The Coalescent in Boundary-Limited Range Expansions

Dissertation
zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades
“Doctor rerum naturalium”
der Georg–August–Universität Göttingen

im Promotionsprogramm
International Max Planck Research School
“Physics of Biological and Complex Systems”
(IMPRS PBCS)
der Göttingen Graduate School for Neurosciences, Biophysics, and Molecular Biosciences
(GGNB)

vorgelegt von

Jens Nullmeier

aus Berlin

Göttingen, 2013

Thesis committee

Dr. Oskar Hallatschek, Biological Physics and Evolutionary Dynamics Group, Max Planck Institute for Dynamics and Self-Organization

Prof. Dr. Anja Sturm, Institut für Mathematische Stochastik, Georg–August–Universität Göttingen

Prof. Dr. Marc Timme, Network Dynamics Group, Max Planck Institute for Dynamics and Self-Organization

I confirm that I have written this thesis independently and with no other sources and aids than quoted.

Goettingen,

Contents

1. Introduction **9**
- 1.1. Structure of this thesis 10
- 1.2. Genetic diversity 11
- 1.3. Genetic drift 12
- 1.4. Spatial structure 12
- 1.5. Range expansions 13
- 1.6. The coalescent 14

I. The coalescent in boundary-limited range expansions **17**

2. Introduction **21**

3. Model **25**
- 3.1. The expanding stepping stone model (linear) 25
- 3.2. The continuous model (planar) 26

4. Results **29**
- 4.1. The expanding stepping stone model (linear) 29
  - 4.1.1. Analytical approximation 29
  - 4.1.2. Data collapse 33
  - 4.1.3. Sampling from different locations. 34
  - 4.1.4. Variable front velocities 34
- 4.2. The continuous model (planar) 36
- 4.3. Population density clines 38

5. Analytical approach **41**
- 5.1. Equations of motion 41
- 5.2. Coalescence time distribution far from the boundary for vanishing convection speeds 42
- 5.3. Effectively well-mixed front population 43
- 5.4. Characteristic scales 44

6. Supplementary information **45**
6.1. Range shifts 45
6.2. Model Details 46
6.2.1. Density gradients at the expansion front 46

7. Simulation design **47**
7.1. The stepping stone model simulations 47
7.1.1. The random walk 47
7.1.2. The boundary movement 47
7.1.3. Coalescence 48
7.1.4. The contraction front 48
7.2. The planar, continuous simulations 48
7.2.1. The Gaussian random walk 48
7.2.2. Coalescence 48

8. Discussion **49**

II. The spatial distribution of alleles in expanding populations **55**

9. Introduction **59**
9.1. The spatial distribution of alleles in stationary habitats 62
9.2. The spatial distribution of alleles in expanding habitats 63

10. Model and methods **65**
10.1. Colony growth 65
10.2. The coalescence process 68
10.2.1. Lineage movement strictly along the colonization paths: the fully quenched coalescent 68
10.2.2. Diffusive lineage movement behind the expansion front: the front–quenched coalescent 69
10.2.3. Diffusive lineage movement within the whole habitat: the unquenched coalescent 70
10.3. Observables 71
10.3.1. The critical patch size 72
10.3.2. Bubble shapes in expanding populations 73
10.3.3. Estimation of the expansion direction 78

11. Results **79**
11.1. Colonization 79

11.2. Coalescence 81
    11.2.1. The unquenched coalescent 81
    11.2.2. The fully quenched coalescent 104
    11.2.3. The front–quenched coalescent 107

12. Discussion **115**
12.1. The unquenched coalescent 115
12.2. The fully quenched coalescent 117
12.3. The front–quenched coalescent 117
12.4. Smaller samples 118
12.5. Conclusions 118
    12.5.1. Hitch–hiker patches 119
    12.5.2. A generalized model 120

Outlook **121**

A. Appendix to part I **123**
A.1. Supplementary Figures 123
A.2. Simulation code 124

B. Appendix to part II **127**
B.1. The phenotypical expansion velocity 127
B.2. The shape of the bubbles 127
    B.2.1. The mean in a periodic interval 127
    B.2.2. Possible values of SD_x and SD_y 128
    B.2.3. Supplementary Figures 128
B.3. Simulation code 145
    B.3.1. The spatial distribution of alleles in expanding populations . 145

Index **155**

Bibliography **157**

1. Introduction

Life on earth is the result of billions of years of evolution including the emergence of species, the colonization of even remote corners of the planet, small and massive extinction events, and subsequent recolonizations. From a contemporary perspective, the distributions of species may seem relatively stable, but over timescales relevant to evolution, the habitat ranges of natural populations change drastically and repeatedly. Such range changes can be caused, for instance, by the alternation between glacial ages and interglacials, by volcanic activity, by the change of the sea level, and by adaptation.

Human societies have faced and are facing unpleasant consequences of range changes such as the advance of pests into regions that were previously spared. Well known examples are the malaria vector *Anopheles* [52], the mite *Varroa destructor* (a major pest of the honey bee *e.g.* [17]), and the fungus *Phytophthora infestans* leading for instance to the Irish potato famine of 1845–57 [34]. The invasion of one species often goes hand in hand with the extinction of other species and thereby threatens biodiversity. Among the more prominent examples are the invasion of the cane toad *Bufo marinus* in Australia [63], the Nile perch *Lates niloticus* to Lake Victoria [79, 37], and the fire ant *Solenopsis invicta* to the United States, the Caribbean, Australia, and New Zealand [73]. The human expansion has certainly caused the extinction of many species.

Apart from such vivid examples, range changes can have a long–lasting impact on genetic diversity within the expanding population. The patterns of genetic diversity in spatially extended populations therefore entail the possibility of reading the footprint of demographic events in the past and to deduce predictions for future developments. Recent advances in biotechnology such as 454 pyrosequencing [69], Illumina sequencing–by–synthesis [9] and IonTorrent non–optical sequencing [87] made it possible to analyze larger samples at more loci with higher accuracy than ever before. To us, it is promising to develop models and methods that take advantage of this cornucopia of genetic data.

Today, the fundamental understanding of evolution is to a large part based on the modern evolutionary synthesis [54] from the middle of the 20th century. In the following we will briefly introduce the concepts of evolution as far as they are necessary for the understanding of this thesis. The questions discussed in this work are part of *population genetics*, a general introduction to the field can be found for instance in [36]. The goal of population genetics consists in quantitatively understanding the action of the fundamental forces of evolution: mutation, selection and genetic drift. Basic definitions and concepts used in this thesis are summarized in Table 1.1.

Name	Definition
allele	genotype at the considered locus
deme	local well-mixed subpopulation or subdivision of the habitat
fixation	process during which an allele reaches frequency 1 in the population
locus	stretch of DNA free of recombination and horizontal gene transfer in the time frame given by the model
mutant / wild–type	When a mutation occurs at a previously not polymorphic site, the mutant refers to the derived genotype. The wild–type refers to the non–mutant genotype.
natal dispersal distance	distance between an individual's place of birth and its parent's place of birth
polymorphic site	locus with more than one allele in the sample
spatially structured population	population in a spatially extended habitat with limited migration

Table 1.1.: Some terms in population genetics are not consistently used in the same way in different publication. The definitions listed here are used throughout this thesis.

