some valuable signal because of that. Moreover, it uses a MCMC approach that is based on a model of Gaussian random Fields, which is computationally too expensive to apply to hundreds or even thousands of individuals or demes. In contrast, our method is well suited to data sets of this magnitude.

Another widely used method to infer barriers to gene flow, Geneland, clusters individuals using their explicit geographic coordinates (Guillot et al., 2005). Safner et al. (2011) identified it as one of the most potent methods to infer barriers to gene flow. Therefore, we tested its performance on some of the datasets which we have generated to test our scheme (Supp. Text 4). As described previously (Guillot and Santos, 2009; Safner et al., 2011), Geneland's ability to accurately infer barriers to gene flow decreases when isolation by distance patterns are present, as its underlying model assumptions of discrete populations with well-defined allele frequencies are violated. Indeed, it fails to detect a barrier to gene flow in our test data sets, that all exhibit such isolation by distance patterns (Supp. Text 4). In contrast, the method introduced here can give accurate estimates of the barrier strength in these scenarios. It is not confounded by isolation by distance patterns; it in fact relies on the presence of this signal. Our method can therefore be seen as complementary to Geneland.

In contrast to BEDASSLE, Geneland and most other existing methods, which all heuristically describe the strength of the parameter, the inference scheme introduced here fits an explicit spatial population genetics model. It corresponds to Nagylaki's γ (Nagylaki, 1978), whose inverse $1/\gamma$ is equal to Barton's barrier strength B (Barton and Bengtsson, 1986). This correspondence makes the inferred barrier strength parameter γ interpretable directly in terms of population genetic theory. For instance, the parameter B/σ^2 , which has dimension of time, must be large to retard the spread of even a neutral allele (Barton and Bengtsson, 1986).

Our method can reliably estimate the presence of strong barrier ($\gamma \lesssim 0.1$), but there

is little power to distinguish between a weak and no barrier (Fig. 3.5 and Fig. 3.6). The reason is not a shortcoming of our inference scheme, but the fact that relatively weak barriers do not significantly affect the spread of ancestry (Fig. 3.2). Therefore, it is infeasible to estimate the strength of weak barriers to gene flow, simply because they do not have a significant effect on allele frequency covariances. This effect was already observed by Barton et al. (2013), who found that in two spatial dimensions the effect of a barrier starts to have appreciable effects on the spatial pattern of genetic marker alleles when barrier strength B (which corresponds to $1/\gamma$) is $\approx \sigma$.

The exact power of the inference method depends on a range of factors. As a barrier mostly affects nearby allele frequency fluctuations, a high sampling density spread evenly on both sides of a putative barrier is ideal (Fig. 3.8). Our method is data hungry, as it needs at least a few dozen SNP markers and at least hundreds of sampled individuals (Supp. Text 2). This wealth of data is required since recent pairwise coalescence events, which constitute the signal for our inference scheme, are rare events in most realistic scenarios. As a rough rule of thumb, our inference scheme can be applied in cases in which there is sufficient power to detect an overall isolation by distance pattern.

Any realistic scenario has its own set of parameters and specific sampling scheme. Our power simulations can only test a tiny fraction of possible combinations of these. They helped to elucidate general underlying patterns, but cannot cover all specific cases. Therefore, we recommend to do customized power simulations. Simulating the specific sampling scheme, the marker number together with likely demographic parameters will help to determine whether the inference scheme has sufficient power to detect putative barriers for the specific scenario of interest.

3.4 Outlook

The inference scheme introduced here fits a linear barrier, the most straight-forward model for a barrier in two dimensions. We used an analytical formula (Eq. 3.5) to model the spread of ancestry, which in turn allows one to reduce calculations for pairwise F to a single numerical integral (Eq. 3.7). However, in practice barriers might be geographically more complex. There could also be multiple barriers in different locations - which would be partly indicated by our method, but also invalidate its underlying model of a single barrier. Such more complicated scenarios will most likely not allow for simple formulas, and calculations for the chance of recent co-ancestry become much more challenging. One could trace the geographic ancestry distribution back with dis-

cretized simulations, and use them to first calculate the expected distributions of recent pairwise coalescence (Eq. 3.6) and consequently identity by descent patterns (Eq. 3.7). This salient extension to our model poses a numerical challenge, but it seems to be within reach of present day computational power.

Our method fits allele frequencies, which can be confounded by deeper ancestral patterns. By filtering loci that show large scale geographic variation, one can in principle remove some of this ancestral genetic structure, but by doing so one might accidentally remove true signal as well. This problem can be a severely confounding factor when applying the inference method to scenarios where ancestral structure is present, for instance zones of putative secondary contact.

One promising way to overcome this problem would be to base inference on identity by descent blocks, the direct genetic traces of recent co-ancestry (Browning and Browning, 2012). As blocks of ancestral genetic material are split up at a constant rate by recombination, the probability of sharing a block of length l decays exponentially back in time (Ralph and Coop, 2013). Therefore, blocks longer than 5 cM, say, are very unlikely to originate from co-ancestry older than 100 generations, even under relatively extreme demographic scenarios. Moreover, the length of the blocks contains information about the time of coalescent. Identifying such blocks is a non-trivial task, in particular when only un-phased genotype data is available (Browning and Browning, 2012). It requires dense genotype data and linkage information. But in cases where IBD blocks can be robustly called - as already possible for humans and some model organisms - an inference scheme based on this signal holds great potential. Our method to model the spread of ancestry can be combined with formulas for block sharing (Ralph and Coop, 2013; Ringbauer et al., 2017a) to calculate the expected number of shared IBD blocks in presence of a barrier. These results could be used to fit observed block sharing data.

Summarizing, our method is only a first step to robustly infer barriers to gene flow from genotype data. The techniques outlined here can be expanded in various directions to better deal with the complexities of real data, and to make full use of opportunities within the era of population genomics. We hope that this will ultimately lead to a better understanding of barriers to gene flow within many natural populations.

3.5 Appendix

Here we give the full formula we fit, and describe the rescaling to a set of independent effective parameters. Let x_1, x_2 denote the x -coordinate of the samples, $\Delta x (= x_1 - x_2)$ and Δy their separation along each axis. For the identity by state on different sides of the barrier, plugging into formula Eq. 3.7 gives for pairwise F :

$$\int_0^\infty \frac{1}{2D_e} \underbrace{\exp\left(\frac{4\kappa}{\sigma^2}(\Delta x + 4kt)\right) \frac{2\kappa}{\sigma^2} \operatorname{erfc}\left(\frac{\Delta x + 8\kappa t}{\sqrt{4\sigma^2 t}}\right)}_{\text{Movement x-axis}} \underbrace{\frac{1}{\sqrt{4\pi\sigma^2 t}} \exp\left(-\frac{(\Delta y)^2}{4\sigma^2 t}\right)}_{\text{Movement y-axis}} \underbrace{\exp(-2\mu t)}_{\text{L.D. migration}} dt$$

We now rescale time, such that $t' = \sigma^2 t$. The integral ($dt' = \sigma^2 dt$) transforms to:

$$\int_0^\infty \frac{\sqrt{\pi}}{\sqrt{t'}} \frac{2\kappa}{4\pi D_e \sigma^4} \exp\left(\frac{4\kappa}{\sigma^2}(\Delta x + \frac{4k}{\sigma^2} t')\right) \operatorname{erfc}\left(\frac{\Delta x + 4\frac{2\kappa}{\sigma^2} t'}{2\sqrt{t'}}\right) \exp\left(-\frac{(\Delta y)^2}{4t'}\right) \exp\left(-\frac{2\mu}{\sigma^2} t'\right) dt'$$

Defining $\text{Nbh} := 4\pi D_e \sigma^2$, $\gamma := \frac{2\kappa}{\sigma^2}$ and $m = \frac{2\mu}{\sigma^2}$ gives the full formula used for inference:

$$\int_0^\infty \frac{1}{\text{Nbh}} \frac{1}{\sqrt{t'}} \exp(2\gamma(\Delta x + 2\gamma t')) \gamma \operatorname{erfc}\left(\frac{\Delta x + 4\gamma t'}{2\sqrt{t'}}\right) \exp\left(-\frac{(\Delta y)^2}{4t'}\right) \exp(-mt) dt'$$

The formula for same sides of the barrier is rescaled analogously. The additional term in the integrand becomes:

$$\frac{1}{\text{Nbh}} \frac{1}{2t} \left\{ \exp\left(-\frac{(x_1 - x_2)^2}{4t'}\right) + \exp\left(\frac{(x_1 + x_2)^2}{4t'}\right) \right\} \exp\left(-\frac{(\Delta y)^2}{4t'}\right) \exp(-mt)$$

3.6 Supplementary Material

3.6.1 Supplementary Information 1: Fitting Allele Frequencies

In the main text, we calculate the chance of pairwise identity by state F , which we defined as the probability of coalescence before a long distance or mutation event. These chances of recent co-ancestry can be only indirectly observed as covariances of allele frequencies, and it is not immediately clear how to best fit these results to observed data. The fitting is also further complicated by the fact that mean allele frequencies are usually unknown. Fitting all of them would introduce many new parameters beyond the small number of demographic parameters and likely lead to over-fitting of the data, as naive allele frequency estimates will be biased towards the most common direction of the allele frequency fluctuations. There are different possible approaches to deal with these problems. We decided to implement and test three different ways to fit allele frequency fluctuations based on our model. Their full implementations are available on the github repository: <https://github.com/hringbauer/BarrierInferPublic.git>

Method 1: Gaussian Random Field Method

In Computer Science and Machine Learning, the so called Gaussian Random Field method is widely used to fit covariance structures (Rasmussen and Williams, 2006). The goal is to fit the covariances themselves, and then use this fit to make new predictions based on the fitted covariance structure. Here, we adapt this method to fit allele frequency covariances. A similar approach has been recently used by Bradburd et al. (2013).

Summarizing briefly, in the Gaussian Random Field method the observed data y_i are modeled to depend on known parameters \vec{x}_i and to covary depending on these known parameters. The covariances affect latent, unobserved variables f_i . These unobserved variables are drawn from a multivariate normal distribution with some mean m and covariance matrix $K : f \sim N(m, K)$. The entries K_{ij} of the covariance matrix depend on x_i and x_j , and a set of so called hyper parameters θ :

$$K_{ij} = K(x_i, x_j, \theta)$$

The Gaussian Random Field approach utilizes the fact that a multivariate Gaussian distribution is fully determined by its mean and its covariance matrix. It therefore possible to write down a full likelihood of the observed data given the covariance matrix, by integrating over all latent variables f :

$$L(y, \theta) = \int_{\mathbf{f}} P(y | \mathbf{f}) P(\mathbf{f} | K(x, \theta)) d\mathbf{f} \quad (3.12)$$

If one assumes that the data y are drawn as a Gaussian around the latent variable, this integral can be solved analytically due to convenient properties of Gaussian probability densities. One can then easily calculate the marginal likelihood of the observed data, and fit the hyper-parameters via maximizing this likelihood.

If the observations are restricted to binary discrete values (w.l.o.g. 0 or 1), it is still possible to apply the Gaussian Random field model. One typically transforms the latent variables f using a so called link function $p(f)$ to take values p_i between 0 and 1 (most commonly the logit or the probit function), and then models the discrete observed values y_i to be drawn binomially with mean p_i . However, integral 3.12 cannot be solved analytically anymore. As it is very high dimensional, direct numerical integration is also computationally infeasible. Therefore, several analytical approximations to 3.12 are widely used (Nickisch and Rasmussen, 2008). For genotype data, we decided to utilize a custom implementation of the Laplace method (Rasmussen and Williams, 2006). This widely used approach is based on a second order Taylor approximation around the most likely latent variables f_i . These and the Hessian are found numerically, and these calculations can be done relatively fast. Using this approximation allows for an analytical approximation of the total likelihood. A full description of the method can be found in Rasmussen and Williams (2006).

Genetic data with geographic information consists of discrete genotypes y_i sampled at positions \vec{x}_i . W.l.o.g. biallelic markers have values 0 or 1. Diploids can be split up into two haploid genotype data points. The Gaussian Random Field method can then be adapted to fit covariance structure within such data, but one has to deal with some peculiarities. Importantly, the magnitude of the allele frequency covariances depends on the mean allele frequency \bar{p} :

$$\text{Cov}(y_i, y_j) = \bar{p}(1 - \bar{p}) F(x_i, x_j), \quad (3.13)$$

while our model predicts the $F(x_i, x_j)$. In order to account for the additional terms, we introduce a custom link function. We utilize the inverse Fisher's angular transformation of allele frequency (Fisher et al., 1947):

$$p(f) := \sin^2\left(\frac{f}{2}\right)$$

This is a valid link function, as its image is confined to the interval $[0, 1]$. Its usefulness stems from the fact that it solves the following differential equation:

$$p'(f) = \sqrt{p(f)(1-p(f))}.$$

Allele frequency fluctuations are usually small, and a first order approximation $p(f) \approx p(f_0) + p'(f_0)\Delta f$ yields:

$$\text{Cov}((p(f_1), p(f_2))) \approx p'(f_0)^2 \text{Cov}(f_1, f_2) = p_0(1-p_0) \text{Cov}(f_1, f_2)$$

Comparing with Eq. 3.13 shows that this link function together with the F -Matrix as Covariance kernel model the covariance structure of discrete genotypes. As we can directly calculate $F(x_i, x_j)$ with our model, this approach can be used to fit the demographic parameters θ to the data.

To deal with the problem of over-fitting by estimating a potentially large number of mean allele frequencies, we adapted the Gaussian Random Field approach. Unknown allele frequencies are not estimated directly, but only the variance of the unknown distribution of mean allele frequency: We model that mean latent variables are randomly drawn from a distribution with Variance σ^2 around some overall mean and that then the latent variables f_i are drawn with covariance matrix K around this means. If the means are drawn from a normal distribution, the overall covariance will also be distributed as a multivariate normal distribution:

$$f \sim N(0, K + J\sigma^2)$$

where J denotes the unit matrix, whose entries are all 1. Using this approach, we can fit the effects of unknown distribution of mean allele frequencies as a single hyperparameter of the covariance matrix. For multiple, independent (unlinked) genotypes, the marginal likelihoods can be multiplied.

Summarizing, the Gaussian Random Field approach allows us to calculate an approximate marginal likelihood of genotype data given the expected co-ancestry structure F based on some demographic hyper parameters θ . Using standard methods to maximize likelihoods, we can find maximum likelihood estimates of these θ . After experimenting with several methods, we found that the standard Nelder-Mead method works very reliably, and used it in all our implementations.

Our approach has two sources of error. First, it is not immediately clear how accurate the Laplace approximation is for genotype data, in particular since allele frequency correlations are typically weak. Second, allele frequency data will not be always distributed as a multivariate Gaussian. Under the model of diffusion of ancestry, there

will also be higher order moments. For instance, having recent co-ancestry with individuals in one geographic direction makes it less likely to have shared co-ancestry with individuals from the opposite direction, and this effect is not captured well by the Gaussian Random Field model. Calculating these higher order moments would go far beyond the pairwise diffusion model that we outline in this work. However, the multivariate Gaussian approximation can be expected to be an accurate approximation as long as fluctuations remain small (Barton et al., 2013).