1.1. Structure of this thesis

In this thesis, we will present predictions for the change of genetic diversity in populations under different scenarios of range expansions and range shifts. We develop methods for the detection and characterization of such scenarios and provide estimators for population parameters.

The remainder of this section is a short introduction to genetic diversity, models in population genetics, range expansions and the coalescent.

In part I, we focus on patterns of diversity along the expansion axis and establish the distinction between two types of range expansions. In part II, we complement the approach of part I by analyzing patterns in two spatial dimensions. We use the differences in the spatial distribution of alleles in expanding versus stationary populations to develop a detection method for range expansions and to illuminate the impact of the colonization history on genetic diversity.

In the final part, we discuss the results especially in the context of their applicability in experiments and give an outlook to follow–up research questions motivated by our results.

1.2. Genetic diversity

The models presented in this thesis describe different aspects of the distribution of genetic diversity in spatially extended habitats. The *genetic diversity* in a population is the amount of genetic variation between individuals from that population and can be quantified by means of different observables: the *heterozygosity*, for instance, describes the probability that two alleles picked randomly from a population are not identical at the considered locus (Historically, the heterozygosity describes the probability for a diploid individual two have different alleles at the considered locus. Here, we compare two alleles and ignore whether they are found in the same individual.). Other classical measures are the number of polymorphic sites in a sample and allele frequencies. The measures of genetic diversity used in this thesis are introduced in the corresponding method section.

The source of genetic diversity are errors in the copying process of genetic data called *mutations*. Mutations can affect the fitness of its carrier in a *beneficial* or *deleterious* way thereby increasing or decreasing the expected number of offspring. Other mutations, called *neutral mutations*, do not confer fitness effects. Note that the effect of a mutation often depends on the genetic background and the environment and can therefore change over time.

The abundance and proper definition of neutral mutations is still debated [23, 38]. Here, we define a mutation as neutral, if it does not influence the fate and reproductive success of its carrier within the limits of time and space addressed in the model. Following this definition, we can safely assume that neutral diversity is abundant.

From an outside perspective, neutral mutations might still seem largely irrelevant. For population genetics, however, they are crucial: neutral mutations do not change the population dynamics and their patterns are used with great success to infer details of the population history such as adaptation processes at neighboring loci [7], population bottlenecks [66], gene flow between subpopulations [50], and range expansions [28]. Nevertheless, different processes can result in very similar patterns and, thus, to misinterpretation [5].

In this thesis, we will deal almost exclusively with neutral genetic diversity and analyze the expected patterns under different scenarios of range expansions.

1.3. Genetic drift

For individuals of most if not all species, life is full of risks and opportunities that can affect their reproductive success drastically. Just think about the influence of the hunting success of predators on their reproductive success — or (more dramatical) think about the reproductive success of its potential prey.

The fate of an allele is essentially determined by the offspring numbers of its carriers and the frequency of the allele will rise or fall as a consequence of the fate of its carriers. In population genetics, the whole of these complex external influences is pooled into the concept of *genetic drift* and modeled as a stochastic process.

The two most fundamental models of population genetics are the *Wright–Fisher model* [31, 104] and the *Moran model* [72]. These models incorporate genetic drift by randomly selecting the individuals that reproduce (and, in case of the Moran model, also by killing individuals at random). Both models increased the general understanding of population genetics and were extended to more complex scenarios, notably to spatially structured populations.

Genetic drift is of particular importance in small populations. Modeled with either the Wright–Fisher model or the Moran model, genetic drift takes an average of N generations to deplete genetic diversity at a neutral locus with N alleles and no new mutations. Alleles that exist only in small numbers can quickly go extinct due to the stochastic fluctuations — even in large populations and even if the allele provides a selective advantage.

1.4. Spatial structure

In *spatially structured populations*, the frequency of alleles varies between different locations in the habitats. Such patterns typically emerge, when migration is limited, that is, if the *natal dispersal distance* of individuals is much smaller than the habitat of the population. See, for instance, [77] for a fascinating application.

The *stepping–stone model* can be considered as one of the most important models for spatially structured populations. Multiple subpopulations are arranged on a lattice and exchange migrants with their direct neighbors. The dynamics within the subpopulations is usually modeled according to the Wright–Fisher model. The free choice of the underlying lattice is the reason for the flexibility of the stepping stone model. On the one hand, fundamental concepts such as the *isolation by distance* can be described analytically based on regular lattices (for instance circular or toroidal habitats, *e.g.* [94, 90], infinite habitat models, *e.g.* [56]). On the other hand, fine tuning of subpopulation sizes and migration rates allow simulation studies for realistic landscapes [19, 60].

1.5. Range expansions

As mentioned in the first paragraph, range expansions are ubiquitous. If the expanding population is spatially structured, range expansions amplify genetic drift: the individuals that colonize the new areas are offspring of a relatively small number of individuals that live close to the expansion front. As a consequence the individuals in the newly colonized area carry only a subset of the populations genetic diversity. When the range expansion proceeds, this *founder effect* occurs continuously and the genetic diversity decreases along the expansion axis.

While well–mixed models cover only population growth (*e.g.* [65, 92]), stepping–stone models are ideal to model range expansions (*e.g.* [29, 28, 1]).

Range expansions have attracted attention in the context of colonizations and invasions long before the advent of DNA sequencing (*e.g.* [3], but see [28] for an excellent review). The most striking genetic consequences of range expansions are the loss of neutral genetic diversity along the expansion axis [2] and so called *gene surfing* [60, 40]. A gene (or allele in our notation) surfs if it fixates locally at the expansion front and travels with the wave of advance. Surfing alleles can increase heavily in number, the local fixation leads to genetic de–mixing apparent in sectoring patterns [40]. Consequently, clines in neutral genetic diversity are used to identify range expansions (*e.g.* [45, 33]). Clear sectoring pattern were, so far, only observed in microbial experiments (and simulations).

Note that range expansions are not the only scenarios that can produce clines of genetic diversity. Nick Barton and colleagues [5] showed that recurred selective sweeps, that is, fixations of new beneficial mutations, can lead to patterns at neutral loci that are commonly interpreted as signs of range expansions.

1.6. The coalescent

The *coalescent* is the ancestral process in population genetics. Starting from a sample of alleles, the coalescent models the sample's genealogical tree. Note that the actual genealogical tree of a particular sample is not random in itself but the result of a series of (almost always) unknown random events in the past. The coalescent is the stochastic model that accounts for our inevitable lack of information.

The coalescent was mathematically established by J. F. C. Kingman [59, 58] in the early 1980's and is based on the concept of identity by descent: two alleles are called *identical by descent* if they are copies of the same ancestral allele. Clearly, this definition is not complete as every two alleles must have a common ancestral allele, even if the ancestor has lived many generations in the past. The definition of identity by descent requires the choice of a timeframe.

Coalescent theory refines the concept by describing (the distribution of) the time to the *most recent common ancestor* (MRCA) of two or more alleles.

The underlying idea of both concepts is the same: the sampled alleles contain genetic material that has been copied and transmitted from one generation to the next from a single 'original' copy since the time of the MRCA (t_{MRCA}). In the MRCA, the genetic information was identical and the genetic information of two sampled alleles differs if and only if a mutation occurred on one of the lineages of the genealogical tree linking the two alleles.