Method 2: Maximizing Pairwise Likelihoods

This method maximizes the likelihood of observing the three states of pairwise genotypes. Given two markers, there are four possible states: 00, 10, 01 and 11. Using our calculations for the co-ancestry matrix F , it is straightforward to write down the probability for each of these for states. Denoting the mean allele frequency of marker 1 by p and marker 0 by q :

$$\begin{aligned} P(00) &= F \cdot p + (1 - F) \cdot p^2 \\ P(10) &= P(01) = (1 - F) \cdot p \cdot q \\ P(11) &= F \cdot q + (1 - F) \cdot q^2. \end{aligned}$$

As the mean allele frequency is usually unknown, we integrate over their unknown distribution:

$$\begin{aligned} P(00) &= F \cdot \bar{p} + (1 - F) \cdot (\text{Var}(p) + \bar{p}^2) \\ P(10) &= P(01) = (1 - F) \cdot (\bar{p} - \text{Var}(p) - \bar{p}^2) \\ P(11) &= F \cdot \bar{q} + (1 - F) \cdot (\text{Var}(q) + \bar{q}^2) \end{aligned}$$

This approach introduces one additional parameter: $v := \text{Var}(p) = \text{Var}(q)$. By multiplying all pairwise likelihoods one gets at a composite likelihood that depends on the demographic parameters θ and the variance parameter v . These pairwise likelihoods are not independent - realized co-ancestry with one individual also increases the probability of co-ancestry with other individuals near the related one. Therefore, multiplying pairwise likelihoods does not yield the total likelihood of the observed data. However, this composite likelihood should be seen as a way to fit the data, and this approach will give consistent parameter estimates in the limit of large amounts of sufficiently uncorrelated data. We implemented the maximization of this likelihood by using the `GenericLikelihoodModel` class of the Python package `statsmodels`.

Method 3: Pairwise Homozygosity

One can also fit identity-by-descent probabilities F based on the signal of pairwise homozygosity. As stated in the main text, the chance of pairwise homozygosity for a single marker is given by:

$$h = F + (1 - F) \cdot (\bar{p}^2 + (1 - \bar{p}^2))$$

Summing over all markers gives the expected fraction of pairwise homozygotes:

$$\mathbb{E}(h) = F + (1 - F) \cdot \underbrace{\sum_{p_i} \Pr(p_i) \cdot (p_i^2 + (1 - p_i)^2)}_{:=s}$$

In order to fit this signal, we minimize the sum of squared difference between the expected and observed pairwise homozygosity for all pairs:

$$\bar{\theta} = \min_{\theta} \sum_{k < l} (\bar{h}_{kl}(\theta, s) - h_{kl})^2$$

In our implementation we use the method `curvefit` from the Python package `Scipy`.

Performance on simulated Data

To test which method performs best in scenarios with realistic parameters, we tested them on simulated data sets. We used the simulation scheme outlined in the main text to generate data with known demographic parameters, and applied the three methods described above. We first simulated and fitted scenarios without a barrier, in order to test the general capability of the methods to accurately fit allele frequency fluctuations. The outcome is visualized in Figure Fig. 3.11. Our results show that the Gaussian Random Field method (Method 1) has a significant downward bias when estimating the neighborhood size, whereas the pairwise likelihood and pairwise homozygosity method are approximately unbiased. Our results also indicate that these two inference methods produce highly correlated estimates and have similar estimation variances. We also found that using the pairwise homozygosity method is a factor of 10 quicker than using the pairwise likelihood method.

Limited Number of Loci and Individuals

Different methods are expected to perform differently when information is limited. Therefore, we tested the three methods on datasets with a varying amount of data. We

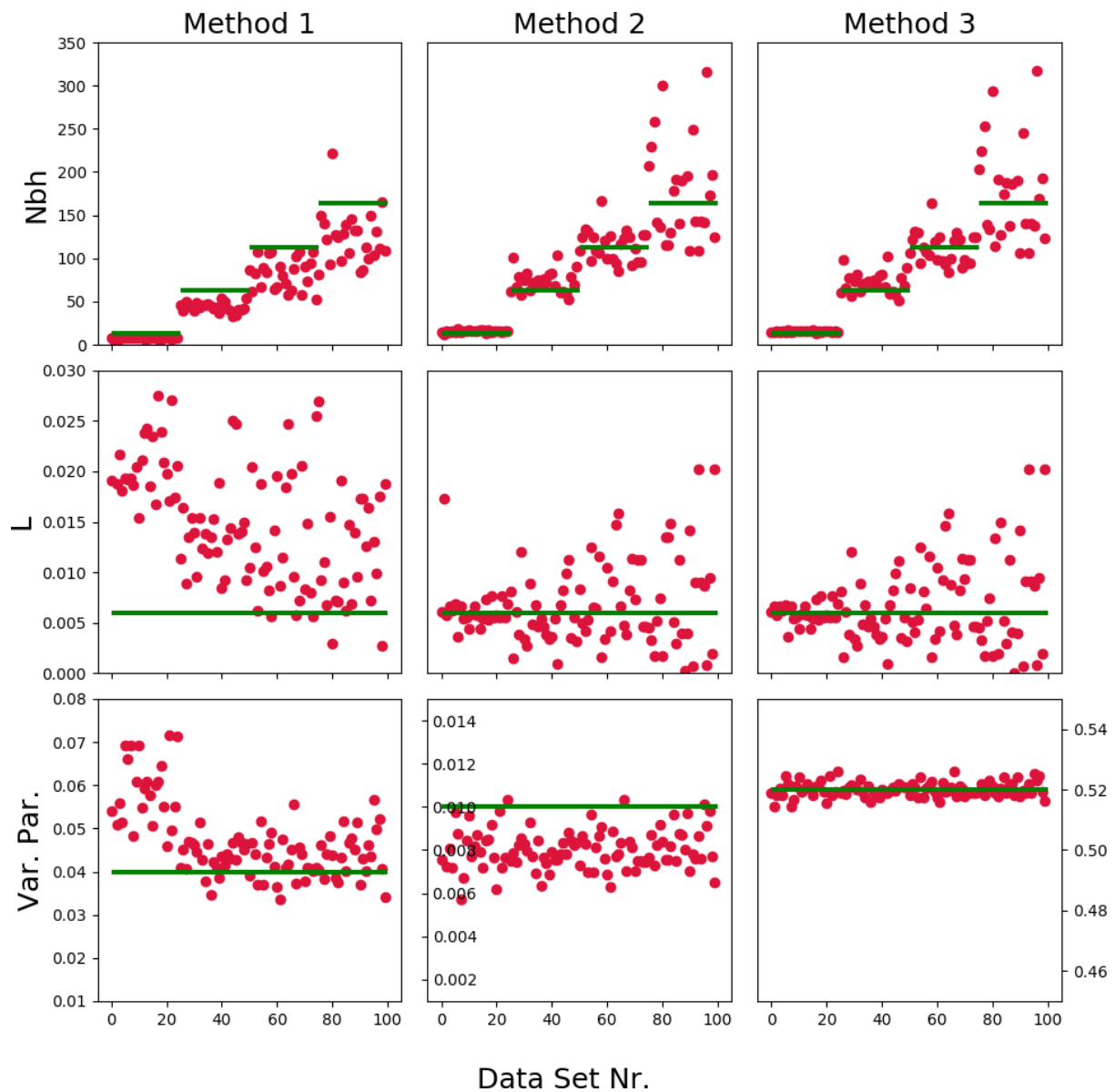


Figure 3.11: Fit to allele frequency fluctuations: We tested the three methods on synthetic datasets generated using the method described in the main text. The datasets consist of 1000 haploid individuals, situated on a grid of 50×20 individuals spaced 2 dispersal units apart along each axis, with genotype information for 200 loci. We simulated 25 replicate data sets for four different neighborhood sizes (and consequently different magnitudes of allele frequency fluctuations). Throughout, $m = 0.006$ and a random distribution of mean allele frequencies with $\sigma(p) = 0.1$. The lower row depicts the estimates for the Variance Parameter that fits the fluctuations of mean allele frequencies, whose true value is different for all three used models.

simulated two types of data sets: One with a varying number of loci, and one with a varying number of individuals. Our results are visualized in Fig. 3.12 and Fig. 3.13.

Interestingly, the Gaussian Random Field method remains biased when the number of loci increases; however this bias vanishes with increasing number of individuals. The estimator variance of the other two methods decreases slowly with increasing information. However, neither increasing the number of loci nor increasing the number of individuals seem to yield dramatic increases in estimation accuracy.

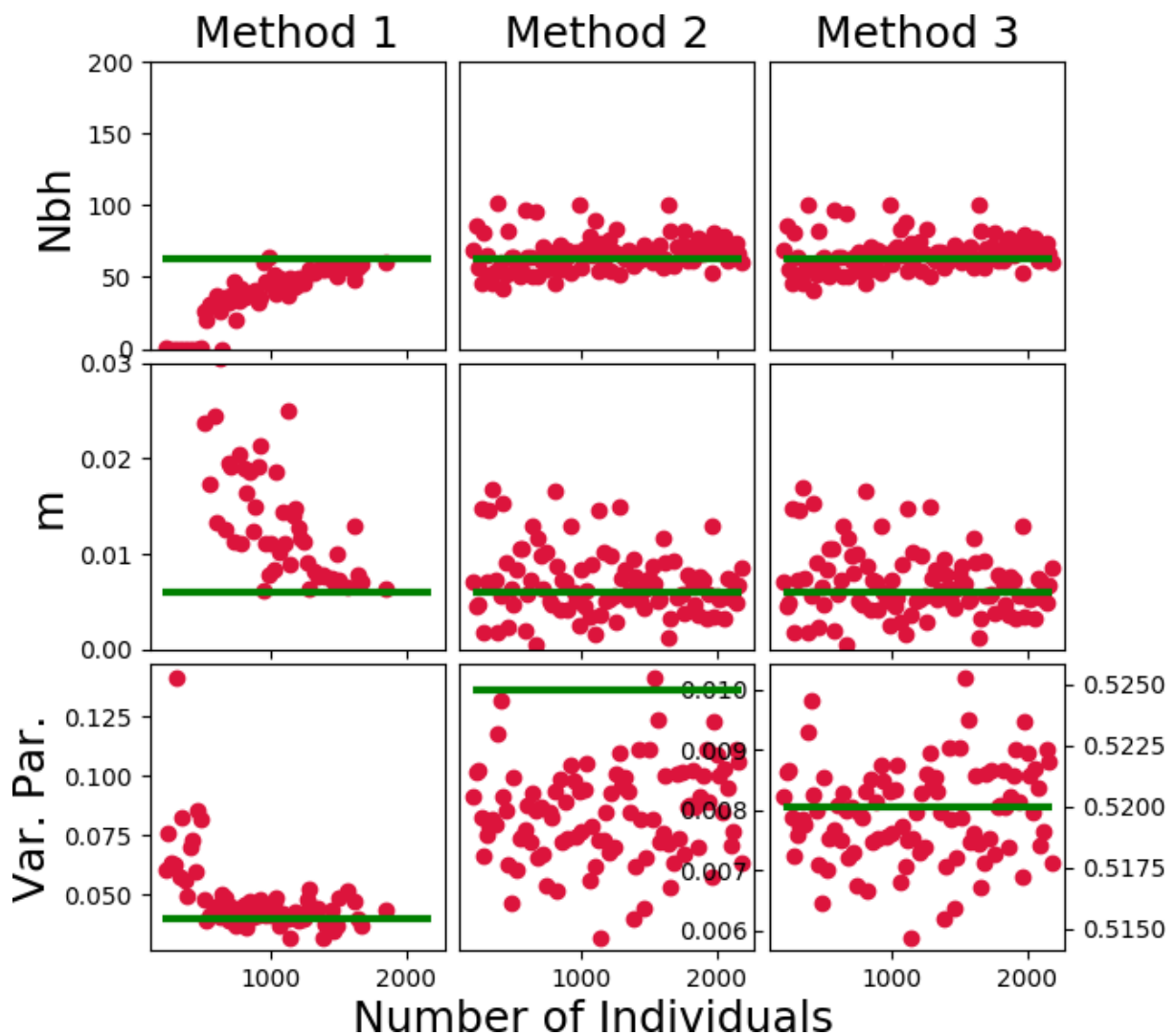


Figure 3.12: Testing the methods on datasets with varying numbers of individuals. We randomly sub-sampled the target number (200–2200 individuals) from a grid of 100×40 haploid individuals spaced 1 dispersal units apart along each axis. We simulated independent data sets with genotype information for 200 loci ($m = 0.006$, $Nbh = 4\pi 5 \approx 62.83$ and $\sigma(p) = 0.1$)

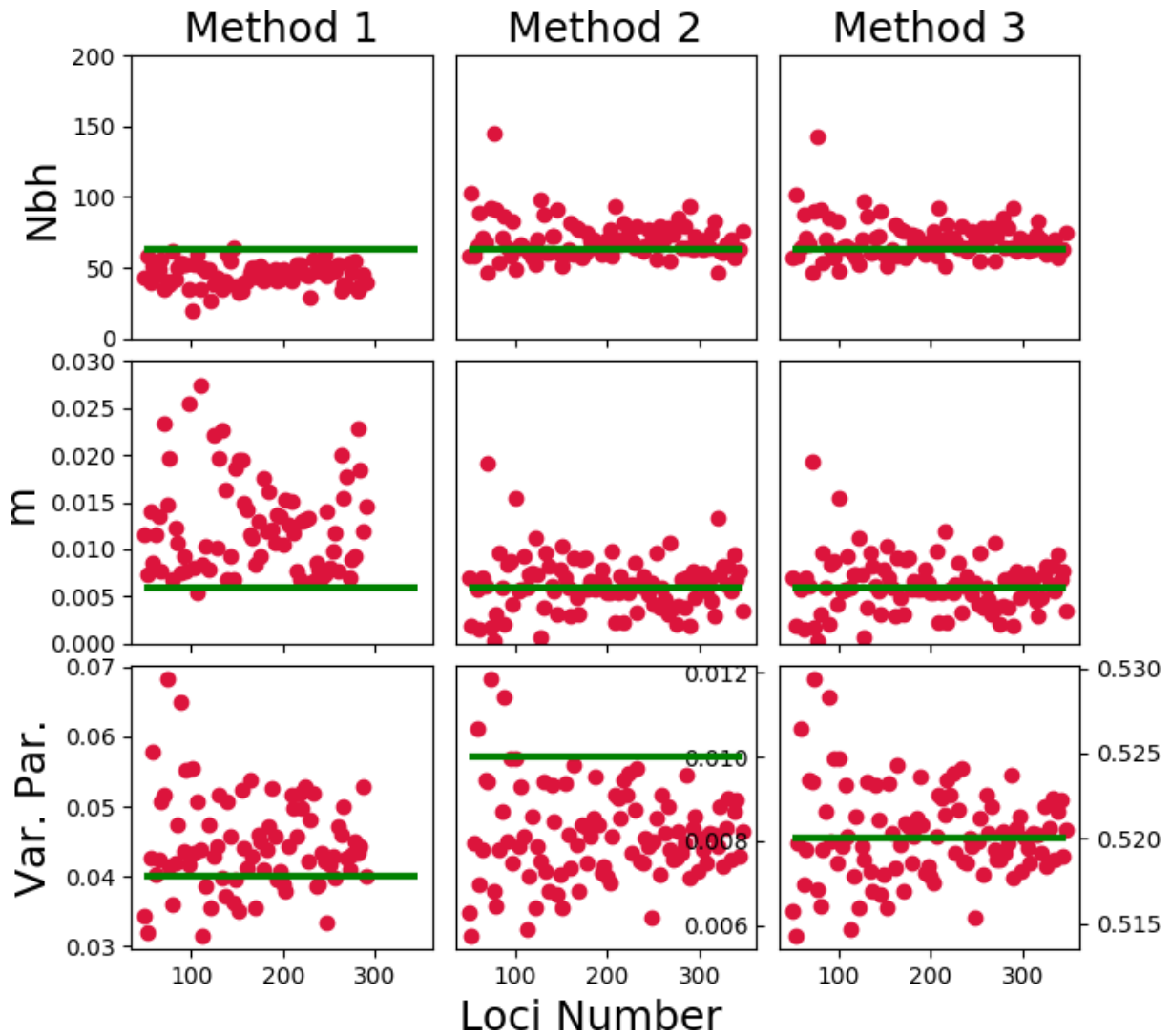


Figure 3.13: Testing the methods on datasets with varying number of loci. We simulated independent data sets with varying number of independent loci (50-350). Haploid individuals were spaced on a grid of 50×20 with a spacing of 1 dispersal unit along each axis. ($m = 0.006$, $Nbh = 4\pi 5 \approx 62.83$ and $\sigma(p) = 0.1$)

Fitting barriers with limited data

One can fit the barrier strength parameter γ while keeping the other parameters fixed. This can be for instance useful if one wants to test the hypothesis of a barrier at a specific subset of loci. One can then estimate the demographic parameters using all loci, and proceed to fit γ based only on the subset of markers.

We therefore tested this approach. We find that fitting the barrier strength alone does not markedly improve inference for estimating γ , at least in the tested scenario, in which there is sufficient information to accurately fit the isolation by distance pattern (Fig. 3.14).

Our results also indicate that even with 2400 individuals and a strong barrier ($\gamma = 0.05$), one would need at least a few dozen independent biallelic markers to reliably estimate a strong barrier. The required number of markers and individuals for a given scenario will of course depend on the exact sampling scheme as well as the strength and shape of isolation by distance in the data.

Binning Individuals into Demes

Method 3 can be used to analyze deme data, as outlined in the main text. Binning into demes of k individuals each speeds up calculations by a factor of k^2 , as all pairwise comparisons for individuals between two demes reduce to a single comparison. On the other hand, binning nearby individuals is not expected to have a big effect on the inference scheme, as only information for pairs within demes is lost. To confirm this intuition we tested our method on simulated data (Fig. 3.15). Our results indicate that small scale binning (with bins are extended up to a few dispersal units) does not have a major effect on the parameter estimates. The variance of the inferred parameters increases as expected, but this increase is slow.

Conclusion

The application to simulate data indicates that the methods based on fitting pairwise statistics (Method 2 and Method 3) are more accurate and less biased than the computationally more elaborate Gaussian Random Field approach (Method 1). As outlined above, the latter suffers from two potential errors: The Laplace approximation and also the multivariate Gaussian approximation could be inaccurate for the spatial covariance patterns typically observed in genotype data. Our datasets were simulated under an explicit population genetics model (see main text) with parameters chosen to match

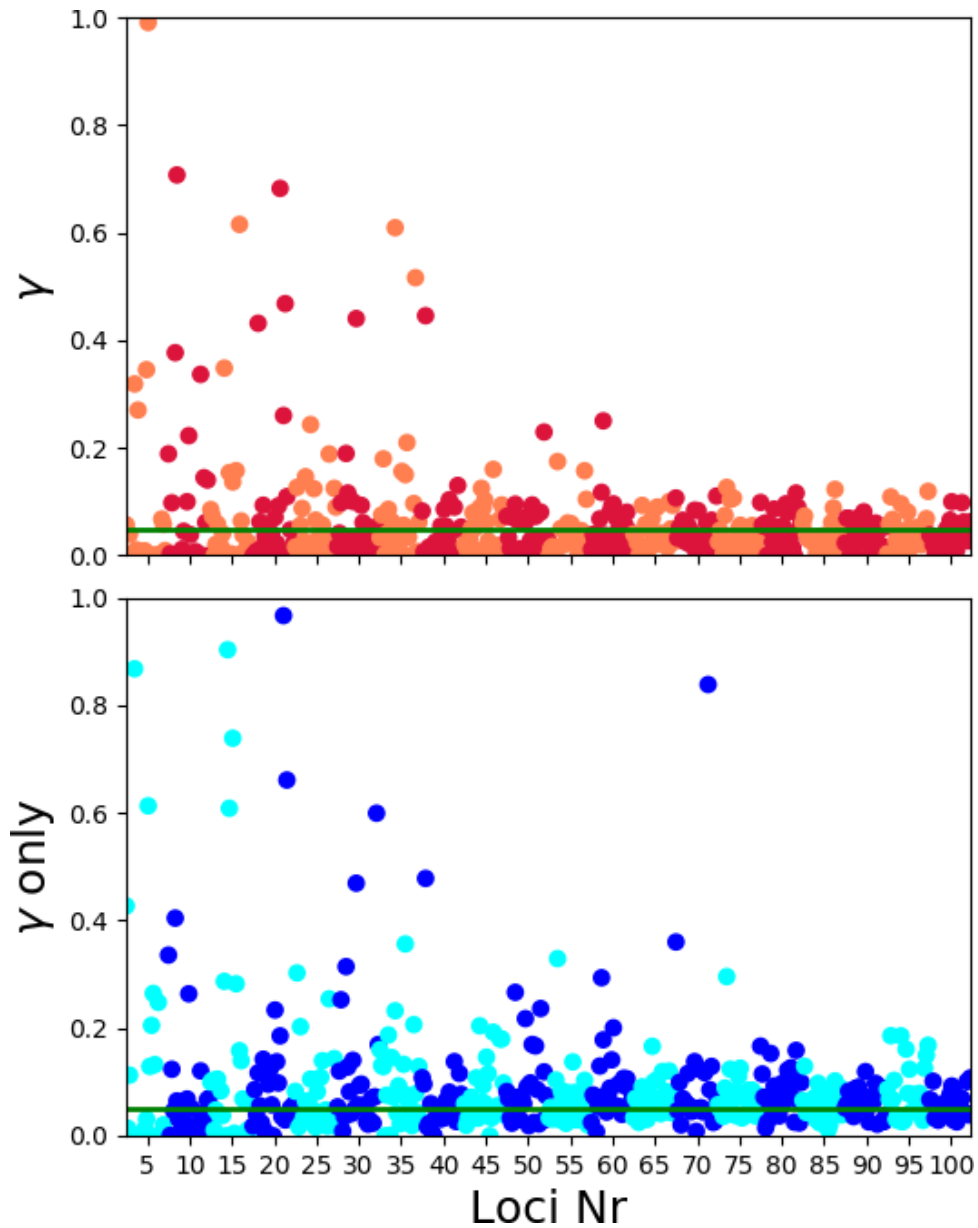


Figure 3.14: Fitting only the barrier parameter: We simulated a dataset of 60×40 individuals spaced on a grid with step size 1σ ($m = 0.006$, $Nbh = 4\pi 5 \approx 62.83$ and $\sigma(p) = 0.1$) and a strong barrier at the middle of the x-axis with $\gamma = 0.05$). We simulated 25 replicates for different loci numbers (5, 10, ..., 100). We applied Method 3 to fit the barrier strength γ by estimating all parameters (upper figure) and to fit only the barrier strength and the fluctuation parameter, with the demographic parameters fixed to their true value (lower figure).

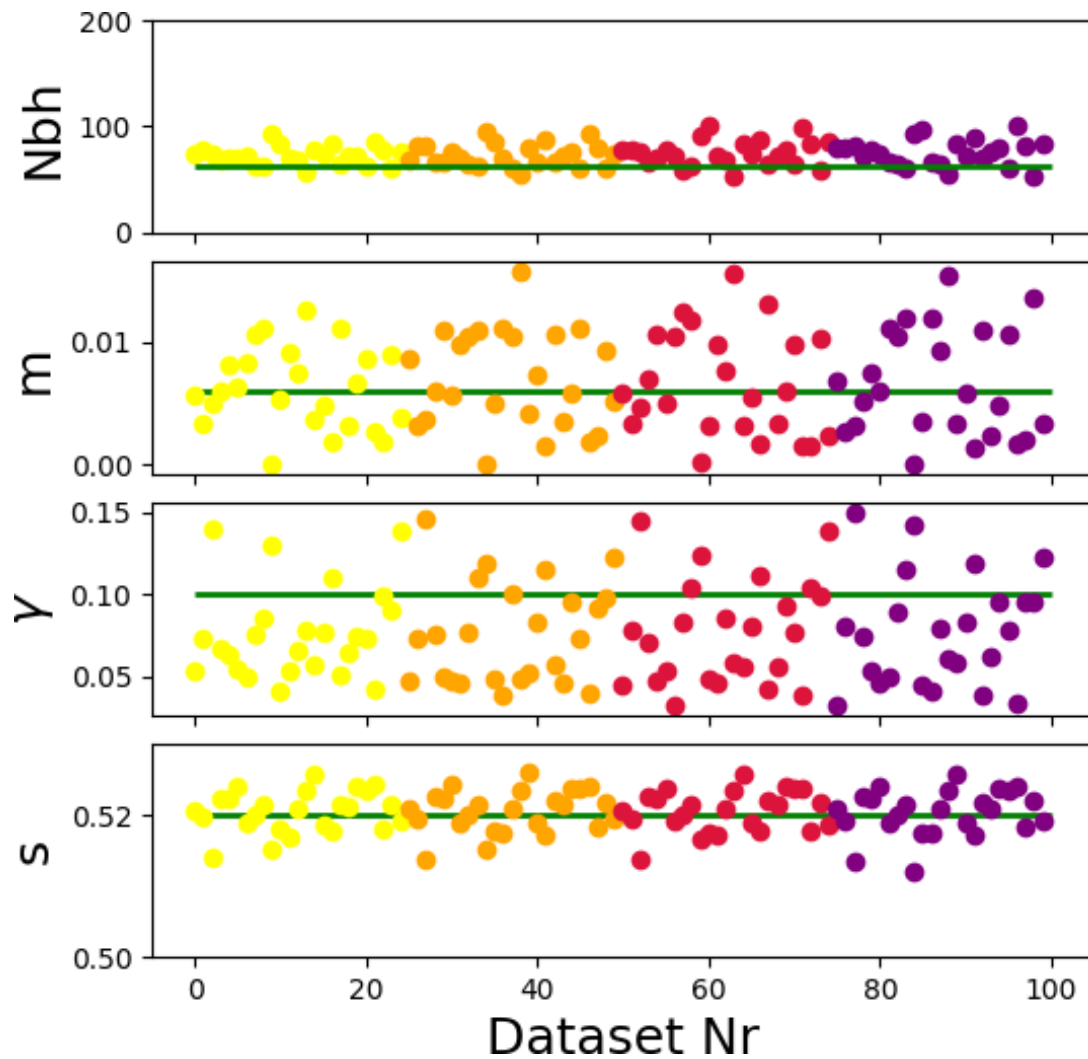


Figure 3.15: Various degrees of binning: We simulated 25 datasets of 60×40 individuals spaced one dispersal unit apart ($m = 0.006$, $Nbh = 4\pi^5 \approx 62.83$ and $\sigma(p) = 0.1$ and a barrier at the middle of the x-axis with $\gamma = 0.1$). For analysis, we binned individual data into demes of 1×1 (yellow), 2×2 (orange), 3×3 (red), 4×4 (purple) individuals. We used Method 3 to analyze the resulting data sets.

typical isolation by distance patterns, and our findings imply that the Gaussian Random Field approach with a Laplace approximation is not suited well to fit those, unless the number of sampled individuals gets very high (Fig. 3.12).

Method 3, which is based on fitting pairwise homozygosity, is additionally faster by a factor of about 10 than Method 2. Our results indicate that it can be used on binned data, without much loss of accuracy. Therefore, we decided to base inference in the main text on this method.

3.6.2 Supplementary Information 2: Comparison to Geneland

A widely used program to detect barriers to gene flow based on spatially explicit data is Geneland (Guillot et al., 2005). In a recent study that compares different methods to detect barriers, it was identified as one of the most potent methods (Safner et al., 2011).

Here we test Geneland on two kinds of simulated data sets. One of them fulfills the model assumptions of Geneland, i.e. a barrier but not further substructure of the populations on either side of the barrier. For the second scenario, we simulated a scenario of a barrier with additional isolation by distance.

Application Details

In all analysis, we ran geneland with an MCMC chain length of 10^5 . For analysis, we used a thinning of 100 and a burn-in of 200, and visually inspected summary statistics to ensure proper convergence. We usually used a fixed population number $K = 2$, and investigated whether Geneland can accurately cluster the two subpopulations on each side of the barrier.

Scenario I: Two Panmictic Populations

In this scenario, we simulated data-sets that met the model assumptions of Geneland. Two equally sized populations on both sides of the barrier were assumed to be panmictic units, i.e. for all individuals alleles were binomially drawn with means p_l and p_r . We simulated datasets of 400 diploid individuals with genotype information for 200 biallelic loci spaced on 20×20 grid. The overall mean allele frequencies for the individuals left and right were randomly drawn:

$$p_l = 0.5 + \Delta p$$

$$p_r = 0.5 + \Delta p + \Delta p_r,$$

where Δp and Δp_r are random normal variables with standard deviation $\sigma = 0.1$.

We simulated 10 replicates. In all of them, Geneland was able to accurately infer the position of the barrier and assign all individuals correctly (see 3.16).