The coalescent is defined for sample sizes $2 < n \leq N$, where n refers to the sample size and N refers to the number of alleles in the population. At diploid loci, there are $2N$ alleles in a population of N individuals. Therefore, $2N$ is often used for the number of alleles. Figure 1.1 shows the typical graphical representation of the coalescence of five alleles.

Under the commonly used infinite sites model [101, 44], each new mutation occurs at a previously non–polymorphic locus. Consequently, each segregating site in a sample corresponds to a mutation on the genealogy (Figure 1.1). Multiple mutations can occur along the same edge of the tree but due to the tree structure, each edge of the tree along which a mutation occurred gives rise to a unique mutant–wildtype bipartition of the sample.

Range expansions remain an active topic of research. It is the major goal of this thesis to develop models and methods that help to disentangle the various impacts of range expansion on genetic diversity and to develop observables that allow to distinguish the impact of range expansions from other influences. Our analysis will, for the largest part, rely on spatial coalescent models.

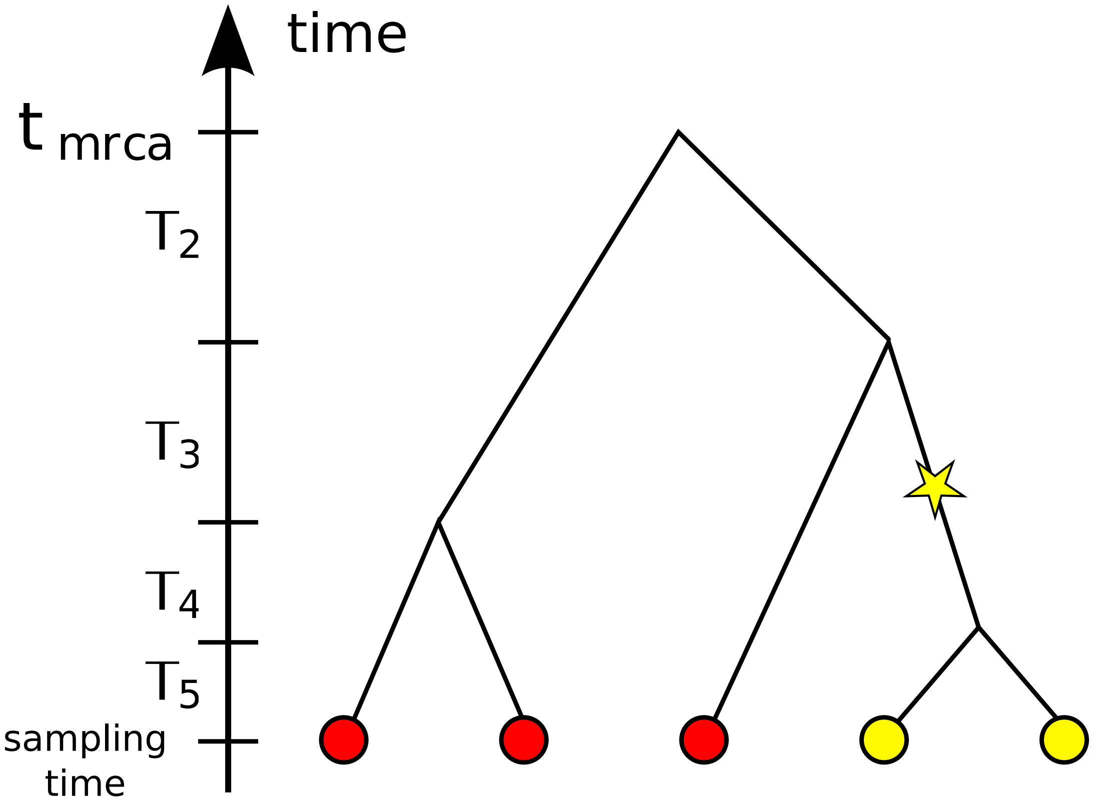


FIGURE 1.1.: **Example of a coalescent tree.** 5 lineages coalesce into the *most recent common ancestor* at time t_{mrca}. The length of the tree can be calculated as $T_{\text{total}} = \sum_{i>1} iT_i$. The mutation (yellow star) bipartitions the sample into 2 mutants (yellow discs) and 3 wild–type individuals (red discs). Each edge corresponds to a different bipartition.

Part I.

The coalescent in boundary-limited range expansions

The research described in this part is published under the title *The Coalescent in Boundary-limited Range Expansions* in *Evolution, International Journal of Organic Evolution* [78]. The part contains to the most part the original text of the paper. Therefore, minor repetitions occur between the general introduction and the introduction to this part.

The second part of the paragraph *'Coalescence time distribution far from the boundary for vanishing convection speeds'* and the paragraph *'Coalescence time distribution close to the boundary for vanishing convection speeds'*, both from the Appendix of the paper, were derived exclusively by Oskar Hallatschek and are therefore not included in this thesis. The supplementary information has been merged into the main text and into the appendix.

Abstract

Habitat ranges of most species shift over time, for instance due to climate change, human intervention, or adaptation. These demographic changes often have drastic effects on the genetic composition of the population, such as a stochastic resampling of the gene pool through the "surfing" phenomenon. Most models assume that the speed of range expansions is only limited by the dispersal ability of the colonizing species and its reproductive potential. While such models of "phenotype-limited" expansions apply for instance to species invasions, it is clear that many range expansions are limited rather by the slow motion of habitat boundaries, as driven for instance by global warming. Here, we develop a coalescent model to study the genetic impact of such "boundary-limited" range expansions. Our simulations and analytical calculations show that the resulting loss of genetic diversity is markedly lower than in species invasions *if* large carrying capacities can be maintained up to the habitat frontier. Counterintuitively, we find that the total loss of diversity does not depend on the speed of the range expansion: Slower expansions have a smaller *rate* of loss, but also last *longer*. Based on our results, we conclude that boundary-limited range expansions have a characteristic genetic footprint and should be distinguished from range expansions limited only by intrinsic characteristics of the species.

2. Introduction

Although the distribution of many common species seems stationary for years or even centuries, habitats do frequently change over the long time scales relevant to evolution. Glacial cycles, for instance, recurrently led to the contraction and expansion of species ranges [46, 47]. The warming after the Last Glacial Maximum gave rise to a massive northward range expansion of temperate species on the northern hemisphere. In the recent past, habitat ranges have started to shift in response to global warming [8, 80, 83, 16]. Human interventions influence species distributions on still faster time scales, for instance by providing new migration opportunities [63, 10], or by transforming landscapes [35](*e.g.* construction of roads, cultivation of fields). This has led to many species invasions in non-native habitats over the last centuries, often with dire consequences for the resident species.