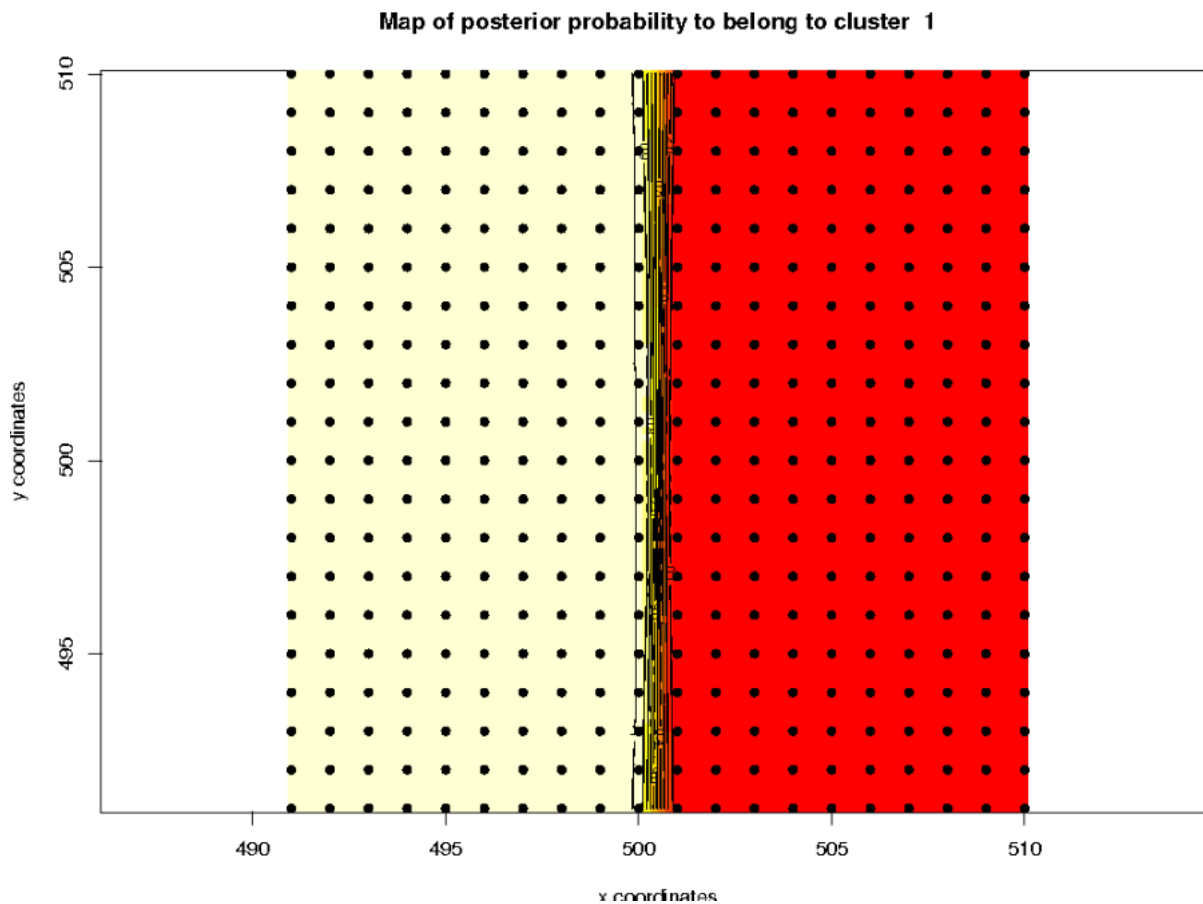


Figure 3.16: Geneland with no isolation by distance: This picture visualizes the typical outcome when Geneland is applied to a model with a barrier but no further substructure (see main text). The figure visualizes posterior probability of population membership. Geneland is able to accurately infer the subdivision of the population into two subpopulations.

Scenario II: Barrier with Isolation by Distance

Second, we applied Geneland on data-sets that we have generated using our explicit spatial population genetics simulations. We simulated 10 replicates of a datasets of 400 individuals for each a complete ($\gamma = 0$) and a weak barrier ($\gamma = 0.1$) with moderate isolation by distance ($m = 0.006$, $Nbh = 4\pi 5 \approx 62.83$ and $\sigma(p) = 0.1$). As expected, an isolation by distance pattern can be observed in this data (Fig. 3.17). In all 10 data-sets, Geneland fails to accurately estimate the barrier, but rather infers 2 patchily distributed

subpopulations (Fig. 3.18). It also cannot infer the barrier if the population number K is not fixed, but inferred as well (Results not shown).

In stark contrast, our method is able to infer the existence and the strength of a barrier in these scenarios (Fig. 3.19).

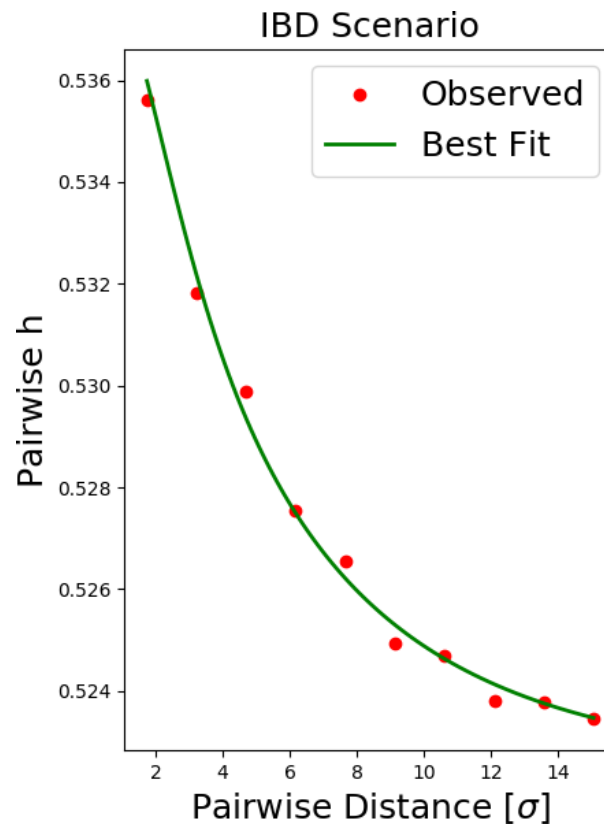


Figure 3.17: Isolation by Distance scenario: The decay of pairwise homozygosity in one dataset simulated under the Isolation by Distance scenarios.

Conclusion

Our results show that Geneland is a powerful tool to infer a barrier when its model assumptions are met (i.e. population is structured into 2 subpopulations without further substructure). However, as observed previously (Safner et al., 2011), it fails in the scenario with additional isolation by distance, which we simulated under an explicit population genetics model. Our findings indicate that caution is warranted when applying Geneland to datasets with isolation by distance patterns. In particular, when the scale of isolation by distance observed on scales smaller than the geographic extension of the subpopulations, Geneland will have very limited power to detect a barrier. In contrast, our method works well in these cases. It can therefore be seen as a complementary approach to Geneland.

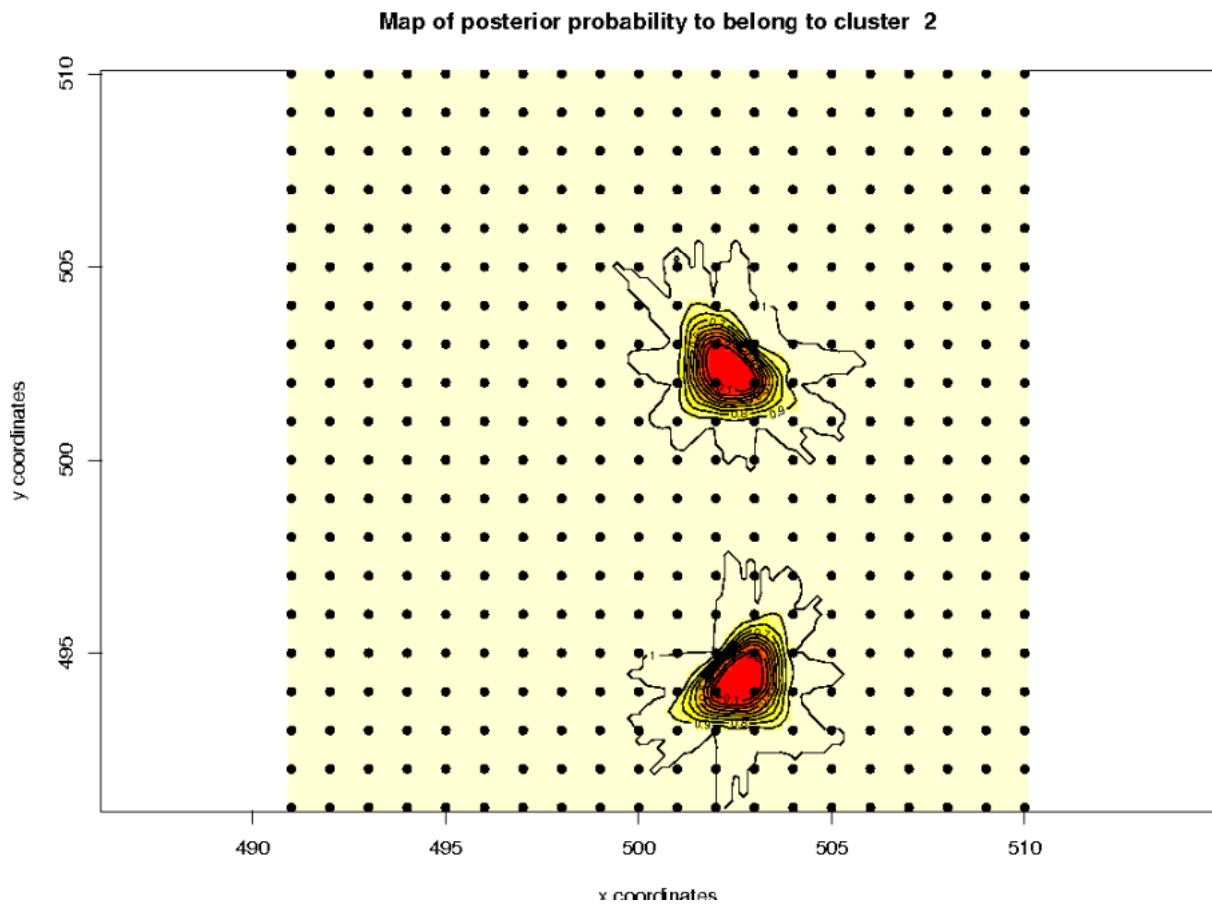


Figure 3.18: Geneland with isolation by distance: This picture shows a typical output of Geneland when applied to simulated data with a complete barrier and isolation by distance (see main text). The figure visualizes the posterior probability of population membership. Geneland fails to accurately infer the subdivision of the population into two subpopulations.

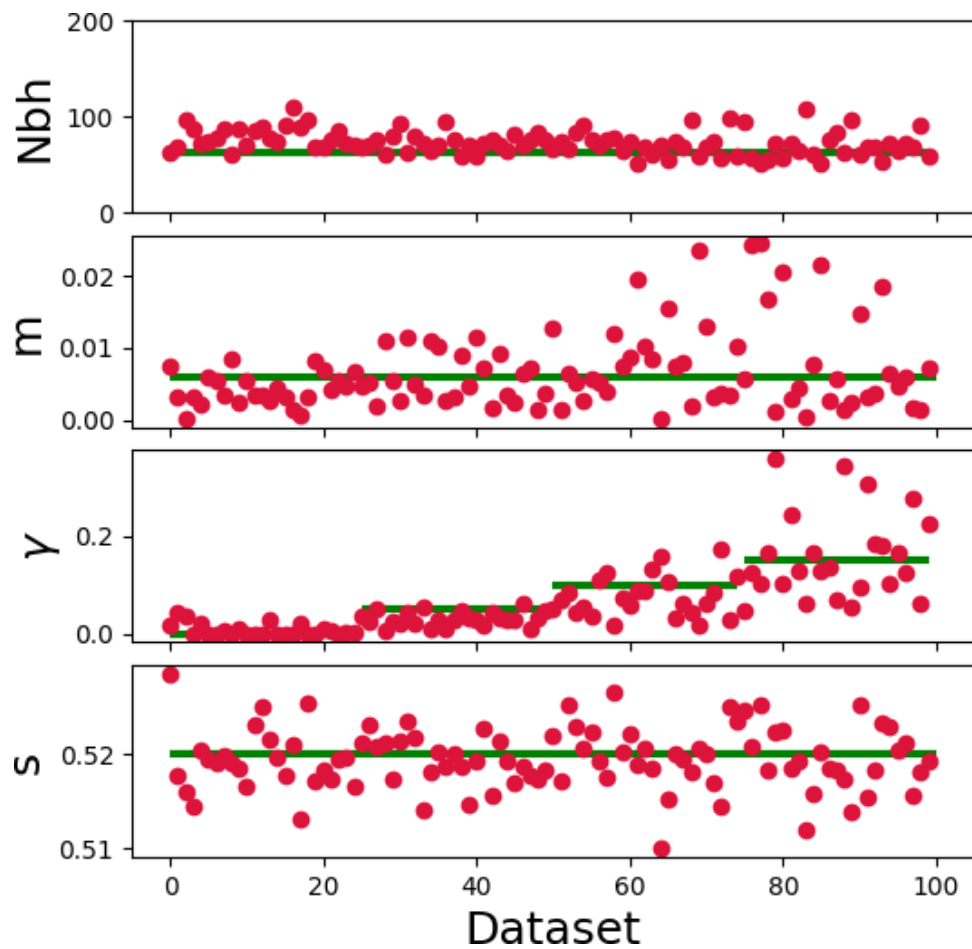


Figure 3.19: Our method on isolation by distance datasets: 25 replicates of four different barrier strengths were simulated $\gamma = 0, 0.05, 0.1, 0.15$. In all datasets Geneland failed to infer the barrier, whereas our method can accurately estimate the strength of the barrier.

3.6.3 Supplementary Information 3: Data Cleaning for Hybrid Zone

In the main text we apply our method to a data set from a hybrid zone between two subspecies of the model plant *Antirrhinum majus*. In the following, we describe sampling, genotyping and filtering criteria we used for selecting data for our analysis.

Sampling and Data Collection

As part of a long-term project examining wild pedigrees and geographic clines, each year from 2009 to 2014 we sampled plants from a hybrid zone between *Antirrhinum majus pseudomajus* and *Antirrhinum majus striatum* in the Spanish Pyrenees located in val di Ribes near the village of Planoles. Individual plants are primarily found within 100 m either side of two roughly parallel roads that run up the valley (Fig. 10). The sampling concentrated on a ≈ 4 km transect encompassing the center of the phenotypic and genetic clines involved in magenta and yellow flower pigmentation (ROSEA; Whibley et al. (2006), and SULFUREA; Desmond et al 2017, respectively¹) and some of the flanking regions in which the pure subspecies mostly reside.

The following data were obtained for each plant:

- A global positioning system (GPS) coordinate
- Leaf material (for DNA extraction)

The collection occurred between May and July, for individuals with open flowers only. Individual's geographic coordinate was collected using a GPS device (Trimble GeoXT datalogger) with a mean accuracy of ≈ 2 m. Four to six fresh leaves from each plant were stored in individual glassine envelope bags, which were placed within a plastic bag containing silica gel (Fisher Scientific) for drying the leaf tissue. Components of the magenta and yellow color of the flowers were scored in the field according to Whibley et al. (2006).

SNP Genotyping

The KASP genotyping platform (LGC genomics) was used to genotype single nucleotide polymorphisms (SNPs) across the *Antirrhinum* genome. In total, we designed ≈ 240 SNP at a subset of polymorphic and divergent loci, but here report just on a subset of 60 polymorphic loci. The remaining markers that clearly violate our model assumptions have been filtered out (see below).

¹ Add citation in final version, this paper is accepted but not published yet

Candidate loci were identified using a draft *A. majus* reference genome (≈ 630 Mb across eight linkage groups; courtesy of Yongbiao Xue, BGI) and allele frequencies obtained from whole-genome Illumina PoolSeq of six pools of $n = 50$ individuals located along a transect through the hybrid zone (unpublished data). All potential SNP loci were identified across the genome with a custom Python script `SNPextract.py` (https://github.com/dfield007/genomics_general) which identified SNPs positions suitable for KASP genotyping platform (LGC Genomics). The script was run with the following parameters: (i) $30 < \text{depth} < 300$ in all pools at the focal SNP (to reduce the probability of false positives and paralogs), (ii) $30 < \text{depth} < 300$ for sequences 50bp upstream and downstream of focal SNP, (iii) < 3 other SNPs within 50bp (to ensure primer efficiency), and (iv) biallelism (a KASP requirement). We also selected loci on the basis of being polymorphic in the hybrid zone ($0.3 < \bar{p} < 0.7$) and selected one locus randomly every couple of mapping units (cM) to maximize marker independence (Figure 3.20). For each candidate, the script extracted the 100bp sequence surrounding each candidate polymorphic site required to design the SNP primers. DNA extractions and SNP genotyping were carried out by LGC Genomics. Replicate DNA extractions and genotyping confirmed relatively low error rates of the KASP platform (mean error rate $< 0.1\%$ per locus).

Data Filtering

Our method requires all individuals to have no missing genotype data. Starting with $n = 13722$ individuals in the core hybrid zone ($\approx \pm 2\text{km}$ around flower color transition), we first removed individuals with more than 8 missing genotypes ($n = 246$). Next, we identified individuals with at most one genotype mismatch, and deleted duplicate individuals ($n = 1087$) to remove intentionally or non-intentionally regentyped plants. For the remaining data ($n = 12389$), we imputed missing genotypes. For this, we first calculated the mean allele frequency per marker averaged over all individuals; and then binomially draw two alleles for missing genotypes at random with the corresponding calculated mean allele frequency. As only a fraction 0.84% of all genotypes had to be imputed, this step does not significantly affect the results of the inference method.

Before applying our method, we filtered markers based on the following 4 criteria:

1. **Geographical Variation:** We removed markers that were correlated to the x or the y coordinate, as such large scale variations could originate from deeper time scales or be the traces of divergent selection or could also be the remnants of secondary contact. We chose a cut-off value of $R^2 = 0.015$.

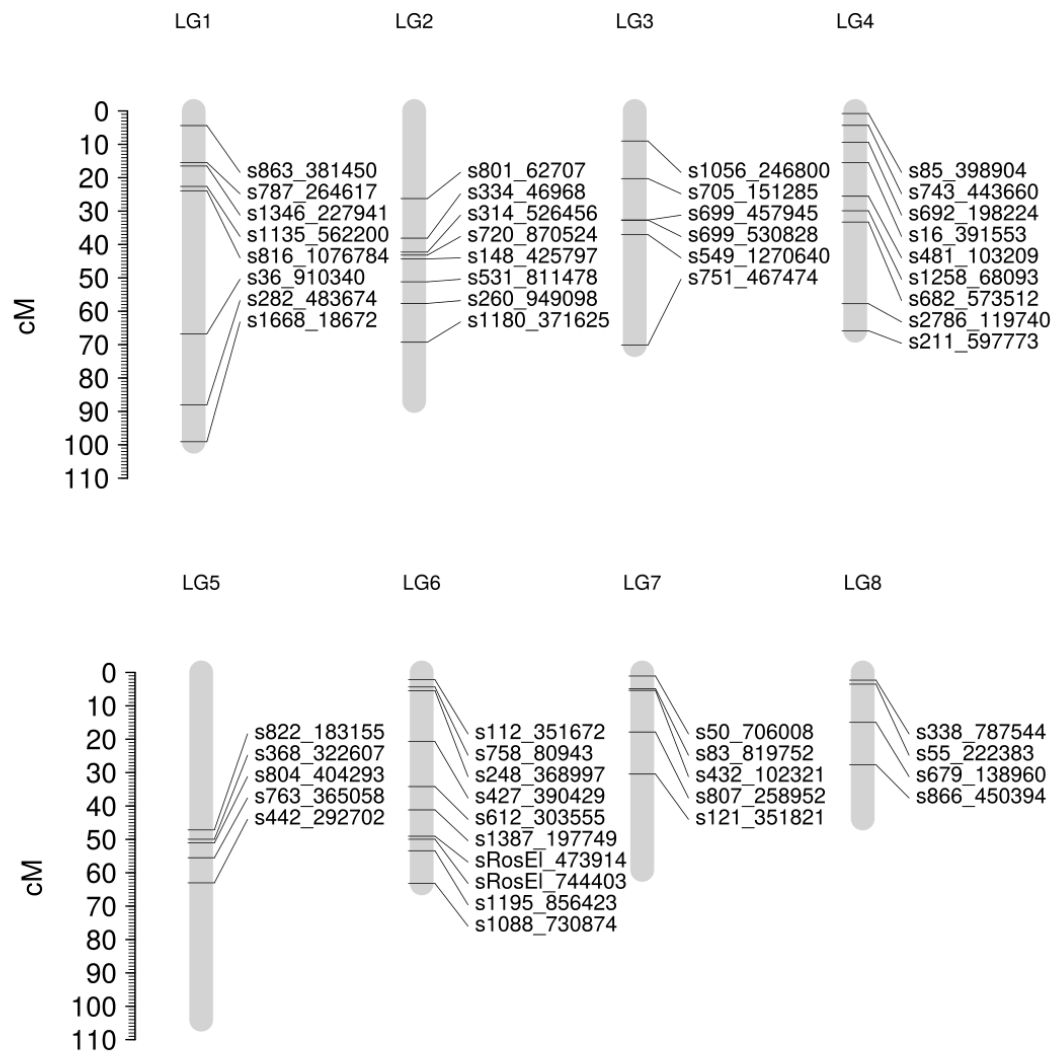


Figure 3.20: Position of KASP SNPs used in our analysis on draft *Antirrhinum majus* linkage map. Positions of 56 loci (black lines) indicated, the remaining 4 SNP loci could not be placed on the linkage map.

2. **Linkage Disequilibrium:** We further filtered markers with strong linkage disequilibrium, because our method assumes that different markers contain independent information. We iteratively pruned markers that were correlated more than $R^2 = 0.03$ with any other marker.

3. **Minor Allele Frequency:** We removed all markers with a minor allele frequency below $\bar{p} = 0.15$, as rare markers can have a dominating influence on pairwise measures of relatedness, but here we aim to base inference on the independent information of many markers. Most markers in our dataset have intermediate allele frequency near $p = 0.5$, and the overall allele frequency distribution is relatively narrow (Fig. 3.21) .

4. **Deviations from Hardy-Weinberg Equilibrium:** We tested for a significant deficit or surplus of heterozygous from random mating expectations. These deviations can have multiple reasons, for instance failed genotyping, strong geographic structure or non-random mating. For filtering, we first calculated local allele frequency estimates by weighting all other individuals with a two-dimensional symmetric Gaussian. After testing the fit of several standard deviations σ , we found that $\sigma = 500$ meters gave the best predictions for local allele frequencies. We first calculated this expected mean allele-frequency for every marker and every individual with this Gaussian. Based on these local frequencies, we then obtained the expected number of heterozygous and homozygous sites for each marker. Using a χ^2 -test, we calculated p -values for deviations from the expected numbers. We then filtered markers that had a \bar{p} -value below a cutoff of 10^{-5} .

After filtering, we were left with a dataset of $n = 12389$ individuals and 60 SNP markers. To ensure that there is no bias towards low or high frequencies, we flipped the 0 and 1 state for every marker with probability 0.5.

Data Availability

The detailed oligo-sequences for SNP genotyping, filtered genotype and geographic data are available at <https://github.com/hringbauer/BarrierInferPublic/tree/master/DataHZ>. The Python scripts used for data filtering are freely accessible at <https://github.com/hringbauer/BarrierInferPublic/tree/master/SNPCleaningScripts>.

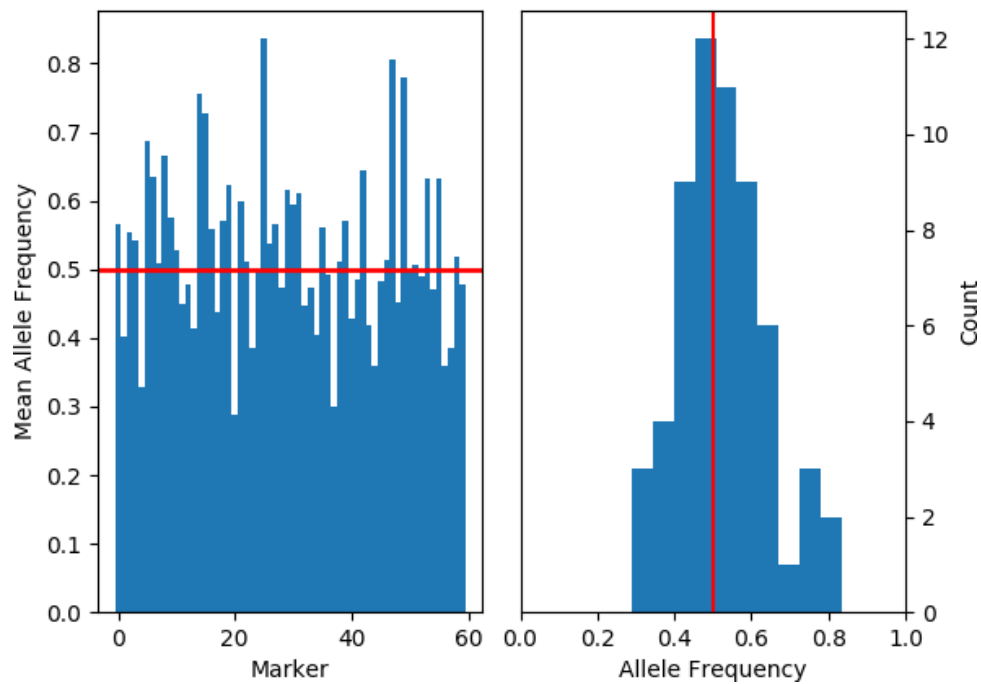


Figure 3.21: Allele frequency distribution. This figure depicts the distribution of the mean allele frequencies of the 60 markers (standard deviation 0.117). Left: Mean allele frequency ordered by marker. Right: Histogram of mean allele frequencies.

Heterogeneity of Isolation by Distance

Our method assumes a uniform isolation by distance pattern in two dimensions. To confirm that this is not grossly violated for the *Antirrhinum* data, we calculated isolation by distance patterns and investigated them for heterogeneity, both with respect to absolute position and angle (Fig. 3.22). Our analysis indicates that there are some spatial fluctuations of isolation by distance. However, they are mostly within the uncertainty estimates obtained by bootstrapping over genetic markers; so there is no indication of gross violations of the model assumptions.

Power Simulation

To test whether our method has sufficient power to detect a strong barrier to gene flow, we simulated a dataset similar to the *Antirrhinum* dataset. We used the same simulation engine described in the main text. We used 60 markers with a standard deviation $\sigma(\bar{p}) = 0.117355$, as in the filtered hybrid zone data set. We simulated a population of a 60×40 demes one dispersal unit apart, with 16 diploid individuals per deme (thus $N_{bh}=201.06$). We sampled one individual per deme, and simulated a strong barrier to gene flow ($\gamma = 0.02$). This synthetic dataset has a similar isolation by distance pattern as the hybrid zone data set (Fig. 3.23). Running our inference scheme on this dataset of

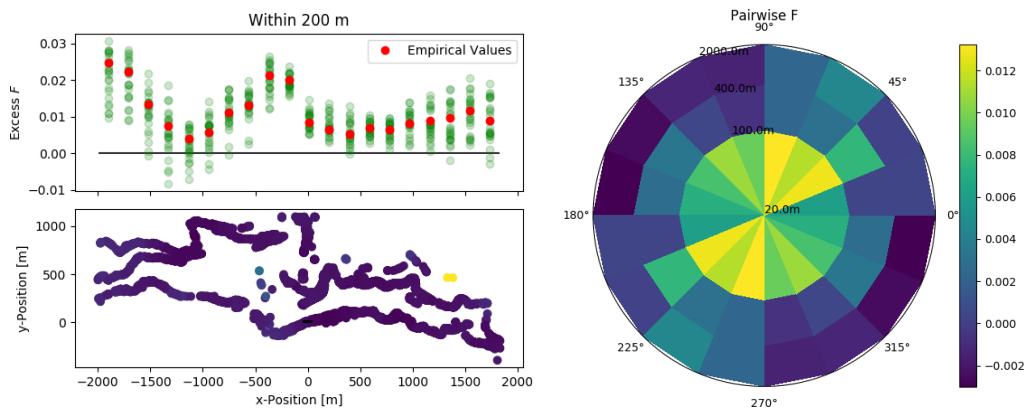


Figure 3.22: Heterogeneity of isolation by distance estimates. As a test statistic for the chance of recent co-ancestry F we used $\frac{(p_1 - \bar{p})(p_2 - \bar{p})}{4}$. This statistic should be a good estimator for F for intermediate mean allele frequencies ($\bar{p} \approx 0.5$). In the figures we depict the excess deviation compared to the average over all pairs. Left: Spatial heterogeneity and 25 bootstraps over loci: We depict mean excess F for all pairs with distance less than 200 meters. Right: Excess F when pairs are binned into 16 angular bins and three distance bins (20 – 100, 100 – 400, 400 – 2000 meter).

2400 samples indicates that there is sufficient power to infer the presence of a strong barrier (Fig. 3.24). At the true position of the barrier, the fit as well as 20 bootstraps over markers estimate a strong barrier to gene flow. For most other putative locations, no strong barrier is estimated. There is variation of bootstrap estimates which indicates that power is limited, but in total only a small number of bootstrap fits estimates a strong barrier. We stress that these power simulations are done for an idealized scenario in which our model assumptions hold.