Population genetics is well equipped for dealing with stable demographies [44]. However, understanding and quantifying evolutionary change of populations far from equilibrium remains one of the major challenges in population genetics. Range expansions are particularly important non-equilibrium scenarios because they are expected to have strong impacts on the gene pool of the population [30, 42]. Existing models of range expansions are applicable mainly to invading populations that expand freely into pristine territories [2, 40, 1]. In the absence of long distance dispersal and major spatial heterogeneities, the population density at the invasion front takes the form of a traveling wave [32, 61]. The velocity of such population expansions primarily depends on the dispersal rates of the species and its reproduction rate, and only weakly on the carrying capacity. Regions close to the front are not at carrying capacity because of the limited reproductive time since first colonization. As a consequence, the population density of a population wave gradually decreases towards the front of the range expansion. As such range expansions depend on phenotypic characteristics of the considered species (dispersal, reproduction, etc.), they will be referred to as *phenotype-limited* range expansions in the following. These range expansions have been shown to have a pronounced effect on genetic diversity. In the absence of long distance dispersal, only the descendants of a small founder population close to the expansion front will contribute to colonization of pristine territory [41, 91]. The population front provides a continual population bottleneck with the consequence to reduce the genetic diversity. The resulting decline in genetic diversity has been observed in various species [28], and demonstrated on the micro scale of expanding bacterial populations [40, 43]. In humans, one has detected a significant though relatively weak decrease in genetic diversity along the presumed migration routes during the expansion out of Africa [85, 84]. This decline in heterozygosity with distance to the source population has been predicted to be linear with the distance to the expansion front [85, 21]. Related to the phenomenon of a small continual bottleneck at the expanding front is the phenomenon of gene surfing, by which neutral variants can rise to high frequency by the action of strong genetic drift [25, 60, 41, 30, 98, 64]. In two dimensions, genetic drift has been shown to give rise to a characteristic sectoring pattern that can mimic very closely the patterns expected for selective sweeps in spatially structured environments [5].

The coalescence process in expanding population waves is still relatively unexplored. In linear habitats, the mean coalescence time has been shown to be controlled by the logarithm of the population size, which might be considered as an effective population size of the expanding front. The coalescence process in the front population, however, is characterized by frequent multiple mergers due to pronounced founder effects [12, 76].

In contrast to these “phenotype-limited” range expansions, many range expansions are limited by a gradual change in environmental conditions rather than any phenotypic trait of the species. A prime example is the slow shifting of species ranges due to a gradual climate change [71]. Often in such cases, the velocity of the range expansion is considerably smaller than the potential invasion speed of the species. The shifting of the climatic isotherms in North America and Europe, for instance, amounts to about 1km per year since 1900 [49]. Many species have shifted their habitat range of the same order [81, 18, 8] but for instance the Sachem Skipper butterfly moved its northern range limit by 75km in a single year with warm winter [80]. The Sachem butterfly habitat is limited by minimal winter temperatures and the strong expansion within one year shows the potential for a much faster expansion than actually realized. Range expansions with an expansion velocity limited by external constraints to values significantly below the potential phenotype-limited velocity will be referred to as “boundary-limited” range expansions.

Boundary-limited range changes are frequent: climate change is a recurrent phenomenon on earth, and leads to gradual shifting of climatic isotherms (longitude and altitude), change of sea levels, and the formation and meltdown of glaciers. Despite the frequency of these events, the associated impact of boundary limited range expansions are to a large extent unexplored theoretically.

To fill this gap, we develop a simple null-model of boundary-limited range expansion, and apply population genetics theory to reveal the resulting patterns of genetic diversity. We take a retrospective view on the dynamics and determine the ancestry of a pair of lineages that are sampled at a certain distance from the expanding frontier. Figure 2.1 illustrates the generic dynamics of the ancestral process [59, 51]. Backward in time the lineages follow a random walk through the habitat, and eventually coalesce in their most recent common ancestor. The time to the most recent common ancestor controls how many genetic changes both lineages could have accumulated, and is therefore a measure of the genetic diversity. Two coalescence scenarios may be distinguished: In the *free* phase of coalescence, lineages coalesce prior to being influenced by the moving population frontier. On the other hand, if lineages avoid coalescence for a sufficiently long time they will be captured by the population front and continually pushed towards the ancestral habitat. The frequent reflections at the front induce frequent encounters of the lineages and thus enforce a large rate of coalescence. The moving front thus divides the coalescence process into a first phase free from short term impact of the front (*free* phase of coalescence) and a second phase in which the front enforces more frequent encounters (*enforced* phase of coalescence). We show that this dichotomy is useful as it allows us to extend the coalescence theory of stationary habitats [68, 102] to the case of moving boundaries.

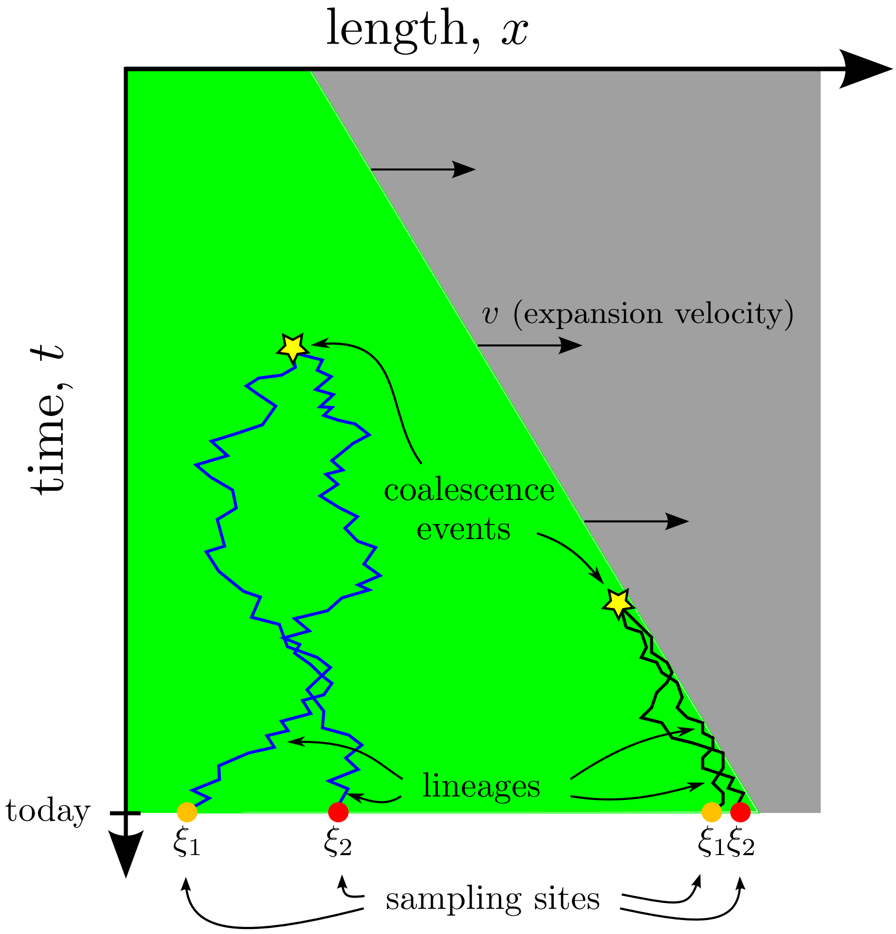


FIGURE 2.1.: **The coalescent in boundary-limited range expansion.** This sketch illustrates the genealogies emerging in a habitat that is slowly expanding, e.g., due to a gradual climate change. The moving habitat boundary is represented in this space-time diagram by the diagonal line separating the habitable region (green) from the empty region (grey). The habitat is largest at present time (bottom) and smallest at earliest time (top). Imagine sampling two lineages at present time from two locations (green and orange circles). Backward in time, these lineages carry out an unbiased random walk through the population of ancestors until they eventually coalesce (stars) in their most recent common ancestor. Two generic coalescence scenarios are depicted. *Free coalescence* is illustrated by the blue pair of lineages, which encounter and coalesce before they are influenced by the moving boundary. If two lineages avoid coalescence for a sufficiently long time (black lines), they are instead "collected" by the moving boundary, which is pushing the lineages into the ancestral habitat. This leads to rapid *enforced coalescence* because the lineages are effectively caged in a small subpopulation in front of the moving boundary. We validate this intuitive coalescence picture in section 4 and show that it can be used to readily characterize the genetic diversity of the expanding population.

3. Model

In order to be able to explore genealogies in boundary-limited range expansions, we trace lineages backward in time within two types of population structures. The first type is a linear stepping stone model [57] with a moving boundary, which is spatially discrete and one dimensional. This model allows us to develop and verify a basic intuitive and mathematical picture of the coalescence process. In a second step, we validate our theory by simulating the coalescent in a more realistic second population structure, which is continuous in space and two-dimensional. In a boundary limited range expansion, migration and population growth is assumed to be faster than the habitat expansion. We therefore impose for both population structures that the populations are everywhere at carrying capacity. For most of our simulations, these carrying capacities are also assumed to be the same everywhere in the habitat and in particular close to the population frontier. This feature has two important consequences for the coalescence process. First, the probability of coalescence is independent of the location at which the lineages intersect. Second, the movement of the lineages can be assumed to not dependent on the movement direction [102]. In the course of our analysis, we will generalize our simulations to the case where the carrying capacities gradually decline (over a given length scale) towards the edge of the boundary. This scenario may apply to the case where the suitability of the habitat deteriorates towards the edge of the habitat, due to a gradient in environmental conditions (temperature, resources, etc.).

3.1. The expanding stepping stone model (linear)

The population consists of a linear array of subpopulations, called demes, that each harbor K individuals if carrying capacities are constant, see Fig. 3.1(a). Migration occurs between neighboring demes at rate m. New demes are added to the moving end of the population at a constant rate v, which leads to a continual expansion of the habitat. All newly added demes are *fully occupied*. Within this demographic structure, our coalescent simulations follow pairs of lineages sampled from specific sampling locations, denoted as ξ_1 and ξ_2, backward in time until they coalesce, as illustrated in Figure 2.1. Thereby, lineages randomly hop between demes, again at rate m, and are reflected when they collide with the moving boundary. Reflecting boundary conditions were chosen because one has to require that every ancestor is born within the habitat. Finally, when two lineages jump into the same deme they undergo coalescence at rate $1/K$, which is the coalescence probability per generation in a well-mixed population of K haploid individuals [59, 51]. Simulations for gradually declining carrying capacities were carried out analogous to the above algorithm, with the exception that the deme sizes K_i were assumed to be decreasing towards the moving boundary, according to a logistic function of characteristic width W (see also Appendix Model Details). Note that a variable deme size K_i not only modifies coalescence rates (given by $1/K_i$) but also changes the migration rates for ancestral lineages: the rate at which lineages jump from deme i into deme $i+1$ is proportional to the number of migrants mK_{i+1} that came from the target deme divided by the size K_i of the source deme.

3.2. The continuous model (planar)

The habitat is a stripe of constant width k and has an expanding front at one side, see Figure 3.1(b). For simplicity, periodic boundary conditions are imposed along the non-moving edges of the habitat. Time is still measured in discrete generations but the displacements of the lineages are now drawn from a two-dimensional Gaussian distribution with vanishing mean and standard deviation σ. If the distance between the two lineages (after dispersal) is smaller than a coalescence distance δ, coalescence occurs with probability $1/K_{2D}$.

linear habitat model		planar habitat model	
Symb.	Meaning	Symb.	Meaning
m	Migration rate	σ	Standard deviation of natal dispersal distance
v	Front velocity	v	Front velocity along expansion axis
		δ	Coalescence distance *
K	Deme size	K_{2D}	Neighborhood size
		k	Habitat width
l	Habitat length	l	Habitat length

* Lineages coalesce at a constant rate $1/K_{2D}$ if the distance between them is smaller than δ.

Table 3.1.: **Parameters of our simulation models.**

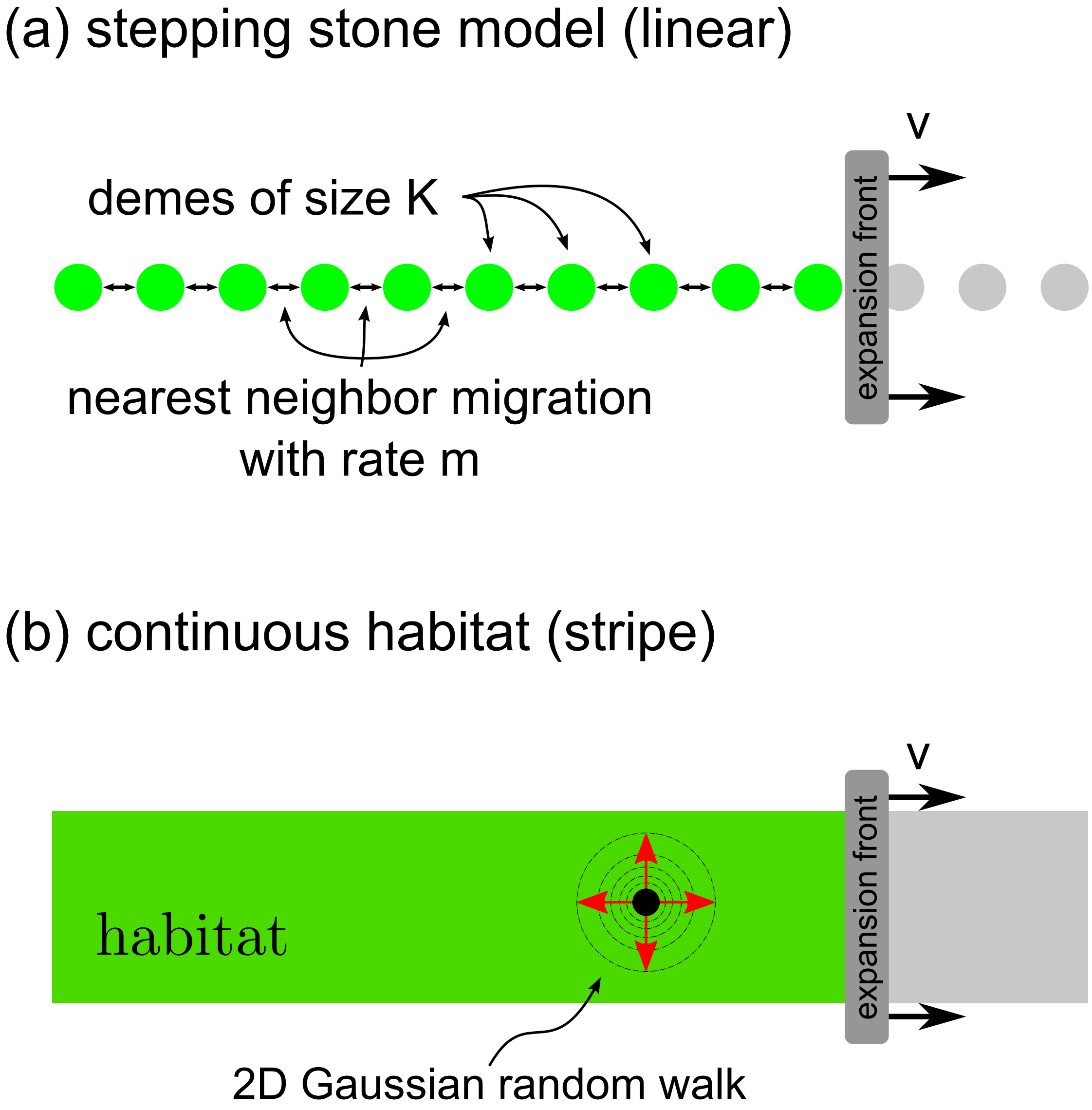

(a) stepping stone model (linear)
v
demes of size K
expansion front
nearest neighbor migration
with rate m
(b) continuous habitat (stripe)
v
habitat
expansion front
2D Gaussian random walk