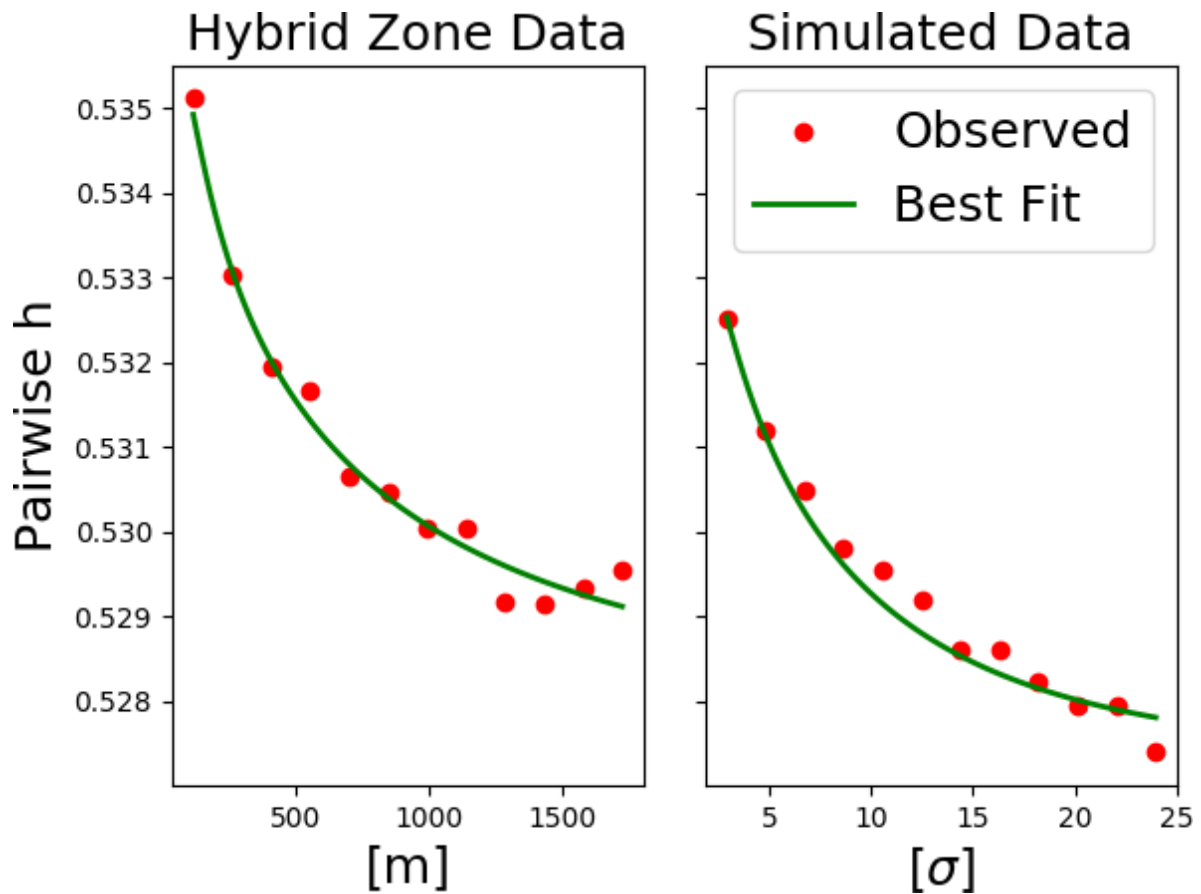


Figure 3.23: IBD of data simulated with hybrid zone parameters. The plot depicts pairwise homozygosity for pairwise distance bins. The small difference in absolute values is due to the randomness with which mean allele frequencies for the synthetic data set were drawn. Left: Hybrid zone data (fit: $Nbh = 192.20$, $m = 0.00839$, $s = 0.528088$). Right: Synthetic data set (fit: $Nbh = 150.6$, $m = 0.0056$, $s = 0.52735$), pairwise distance is measured in standard deviations σ of the dispersal kernel.

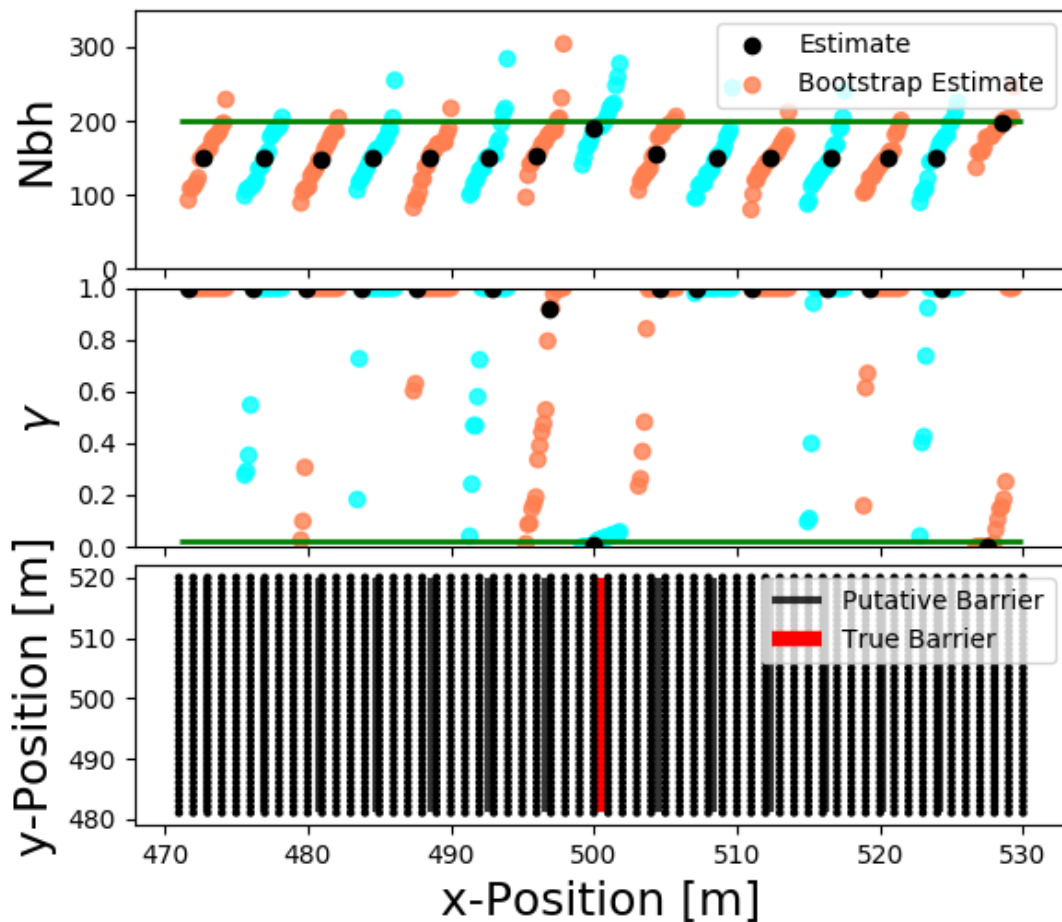


Figure 3.24: Power of inference scheme on simulated data set. Black dots indicate best fit estimates, colored dots are estimates after bootstrapping over genetic markers. Marker number, mean allele frequency distribution and demographic parameters were chosen to approximate the parameters from the *Antirrhinum* hybrid zone. We simulated a strong barrier ($\gamma = 0.02$), and run the inference scheme for multiple putative barrier locations (indicated by black lines).



4. Future Directions

THE methods developed in this thesis are implemented and are ready to be applied to more datasets. However, there are some necessary steps needed to provide other researchers a comfortable option to use these inference schemes. Perhaps the most pressing task is to develop software that makes the inference methods easily accessible. Further down, I outline a road-map to a full software package that helps to address these issues.

There are also many possible extensions. The underlying models can be expanded to fit other scenarios, in particular for IBD block sharing. Since the inference schemes introduced here are based on analytical approximations, such extensions to more complex models will lead to novel computational challenges. However, the methods to fit the model to data are very general and can be easily adapted to fit a wider range of models. In the following, I will outline some of these possible extensions.

4.1 Generalized Parametric Inference for IBD Blocks

The inference method of Chapter 2 fits isolation by distance patterns of IBD blocks to a scenario of homogeneous diffusion of ancestry. While this scenario is designed to be a general approximation to many two-dimensional populations, it is easy to imagine other scenarios and models for which IBD block data can be utilized for estimation of population structure. For instance, one might be interested to infer the effect of a barrier to gene flow (as in Chapter 3), or in scenarios where demographic parameters vary throughout a habitat and perhaps even time.

The methods used to fit observed IBD blocks to a model do not depend on the specifics of the demographic model. Recapitulating briefly, the inference scheme predicts expected pairwise IBD block sharing for given demographic parameters, and finds the parameter values that maximize the fit to observed IBD block sharing. In principle, one can use the same technique to fit the parameters of any other model that

predicts expected IBD block sharing.

The only big advantage of the diffusion model is its analytical tractability. Pairwise sharing of IBD blocks of a given length can be easily calculated with an analytical formula (Eq. 3), and this leads to a huge speed up of computational runtime. To calculate expected IBD block sharing for more complex models, numerical techniques have to be applied. These can be a huge computational challenge.

To address this issue, I outline a general numerical approach which can be used to calculate this expected pairwise sharing. It is based on expanding some of the key techniques of Chapter 2. This proposed numerical approach is joint work with **Raphael Forien**, who helped to design and implement a first pilot project.

4.1.1 Parametric Model for Spread of Ancestry

The goal is to calculate pairwise IBD block sharing of a given segment length. In Chapter 2, we used a diffusion approximation to model the movement of ancestral lineages, and we combined this with Poisson recombination to calculate expected sharing of IBD blocks (Eq. 2.3). Importantly, the specific demographic model only entered via the pairwise coalescence time distribution $\psi(t)$ in these calculations. Any numerical scheme that can predict pairwise coalescence probabilities $\psi(t)$ can therefore be straightforwardly combined with the method of Chapter 2.

Here, we introduce a numerical method to calculate this pairwise coalescence time distribution. We use a model discretized both in space and time, and follow the probability density of ancestry for an initial position on a grid backwards in time. In every time step t , ancestry is updated according to a migration matrix $M(t)$ that describes the migration probabilities between all grid points. Given an initial set of samples $i = 1 \dots n$ at initial positions $x_i = 1 \dots n$, this approach yields the spatial probability density distribution $G_{x_i}^t$ of ancestry at time t back:

$$G_{x_i}^t = M(t)M(t-1) \cdots M(2)M(1)x_i. \quad (4.1)$$

The chance that two lineages starting at position x and y coalesce at time t is given by the sum over the probabilities that lineages coalesce in any local deme. Denoting the number of individuals in deme k by $N_e(k)$, this sum is:

$$\psi_{xy}(t) = \sum_k \frac{1}{N_e(k)} G_x^t(k) G_y^t(k). \quad (4.2)$$

This recursion neglects previous coalescence events, similar to the approximation in 2. There, this simplification was applied to yield analytical formulas. When doing

numerical calculations of the coalescence time distribution, one could in principal also account for previous coalescence events. For instance, the spatial probability density distributions, $G_{x_i}^t$ could be updated to delete previously coalesced lineages. However, these updates would have to be applied for every pair of samples, and this scaling could be numerically problematic. Since double coalescence events are usually not problematic for all but very low neighborhood sizes (Chapter 2, (Wilkins, 2004)) neglecting them for sake of speed of calculations seems to be a valid approximation for most practical use cases. However, one should always keep the approximative nature of Eq. 4.2 in mind, in particular when dealing with extreme cases.

This discrete numerical model is very general. One can vary the migration matrices $M(t)$ as well as the local coalescence probabilities D_e to model spatial and also temporal heterogeneities. For a given set of demographic parameters, the expected IBD block sharing can be calculated, which can in turn can be combined with our method to fit these parameters to observed block sharing. In principal, this approach allows one to use arbitrary complex parametric models. However, caution with the number of parameters should be exercised in order to avoid over-fitting. Moreover, the parameters of the model should be sufficiently independent in order to avoid degeneracies and to ensure that they can be estimated independently.

This numerical model has some useful computational properties. The calculations of Eq. 4.1 and Eq. 4.2 can be efficiently implemented as matrix calculations, which helps to speed up calculations in many programming languages. For many scenarios, these matrices are very sparse, which further speeds up calculations. For multiple samples, one also only needs to trace the ancestry of a sample at x only once, and one can reuse the stored results for intermediate timesteps. The overall complexity of this algorithm therefore scales linearly with the number of demes and number of time steps, and quadratically with the number of initial samples. For many realistic discretizations, this allows numerically tractable calculations.

4.1.2 Example: Heterogeneous Gene Flow and Population Density

To demonstrate the generality and computational tractability of this discrete numerical approach, we used it to implement a scenario where migration and population densities vary across a linear interface in a two-dimensional population (Fig. 4.1). In this model, individuals occupy a grid of demes, and migrate between them. Left of the linear interface, the migration kernel has axial variance σ_L^2 and the number of individuals per deme is given by D_e^l , whereas the parameters are different on the right hand side of the barrier (σ_R^2 and D_e^r).

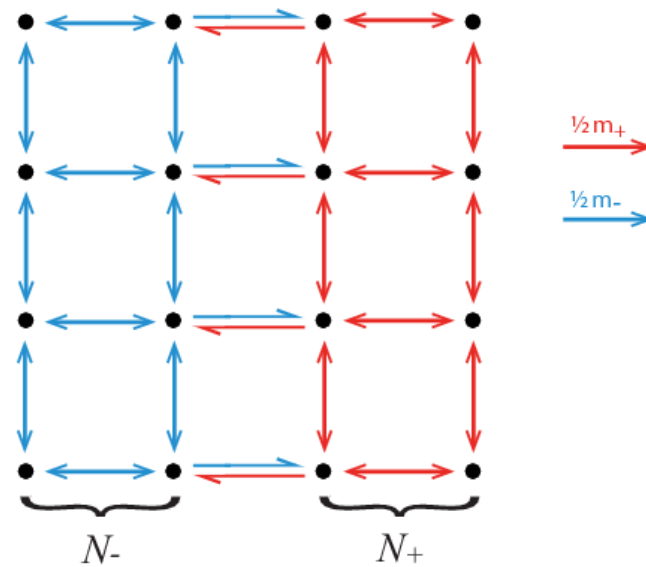


Figure 4.1: Spatial heterogeneous model: There are two different regimes in a two-dimensional population. Two areas with different population densities (D_e^l/D_e^r) and diffusion constants (σ_l/σ_r) are separated by a linear divide. We simulate this model by using a stepping stone model demes, in which neighboring demes exchange migrants at a rate m_- resp. m_+ , and demes contain N_- resp. N_+ individuals.