FIGURE 3.1.: **Two models of boundary-limited range expansion.** In the expanding stepping stone model (a), the population is represented by a linear array of demes, which harbor K haploid individuals. Individuals jump to neighboring demes at rate m. At the moving end of the habitat, new demes are added at a constant rate v. In the continuous model (b), the population is represented by a stripe like habitat of width k with a constant population density. Periodic boundaries are imposed at the non-moving edges. Individuals migrate according to a two-dimensional Gaussian kernel with variance σ^2. The basic parameters of our model are summarized in table 3.1.

4. Results

4.1. The expanding stepping stone model (linear)

In a first step, we sampled lineages from the same deme at a distance ξ from the boundary, ran our coalescence simulations of the linear expansion model at least 10^4 times and recorded the coalescence times T_c. Figure 4.1 depicts the simulated mean coalescence times $\langle T_c \rangle$ averaged over all runs. As expected, $\langle T_c(\xi) \rangle$ increases monotonically with sampling distance ξ to the moving boundary. We observe two qualitatively different regimes: i) a plateau regime close to the expansion front and ii) a regime with a shoulder and a (seemingly) square-root relationship between the coalescence time and sampling distance, $\langle T_c \rangle \sim \xi^{1/2}$. Our analytical results will indeed show that this power law relationship is to be expected for large sampling distances.

Next, we display for several sampling locations the cumulative distributions of coalescence times. Figure 4.2 shows the probability $p(t, \xi)$ of no coalescence before time t for a pair of lineages sampled from the same deme at a distance ξ from the front. The data confirms the hypothesized division of the coalescence process into two phases: At short times the data is perfectly described by the known analytical results for the coalescent in an infinite habitat without boundaries (c.f. equation (4.1)) [68]. At a certain time, which depends on the sampling position, there is a sharp drop in $p(t, \xi)$ indicating rapid coalescence of lineages that have survived up to this time. The crossover happens close to the time $t_0 \equiv \xi/v$ at which the expansion front reaches the sampling site ξ of the two alleles. This time is also the expected time for the front to reach the lineages, as they carry out an unbiased random walk starting from ξ. The data is thus consistent with the view that coalescence is unaffected by the moving boundary until the boundary reaches the sampling sites.

4.1.1. Analytical approximation

Based on these observations, we can develop a simple approximation for the coalescent in boundary-limited range expansions. At early times, where the influence

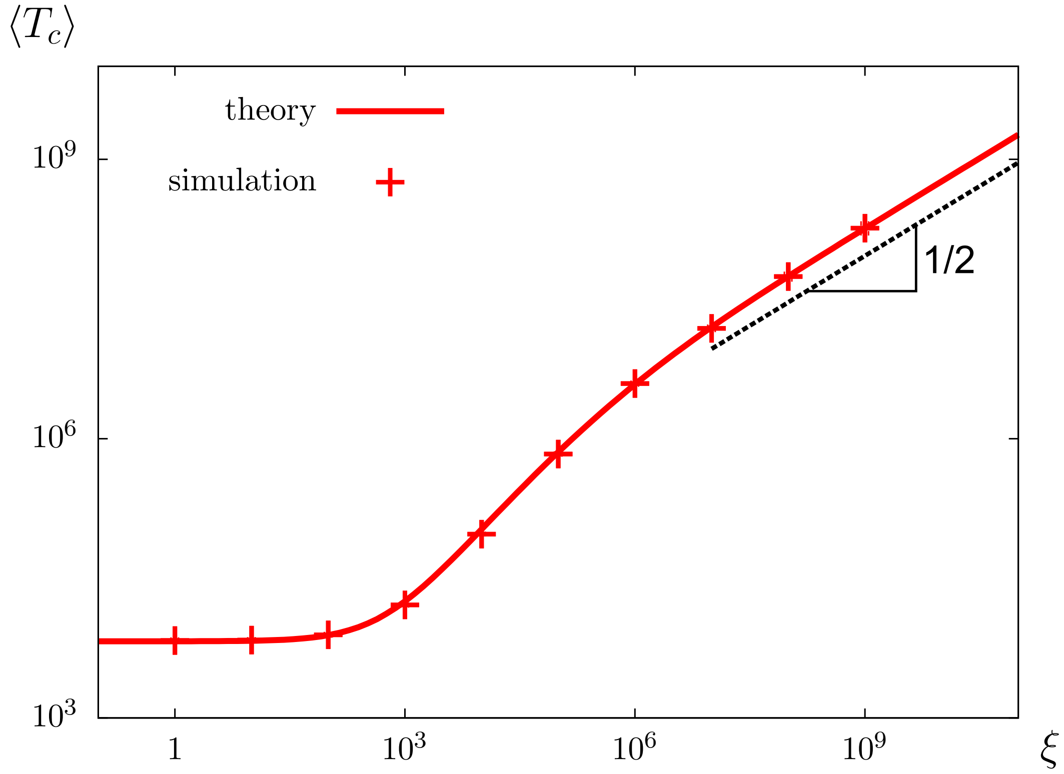


FIGURE 4.1.: **Mean coalescence times in the expanding stepping stone model (linear).** The mean time to common ancestry, or coalescence time $\langle T_c \rangle$, is an important measure for genetic diversity, as it is proportional to the expected number of pairwise nucleotide differences. The plot depicts $\langle T_c \rangle$ for a pair of lineages sampled from the same deme a distance ξ from the moving boundary. The simulations were run for an expansion velocity of $v = 0.1$, deme size $K = 1000$ and a migration rate of $m = 0.33$. Averages have been taken over 10^4 simulation runs. The size of the symbols represent the standard deviations of our estimates. Notice two qualitatively different regimes: For sampling distances $\xi \leq 10^3$, the mean coalescence time is almost independent of the sampling distance (plateau regime). For large sampling distances, we observe an apparent power law with exponent 1/2. The solid red line is our analytical approximation derived from equation (4.2).

of the moving boundary is negligible, we can describe the coalescent by known results for the coalescent in infinite linear habitats *without* boundaries. Assuming that lineages carry out an unbiased diffusive random walk, the probability of