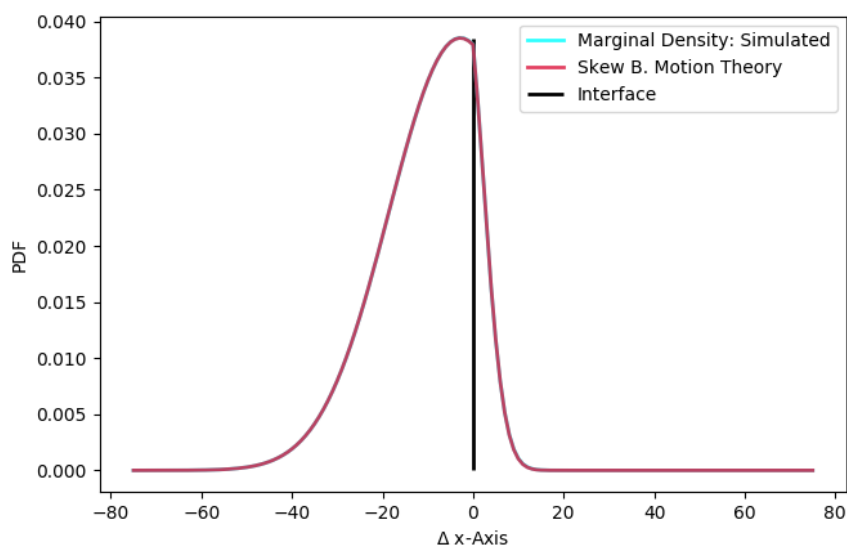


Figure 4.2: Spread of ancestry in numerical simulation compared to analytical formulae. Spread of ancestry was simulated for a two-dimensional discrete stepping stone model. The figure shows the marginal density summed up along the axis parallel to the linear divide. The analytical formula is a solution for skew Brownian motion (Harrison and Shepp, 1981).

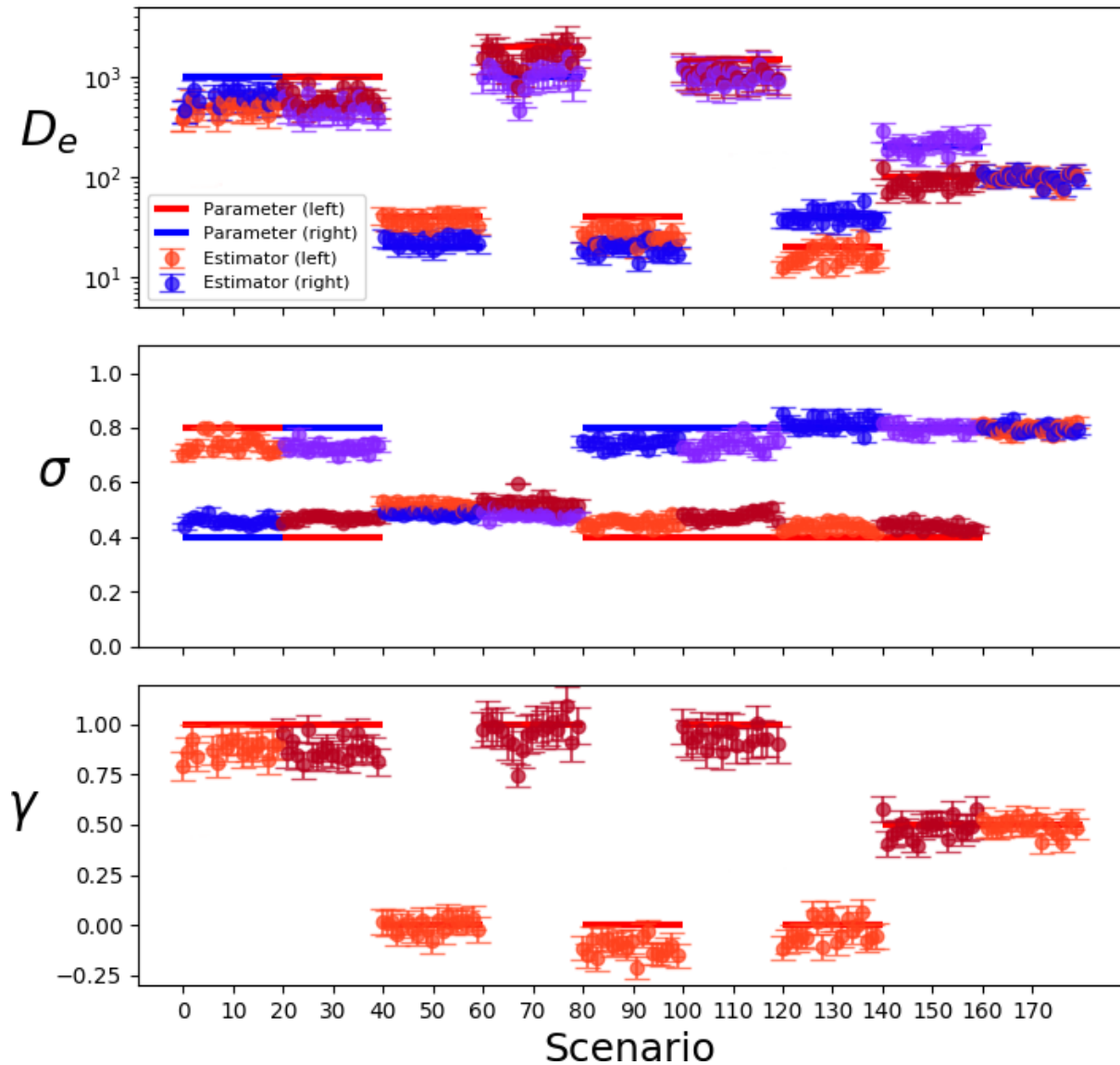


Figure 4.3: We ran 180 simulations of block sharing between of 10 chromosomes of length 5000 cM from 24 population spaced on a 6×4 grid symmetric around the divide. The spacing between neighboring sampled populations was 4 demes along each axis. The underlying grid consisted of 200×200 demes with a number of diploid individuals calculated from the population density function $D_e \cdot t^{-\gamma}$. We simulated 20 independent replicates of 9 different scenarios - the lines mark the parameters used in the simulations. The dots are the best estimates for each simulation, the error bars visualize 95 % confidence intervals based on the curvature of the numerically inferred likelihood surface.

Results on simulated data

We simulated data using the same simulation scheme developed in Chapter 2, by tracing ancestral parts of chromosomes back in time. It was straightforward to update the existing simulation engine, since it is already based on a model of grid of demes that exchange migrants. By updating the dispersal kernel resp. the migration matrix together with a model for past population density model, in principal arbitrary complex scenarios can be simulated with this powerful engine.

We ran a number of IBD block sharing simulations. We used different demographic parameters for dispersal rates, population densities and a global population growth parameter, and then inferred these parameters based on observed pairwise IBD block sharing. The calculations are now computationally more challenging than in Chapter 2, as ancestry has to be traced back numerically. Therefore, we heavily utilized the cluster at IST Austria to run simulations and inference on replicate scenarios in parallel on many nodes.

Our results demonstrate that our inference scheme can robustly recover the demographic parameters used in the simulated data sets. It can infer population density and growth as well as dispersal. There are slight estimation biases for some scenarios, which require further investigation. But overall, the method can robustly distinguish between different population densities and dispersal distances on different sides of the divide. These first tests of this scheme on simulated data are promising. However further thorough testing is needed to fully confirm that the inference scheme and the numerical implementation work as intended.

4.1.3 Further Directions

The scenario of a heterogenous habitat represents merely a first showcase application. The numerical scheme we presented here can be used to calculate and fit block sharing for many other interesting scenarios. For instance, it can be used to infer linear barriers to gene flow. But also more complex shapes of divides can be fit. This will be of much use in cases where the geographic shape of the barrier is known a priori, for instance in the case of mountain ridges or rivers. In principle any habitat can be discretized, and as long as the migration matrix and the past population densities can be supplied by a parametric model, these parameters can be fit by using our composite likelihood approach. Inference schemes like this will hopefully open a completely new avenue for demographic inference.

One salient goal is to apply these more complex IBD block sharing models to the

human data that was already analyzed in Chapter 2. There, we found systematic large scale deviations from the best fit of simple diffusion of ancestry (Fig. 2.8), with more block sharing toward the southeastern part of Europe. It is a tempting task to test whether one can further disentangle the causes of these spatially heterogeneity, and check whether it can be explained better by a model of differential migration or differential population density. The spatially heterogeneous model will help to yield interesting novel insights into this question.

4.2 Software Package for IBD Block Analysis

The above described inference methods were designed with the long term goal to be applied to real data. However, as of now the methods have to be deployed by directly interacting with code, and this burden is likely too prohibitive for many other researchers who are not experts in programming and computational analysis. Removing this hurdle will likely lead to a much more widespread application of the methods developed here.

Therefore, I plan to implement the above methods to simulate and fit IBD block sharing in a versatile Python package. This program will likely include a Graphical User Interface (GUI), that helps to reduce the interaction of the user with the underlying code. I plan to use the Python package management system *pip* for distributing this software.

The foundation of this package will be the following two main features:

1. **Run custom block sharing simulations** With this feature, the user can simulate IBD block data sets with the simulation engine outlined in Chapter 2. He can place samples on a grid of demes, and run numerical simulations that yield the IBD sharing between these samples. Moreover, the user will be enabled to supply his own demographic scenario, as outlined in subsection 4.1.2. He can choose custom migration matrices and population densities, which he can supply to the package in form of python functions. In combination with running inference schemes on this simulated data sets (see below), these custom scenarios will help researchers to determine the power of the inference scheme given a specified sampling scheme and some likely demographic scenarios. Such tests will help researchers to determine the ideal sampling scheme for each specific inference task.
2. **Fit custom demographic scenarios** A main part of this program will be the implementation of the inference method of Chapter 2. In addition, I plan to give the

user the opportunity to supply custom demographic scenarios. The user will be able to fit arbitrary scenarios with the method outlined in subsection 4.1.2. He can provide custom parametrized migration matrices and population densities, and the program will find the parameters that best fit observed IBD block sharing patterns. To obtain uncertainty estimates, the user will be given the confidence intervals obtained by the likelihood surface, and also the possibility to get bootstrap estimates.

To be most useful for the user, such a package should ideally possess the following key properties:

- **Graphical presentation of the results:** An appealing visualization of the results will be helpful. Presenting the results in an informative way will aid the user to quickly explore the fit of the model to the data, and help him to obtain a better overview of the results of the analysis.
- **Sample files:** Accompanying sample files will help the user to get an overview about the execution speed, the presentation of the results and the range of computations the method can perform.
- **A detailed user manual:** A user manual will inform the user about the possibilities of the program, and can be consulted in case help is needed.

However, while these features will provide a better experience for the user, they will not empower him to use the software package in a black box manner. He will always be required to make himself familiar with the underlying theory paper (Ringbauer et al., 2017a), and the overall limitations of the method to avoid interpretation errors and inconsistent program settings.

4.3 Outlook

Demographic inference based on IBD blocks is likely the future of demographic inference. In the era of population genomics, datasets that allow one to reliably call long IBD blocks will become more and more widespread. As one can study the direct genetic traces of distant relatedness, this signal is ideally suited for the inference of recent demographic structure. IBD blocks contain all genetic signals of relatedness in the recent history of a population, and therefore the full genetic information for the inference of recent population structure.

As of now, calling IBD blocks is only feasible for humans and a few model organisms. However, in light of the falling costs of next-generation sequencing technologies, such studies will soon become common in many other organisms as well. This foreseeable advance will make methods based on IBD blocks, such as the one introduced here, a powerful tool to infer population structure. As outlined above, possible applications go far beyond the simple analytical model that has been treated in Chapter 2. One can use this signal to fit more complex structure, such as heterogeneities across the habitat or barriers to gene flow. The methods outlined here are only a first step in a journey to a better understanding of the demographic structure of natural populations. But I hope that they will provide a solid foundation for future work, and that this work will become a valuable building block in the wall of human knowledge.

Bibliography

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Number 55. Courier Corporation.

Aguillon, S. M., Fitzpatrick, J. W., Bowman, R., Schoech, S. J., Clark, A. G., Coop, G., and Chen, N. (2017). Deconstructing isolation-by-distance: The genomic consequences of limited dispersal. *PLoS Genetics*, 13(8):e1006911.

Baharian, S., Barakatt, M., Gignoux, C. R., Shringarpure, S., Errington, J., Blot, W. J., Bustamante, C. D., Kenny, E. E., Williams, S. M., Aldrich, M. C., et al. (2016). The great migration and African-American genomic diversity. *PLoS Genetics*, 12(5):e1006059.

Barton, N. (1979). Gene flow past a cline. *Heredity*, 43(3):333–339.

Barton, N. (2008). The effect of a barrier to gene flow on patterns of geographic variation. *Genetics Research*, 90(01):139–149.

Barton, N. and Bengtsson, B. O. (1986). The barrier to genetic exchange between hybridising populations. *Heredity*, 57(3):357–376.

Barton, N., Etheridge, A., Kelleher, J., and Véber, A. (2013). Inference in two dimensions: Allele frequencies versus lengths of shared sequence blocks. *Theoretical Population Biology*, 87:105–119.

Barton, N., Halliday, R., and Hewitt, G. (1983). Rare electrophoretic variants in a hybrid zone. *Heredity*, 50(2):139–146.

Barton, N. H., Depaulis, F., and Etheridge, A. M. (2002). Neutral evolution in spatially continuous populations. *Theoretical Population Biology*, 61(1):31–48.

Barton, N. H. and Gale, K. S. (1993). Genetic analysis of hybrid zones. *Hybrid zones and the evolutionary process*, pages 13–45.

Barton, N. H. and Hewitt, G. M. (1985). Analysis of hybrid zones. *Annual Review of Ecology and Systematics*, 16(1):113–148.

- Barton, N. H., Kelleher, J., and Etheridge, A. M. (2010). A new model for extinction and recolonization in two dimensions: quantifying phylogeography. *Evolution*, 64(9):2701–2715.
- Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.
- Beerli, P. and Felsenstein, J. (2001). Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences*, 98(8):4563–4568.
- Blouin, M. S. (2003). Dna-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends in Ecology & Evolution*, 18(10):503–511.
- Bossart, J. and Prowell, D. P. (1998). Genetic estimates of population structure and gene flow: limitations, lessons and new directions. *Trends in Ecology & Evolution*, 13(5):202–206.
- Bradburd, G. S., Ralph, P. L., and Coop, G. M. (2013). Disentangling the effects of geographic and ecological isolation on genetic differentiation. *Evolution*, 67(11):3258–3273.
- Bradley, D., Xu, P., Mohorianu, I.-I., Whibley, A., Field, D., Tavares, H., Couchman, M., Copsey, L., Carpenter, R., Li, M., et al. (2017). Evolution of flower color pattern through selection on regulatory small RNAs. *Science*, 358(6365):925–928.
- Brinkmann, B., Klintschar, M., Neuhuber, F., Hühne, J., and Rolf, B. (1998). Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *The American Journal of Human Genetics*, 62(6):1408–1415.
- Browning, B. L. and Browning, S. R. (2011). A fast, powerful method for detecting identity by descent. *The American Journal of Human Genetics*, 88(2):173–182.
- Browning, S. R. and Browning, B. L. (2012). Identity by descent between distant relatives: detection and applications. *Annual Review of Genetics*, 46:617–633.
- Browning, S. R. and Browning, B. L. (2015). Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *The American Journal of Human Genetics*, 97(3):404–418.
- Cercueil, A., François, O., and Manel, S. (2007). The genetical bandwidth mapping: a spatial and graphical representation of population genetic structure based on the wombling method. *Theoretical Population Biology*, 71(3):332–341.