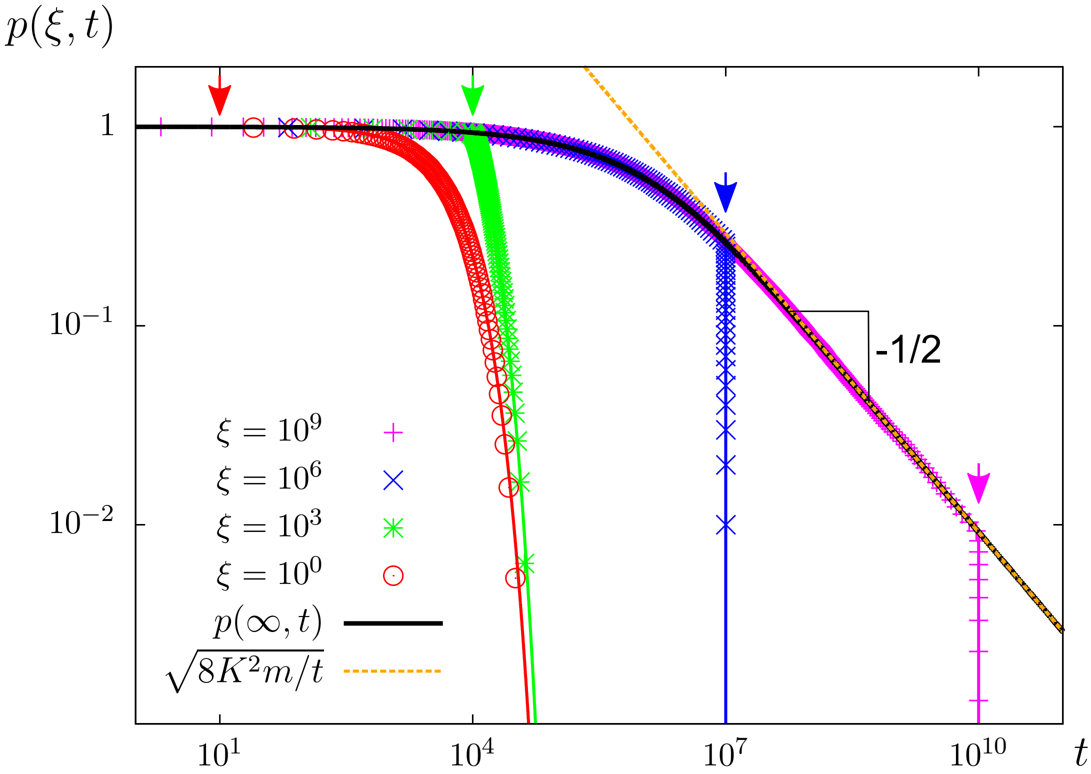


FIGURE 4.2.: **Characterization of the full coalescence time statistics in the expanding stepping stone model (linear).** The graph depicts the probability $p(\xi,t)$ that lineages do not coalesce up until time t when they were both sampled from the same deme at a distance ξ from the expansion front (see legend for the sampling locations). The parameters characterizing the range expansion are the same as in Fig. 4.1 ($K = 1000$, $v = 0.1$, $m = 0.33$). Simulation data of four different sampling distances (crosses, slanted crosses, stars, and circles) are shown along with analytical approximations (solid black line), Eq. (4.2). The asymptotic power law with exponent $-1/2$ is indicated by the yellow dashed line. The arrows indicate for each parameter set the time at which the moving boundary reaches the sampling positions.

non-coalescence up to time t is given by [68, 75, 6]

$$p_{\text{free}}(t) = \operatorname{erfc}\left(\sqrt{\frac{t}{8mK^2}}\right) e^{t/\left(8mK^2\right)}, \tag{4.1}$$

see the Appendix "Analytical Approach" for a derivation. Note that the above formula implies a diverging expected coalescence time in an infinite habitat. This is consistent with the more general result that mean coalescence time in a finite linear habitat is given by the total number of alleles in the population [95, 90, 15], and hence clearly diverges as the habitat ranges are sent to infinity.

At the time $t_0 \equiv \xi/v$ when the moving boundary reaches the sampling site, equation (4.1) ceases to be valid. Instead, the probability $p(\xi, t)$ of non-coalescence in Fig. 4.2 drops sharply because the boundary collects the surviving lineages and forces them to coalesce. This final stage of the coalescence process of a sample of two can be approximated as follows: The lineages are caged in a small collection zone in the vicinity of the front that they explore quite rapidly by random migration. For large deme sizes, this "cloud" of diffusing lineages is therefore effectively well-mixed. The corresponding *effective* population size $N_e \equiv 2Km/v$ can be estimated from the rate at which the lineages meet inside the collection zone, as detailed in the Appendix "Analytical Approach". Thus, once the lineages have arrived in the well-mixed zone, coalescence occurs at rate N_e^{-1}, according to Kingman's coalescent. Under these assumptions, the probability of non-coalescence up to time t is approximated by

$$p(\xi,t) = \begin{cases} p_{\text{free}}\left(t\right), & \text{if } t < t_0 \equiv \xi/v, \\ p_{\text{free}}\left(t_0\right) \exp\left(-\frac{t-t_0}{N_e}\right), & \text{if } t > t_0 \ . \end{cases} \tag{4.2}$$

The case $t > t_0$ consists of a product of two probabilities, firstly the probability to survive the phase of free coalescence and secondly the probability to survive up to time t in the well-mixed phase. Notice that a ξ dependence only enters through the ξ–dependence of the time $t_0 = \xi/v$ at which the boundary arrives at the sampling location.

Our approximation (4.2) for $p(\xi, t)$ is plotted as solid lines in Figure 4.2 and shows very good agreement with our simulation results. By integrating $p(\xi, t)$ over the time variable, we can now derive the mean coalescence time as a function of sampling distance. The resulting predictions reproduce the simulation data as can be seen from Fig. 4.1, where the theory is plotted as a red solid line. The closed form for the mean coalescence time reveals the behavior of the coalescent in the plateau and the power law regimes observed above: At large ξ one indeed finds an asymptotic power law with exponent 1/2. The approach to that regime is, however, rather slow as the marked shoulder in Fig. 4.1 indicates. It is therefore advised to use the full expression rather than the asymptotic results for numerical comparisons. The plateau at short sampling distances corresponds to the effectively well-mixed cloud where both lineages are collected in front of the moving boundary. The fact that coalescence hardly depends on sampling location in this regime is consistent with the concept of an effectively well-mixed collection zone.

The typical size of the region in which lineages are caged once they have been collected by the moving boundary is given by the characteristic length scale

$\lambda = m/v$, which results from the competition of random migration at rate m and the deterministic motion of the boundary with velocity v. Notice that the cage can be a rather loose one, because λ can become large when the expansion velocities are small or the migration rates large. The characteristic time that lineages need to explore the well-mixed region into which they are caged once the boundary has arrived can be estimated by $\tau \equiv \lambda^2/m = m/v^2$. Interestingly, the total length of the plateau region is not set by λ, but instead given by the length $2Km$. This length defines the region for which the waiting time for the moving boundary $t_0 = \xi/v$ is smaller than the coalescence time $N_e \equiv 2Km/v$ in the cloud. Therefore, lineages that are sampled from within a distance of $L_{\text{plateau}} = 2Km$ demes of the moving boundary typically do not coalesce prior to the arrival of the moving boundary, which eliminates the dependence of the coalescence time on sampling location. The characteristic scales in our problem are summarized in table 3.1.