- Chiang, C. W., Ralph, P., and Novembre, J. (2016). Conflation of short identity-by-descent segments bias their inferred length distribution. *G3: Genes | Genomes | Genetics*, 6(5):1287–1296.
- Coffman, A. J., Hsieh, P. H., Gravel, S., and Gutenkunst, R. N. (2016). Computationally efficient composite likelihood statistics for demographic inference. *Molecular Biology and Evolution*, 33(2):591–593.
- Davies, N. (2014). *Europe: A history*. Random House.
- Duforet-Frebourg, N. and Blum, M. G. (2014). Nonstationary patterns of isolation-by-distance: Inferring measures of local genetic differentiation with bayesian kriging. *Evolution*, 68(4):1110–1123.
- Dupanloup, I., Schneider, S., and Excoffier, L. (2002). A simulated annealing approach to define the genetic structure of populations. *Molecular Ecology*, 11(12):2571–2581.
- Ellis, T. J. (2016). *The role of pollinator-mediated selection in the maintenance of a flower color polymorphism in an *Antirrhinum majus* hybrid zone*. PhD thesis, IST Austria.
- Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587.
- Fearnhead, P. and Donnelly, P. (2002). Approximate likelihood methods for estimating local recombination rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):657–680.
- Felsenstein, J. (1975). A pain in the torus: some difficulties with models of isolation by distance. *American Naturalist*, pages 359–368.
- Fenner, J. N. (2005). Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *American journal of physical anthropology*, 128(2):415–423.
- Fisher, R. A. (1937). The wave of advance of advantageous genes. *Annals of Eugenics*, 7(4):355–369.
- Fisher, R. A. et al. (1947). 219: The spread of a gene in natural conditions in a colony of the moth *panaxia dominula* l.

- Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., et al. (2002). The structure of haplotype blocks in the human genome. *Science*, 296(5576):2225–2229.
- Gao, F. and Keinan, A. (2016). Inference of super-exponential human population growth via efficient computation of the site frequency spectrum for generalized models. *Genetics*, 202(1):235–245.
- Goldberg, A., Verdu, P., and Rosenberg, N. A. (2014). Autosomal admixture levels are informative about sex bias in admixed populations. *Genetics*, 198(3):1209–1229.
- Grebenkov, D. S., Van Nguyen, D., and Li, J.-R. (2014). Exploring diffusion across permeable barriers at high gradients. i. narrow pulse approximation. *Journal of Magnetic Resonance*, 248:153–163.
- Guillot, G., Estoup, A., Mortier, F., and Cosson, J. F. (2005). A spatial statistical model for landscape genetics. *Genetics*, 170(3):1261–1280.
- Guillot, G., Leblois, R., Coulon, A., and Frantz, A. C. (2009). Statistical methods in spatial genetics. *Molecular Ecology*, 18(23):4734–4756.
- Guillot, G. and Rousset, F. (2013). Dismantling the Mantel tests. *Methods in Ecology and Evolution*, 4(4):336–344.
- Guillot, G. and Santos, F. (2009). A computer program to simulate multilocus genotype data with spatially autocorrelated allele frequencies. *Molecular Ecology Resources*, 9(4):1112–1120.
- Gusev, A., Lowe, J. K., Stoffel, M., Daly, M. J., Altshuler, D., Breslow, J. L., Friedman, J. M., and Pe'er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. *Genome Research*, 19(2):318–326.
- Hardy, O. J. and Vekemans, X. (1999). Isolation by distance in a continuous population: reconciliation between spatial autocorrelation analysis and population genetics models. *Heredity*, 83(2):145–154.
- Harris, K. and Nielsen, R. (2013). Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genetics*, 9(6):e1003521.
- Harrison, J.-M. and Shepp, L.-A. (1981). On skew brownian motion. *The Annals of Probability*, pages 309–313.

- Hellenthal, G., Busby, G. B., Band, G., Wilson, J. F., Capelli, C., Falush, D., and Myers, S. (2014). A genetic atlas of human admixture history. *Science*, 343(6172):747–751.
- Hill, W. G. (1981). Estimation of effective population size from data on linkage disequilibrium. *Genetics Research*, 38(3):209–216.
- Hubby, J. L. and Lewontin, R. C. (1966). A molecular approach to the study of genic heterozygosity in natural populations. i. the number of alleles at different loci in *Drosophila pseudoobscura*. *Genetics*, 54(2):577.
- Jay, F., Sjödin, P., Jakobsson, M., and Blum, M. G. (2013). Anisotropic isolation by distance: the main orientations of human genetic differentiation. *Molecular Biology and Evolution*, 30(3):513–525.
- Joseph, T., Hickerson, M., and Alvarado-Serrano, D. (2016). Demographic inference under a spatially continuous coalescent model. *Heredity*.
- Keinan, A. and Clark, A. G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*, 336(6082):740–743.
- Kelleher, J., Etheridge, A., and Barton, N. (2014). Coalescent simulation in continuous space: Algorithms for large neighbourhood size. *Theoretical Population Biology*, 95:13–23.
- Kimura, M. and Weiss, G. H. (1964). The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, 49(4):561.
- Kingman, J. F. C. (1982). The coalescent. *Stochastic processes and their applications*, 13(3):235–248.
- Knowles, L. L. (2009). Statistical phylogeography. *Annual Review of Ecology, Evolution, and Systematics*, 40:593–612.
- Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., et al. (2002). A high-resolution recombination map of the human genome. *Nature Genetics*, 31(3):241–247.
- Leblois, R., Estoup, A., and Rousset, F. (2003). Influence of mutational and sampling factors on the estimation of demographic parameters in a "continuous" population under isolation by distance. *Molecular Biology and Evolution*, 20(4):491–502.

- Leblois, R., Rousset, F., and Estoup, A. (2004). Influence of spatial and temporal heterogeneities on the estimation of demographic parameters in a continuous population using individual microsatellite data. *Genetics*, 166(2):1081–1092.
- Lemey, P., Rambaut, A., Welch, J. J., and Suchard, M. A. (2010). Phylogeography takes a relaxed random walk in continuous space and time. *Molecular Biology and Evolution*, 27(8):1877–1885.
- Li, H. and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496.
- Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics*, 80(1):221–39.
- Lynch, M., Xu, S., Maruki, T., Jiang, X., Pfaffelhuber, P., and Haubold, B. (2014). Genome-wide linkage-disequilibrium profiles from single individuals. *Genetics*, 198(1):269–281.
- Malécot, G. (1948). Les mathématiques de l'hérédité. *Masson et Cie*.
- Manel, S., Schwartz, M. K., Luikart, G., and Taberlet, P. (2003). Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology & Evolution*, 18(4):189–197.
- Manni, F., Guerard, E., and Heyer, E. (2004). Geographic patterns of (genetic, morphologic, linguistic) variation: how barriers can be detected by using monmonier's algorithm. *Human Biology*, 76(2):173–190.
- McEvedy, C., Jones, R., et al. (1978). *Atlas of world population history*. Penguin Books Ltd, Harmondsworth, Middlesex, England.
- McRae, B. H. (2006). Isolation by resistance. *Evolution*, 60(8):1551–1561.
- McRae, B. H., Dickson, B. G., Keitt, T. H., and Shah, V. B. (2008). Using circuit theory to model connectivity in ecology, evolution, and conservation. *Ecology*, 89(10):2712–2724.
- Meirmans, P. G. (2012). The trouble with isolation by distance. *Molecular Ecology*, 21(12):2839–2846.
- Nagylaki, T. (1978). A diffusion model for geographically structured populations. *Journal of Mathematical Biology*, 6(4):375–382.

- Nagylaki, T. (1988). The influence of spatial inhomogeneities on neutral models of geographical variation: I. formulation. *Theoretical Population Biology*, 33(3):291–310.
- Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences*, 70(12):3321–3323.
- Nelson, M. R., Bryc, K., King, K. S., Indap, A., Boyko, A. R., Novembre, J., Briley, L. P., Maruyama, Y., Waterworth, D. M., Waeber, G., et al. (2008). The population reference sample, popres: a resource for population, disease, and pharmacological genetics research. *The American Journal of Human Genetics*, 83(3):347–358.
- Nickisch, H. and Rasmussen, C. E. (2008). Approximations for binary Gaussian process classification. *Journal of Machine Learning Research*, 9(Oct):2035–2078.
- Nielsen, R. and Beaumont, M. A. (2009). Statistical inferences in phylogeography. *Molecular Ecology*, 18(6):1034–1047.
- Novembre, J. and Peter, B. M. (2016). Recent advances in the study of fine-scale population structure in humans. *Current Opinion in Genetics & Development*, 41:98–105.
- Novembre, J. and Slatkin, M. (2009). Likelihood-based inference in isolation-by-distance models using the spatial distribution of low-frequency alleles. *Evolution*, 63(11):2914–2925.
- Palamara, P. F., Lencz, T., Darvasi, A., and Pe'er, I. (2012). Length distributions of identity by descent reveal fine-scale demographic history. *The American Journal of Human Genetics*, 91(5):809–822.
- Palamara, P. F. and Pe'er, I. (2013). Inference of historical migration rates via haplotype sharing. *Bioinformatics*, 29(13):i180–i188.
- Petkova, D., Novembre, J., and Stephens, M. (2015). Visualizing spatial population structure with estimated effective migration surfaces. *Nature Genetics*.
- Pooley, C. and Turnbull, J. (2005). *Migration and mobility in Britain since the eighteenth century*. Routledge.
- Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.
- Ralph, P. and Coop, G. (2013). The geography of recent genetic ancestry across europe. *PLoS Biology*, 11(5):e1001555.

- Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian processes for machine learning*, volume 1. MIT press Cambridge.
- Ringbauer, H., Coop, G., and Barton, N. H. (2017a). Inferring recent demography from isolation by distance of long shared sequence blocks. *Genetics*, 205(3):1335–1351.
- Ringbauer, H., Kolesnikov, A., Field, D., and Barton, N. H. (2017b). Estimating barriers to gene flow from distorted isolation by distance patterns. *bioRxiv*, page 205484.
- Robledo-Arnuncio, J. and Rousset, F. (2010). Isolation by distance in a continuous population under stochastic demographic fluctuations. *Journal of Evolutionary Biology*, 23(1):53–71.
- Rousset, F. (1997). Genetic differentiation and estimation of gene flow from f-statistics under isolation by distance. *Genetics*, 145(4):1219–1228.
- Rousset, F. (2000). Genetic differentiation between individuals. *Journal of Evolutionary Biology*, 13(1):58–62.
- Rousset, F. (2002). Inbreeding and relatedness coefficients: What do they measure? *Heredity*, 88(5):371–380.
- Safner, T., Miller, M. P., McRae, B. H., Fortin, M.-J., and Manel, S. (2011). Comparison of Bayesian clustering and edge detection methods for inferring boundaries in landscape genetics. *International Journal of Molecular Sciences*, 12(2):865–889.
- Scally, A. (2016). The mutation rate in human evolution and demographic inference. *Current Opinion in Genetics & Development*, 41:36–43.
- Sedghifar, A., Brandvain, Y., Ralph, P., and Coop, G. (2015). The spatial mixing of genomes in secondary contact zones. *Genetics*, 201(1):243–261.
- Slatkin, M. (1985). Rare alleles as indicators of gene flow. *Evolution*, pages 53–65.
- Slatkin, M. (1991). Inbreeding coefficients and coalescence times. *Genetical Research*, 58(02):167–175.
- Slatkin, M. (1993). Isolation by distance in equilibrium and non-equilibrium populations. *Evolution*, 47(1):264–279.
- Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, 139(1):457–462.

- Slatkin, M. and Barton, N. H. (1989). A comparison of three indirect methods for estimating average levels of gene flow. *Evolution*, pages 1349–1368.
- Smouse, P. E., Long, J. C., and Sokal, R. R. (1986). Multiple regression and correlation extensions of the mantel test of matrix correspondence. *Systematic Zoology*, 35(4):627–632.
- Storfer, A., Murphy, M., Evans, J., Goldberg, C., Robinson, S., Spear, S., Dezzani, R., Delmelle, E., Vierling, L., and Waits, L. (2007). Putting the 'landscape' in landscape genetics. *Heredity*, 98(3):128.
- Storfer, A., Murphy, M. A., Spear, S. F., Holderegger, R., and Waits, L. P. (2010). Landscape genetics: where are we now? *Molecular Ecology*, 19(17):3496–3514.
- Vekemans, X. and Hardy, O. (2004). New insights from fine-scale spatial genetic structure analyses in plant populations. *Molecular Ecology*, 13(4):921–935.
- Von Foerster, H., Mora, P. M., and Amiot, L. W. (1960). Doomsday: Friday, 13 november, ad 2026. *Science*, 132(3436):1291–1295.
- Waples, R. S. and Do, C. (2008). Ldne: a program for estimating effective population size from data on linkage disequilibrium. *Molecular Ecology Resources*, 8(4):753–756.
- Watts, P. C., Rousset, F., Saccheri, I. J., Leblois, R., Kemp, S. J., and Thompson, D. J. (2007). Compatible genetic and ecological estimates of dispersal rates in insect (coenagrion mercuriale: Odonata: Zygoptera) populations: analysis of 'neighbourhood size' using a more precise estimator. *Molecular Ecology*, 16(4):737–751.
- Weir, B. S. and Cockerham, C. C. (1984). Estimating f-statistics for the analysis of population structure. *Evolution*, pages 1358–1370.
- Wetterstrand, K. A. (2013). Dna sequencing costs: data from the NHGRI genome sequencing program.
- Whibley, A. C., Langlade, N. B., Andalo, C., Hanna, A. I., Bangham, A., Thébaud, C., and Coen, E. (2006). Evolutionary paths underlying flower color variation in antirrhinum. *Science*, 313(5789):963–966.
- Whitlock, M. C. and McCauley, D. E. (1999). Indirect measures of gene flow and migration: $F_{ST} \neq 1/(4Nm + 1)$. *Heredity*, 82(2):117–125.
- Wijsman, E. M. and Cavalli-Sforza, L. L. (1984). Migration and genetic population structure with special reference to humans. *Annual Review of Ecology and Systematics*, 15:279–301.

Wilkins, J. F. (2004). A separation-of-timescales approach to the coalescent in a continuous population. *Genetics*, 168(4):2227–2244.

Wilkins, J. F. and Wakeley, J. (2002). The coalescent in a continuous, finite, linear population. *Genetics*, 161(2):873–888.

Womble, W. H. (1951). Differential systematics. *Science*, 114(2961):315–322.

Wright, S. (1938). The distribution of gene frequencies under irreversible mutation. *Proceedings of the National Academy of Sciences*, 24(7):253–259.

Wright, S. (1943). Isolation by distance. *Genetics*, 28(2):114.

Wright, S. (1949). The genetical structure of populations. *Annals of Human Genetics*, 15(1):323–354.

Wright, S., Dobzhansky, T., and Hovanitz, W. (1942). Genetics of natural populations. vii. the allelism of lethals in the third chromosome of *Drosophila pseudoobscura*. *Genetics*, 27(4):363.