4.1.2. Data collapse

The characteristic time and length scales, τ and λ, define natural units of time and length for our model. This becomes evident, when we present our data in these units of time and space. Figure 4.3 displays the mean coalescence time as a function of sampling location for 6 different parameter sets. We find that all curves that have the same value of Kv fall onto the same curve. This data collapse indicates that Kv represents the only relevant control parameter of the coalescence process. That is, the coalescence process behaves qualitatively similar if one either increases the carrying capacity or the expansion velocity. These intuitive considerations can be further justified by a mathematical description of the coalescence process developed in the Appendix "Analytical Approach". From Fig. 4.3, one can further observe that as one lowers Kv, the plateau region tends to disappear while the asymptotic power law region remains unchanged. The agreement of simulation and the approximation, described above, is still quite good except if lineages are sampled very close to the boundary, where the plateau region seems to disappear.

The disagreement for sampling close to the plateau for $Kv < 1$ is a consequence of the breakdown of our well-mixed approximation close to the front. Our approximation only holds if the characteristic time $\tau \equiv m/v^2$ to explore the cloud is smaller than the effective population size $N_e \equiv 2Km/v$ of that well-mixed region, which sets the time scale for the coalescent. In the regime of very small carrying capacities, such that $Kv < 1$, coalescence occurs essentially immediately when the boundary arrives at the sampling sites. Indeed, if we assume a vanishing effective front population size, $N_e = 0$, we obtain a good description of the data (dashed line in the Figure 4.3, a vanishing front effective population size corresponds to using $p(\xi, t) = 0$ for $t > t_0$ in (4.2).) . Simulations start to deviate close to the boundary

Scale	Explanation
$\lambda = m/v$	Typical range in which the lineages are trapped by the moving population front
$\tau = \lambda/v = m/v^2$	Time that lineages need to explore the well-mixed region of size λ
$L_{\text{plateau}} = 2Km$	Length of the front region of nearly constant genetic diversity (plateau length)
$N_e = 2Km/v$	Effective population size of the well-mixed front population
$t_0 = \xi/v$	Crossover time from the free phase to the enforced phase (cutoff time)

Table 4.1.: **Characteristic time and length scales**

because because our convection-free description of the coalescence process in Eq. (4.1) does not obey the reflecting boundary conditions at the moving frontier. This can be improved using the approach of Ref. [102], as described in the Appendix "Analytical Approach". The resulting expression for the mean coalescence time is shown as dotted line in Fig. 4.3.

4.1.3. Sampling from different locations.

So far, we have explored the coalescence of lineages that were sampled from the same deme, as a measure for the expected number of pairwise genetic differences *within demes*. Now, to determine the diversity *between demes*, we consider the coalescence of lineages sampled from different locations. For definiteness, we sample one lineage directly at the front and the other one at a location ξ_2. As a consequence, no quick coalescence can occur as it was possible for the case of within deme sampling. Accordingly, the probability $p(t, \xi_1 = 0, \xi_2)$ of having no coalescence before time t for a pair of alleles sampled at distance $\xi_1 = 0$ and ξ_2 from the front, respectively, is close to 1 before the cutoff time and decays quickly thereafter. A good approximation for the mean coalescence times can be obtained by simply adding the plateau value and the cutoff time, $\langle T_c \rangle \approx 2mK/v + \xi_2/v$, see Figure 4.4.

4.1.4. Variable front velocities

Our model can be extended to boundaries that do not move at a strictly constant pace, but move according to a time dependent velocity $v(t)$, if variations in speed

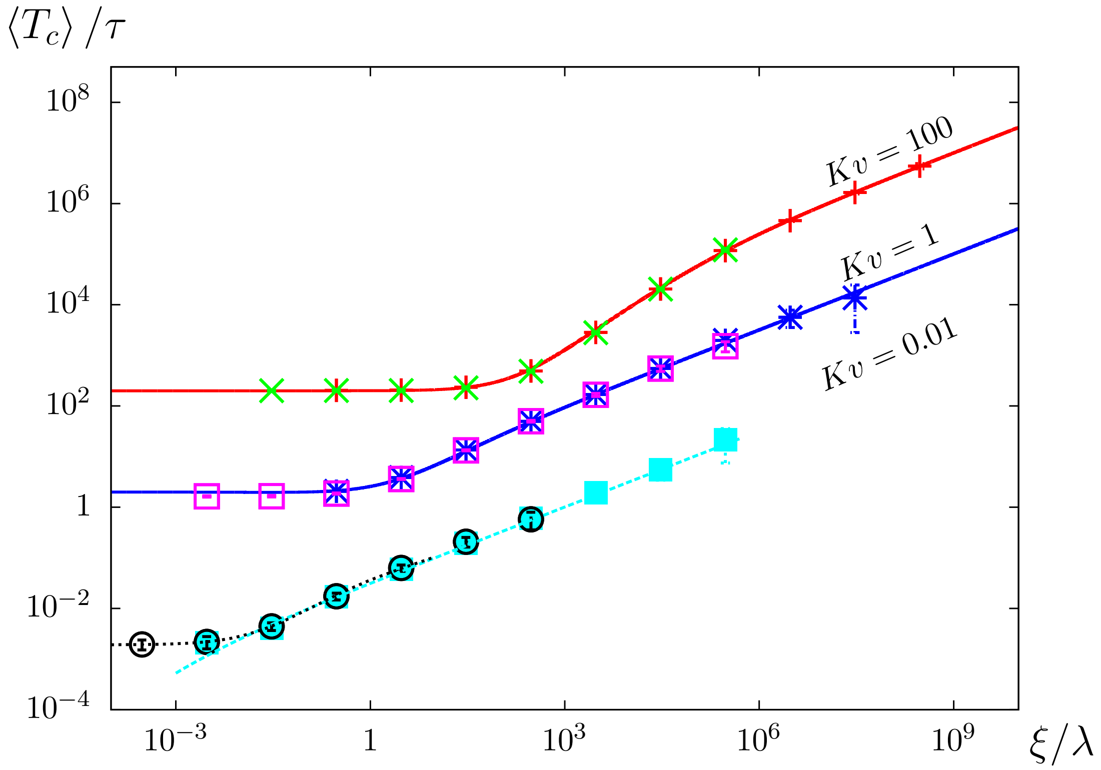


FIGURE 4.3.: **Data collapse for the expanding stepping stone model.** By measuring coalescence times in units of $\tau = m/v$ and sampling distances in units of $\lambda = m/v$, simulation results for given value of Kv collapse on the same curve. The parameter Kv thus controls the qualitative behavior of the coalescence. For $Kv \geq 1$, the coalescence times have a marked plateau for small sampling distances indicating a well-mixed collection zone in front of the boundaries. Our approximation (solid lines) works excellent in this regime. For $Kv \gg 1$, the plateau at short sampling distances disappears, and lineages coalesce essentially immediately when the boundary arrives at the sampling locations. Accordingly, by using a vanishing front population size N_e in our approximation in equation (4.2), we obtain a good description of the data (dashed line). The agreement can be improved even further by taking into account the reflecting boundary conditions at the moving frontier (black dotted line). To demonstrate the data collapse, we have plotted for each value of Kv data from runs with two different parameter sets: $K \in \{10^3, 10^4\}$ for $Kv = 100$; $K \in \{10, 10^3\}$ for $Kv = 1$; $K \in \{10, 10^2\}$ for $Kv = 0.01$. The migration rate was set to $m = 1/3$ in all cases.

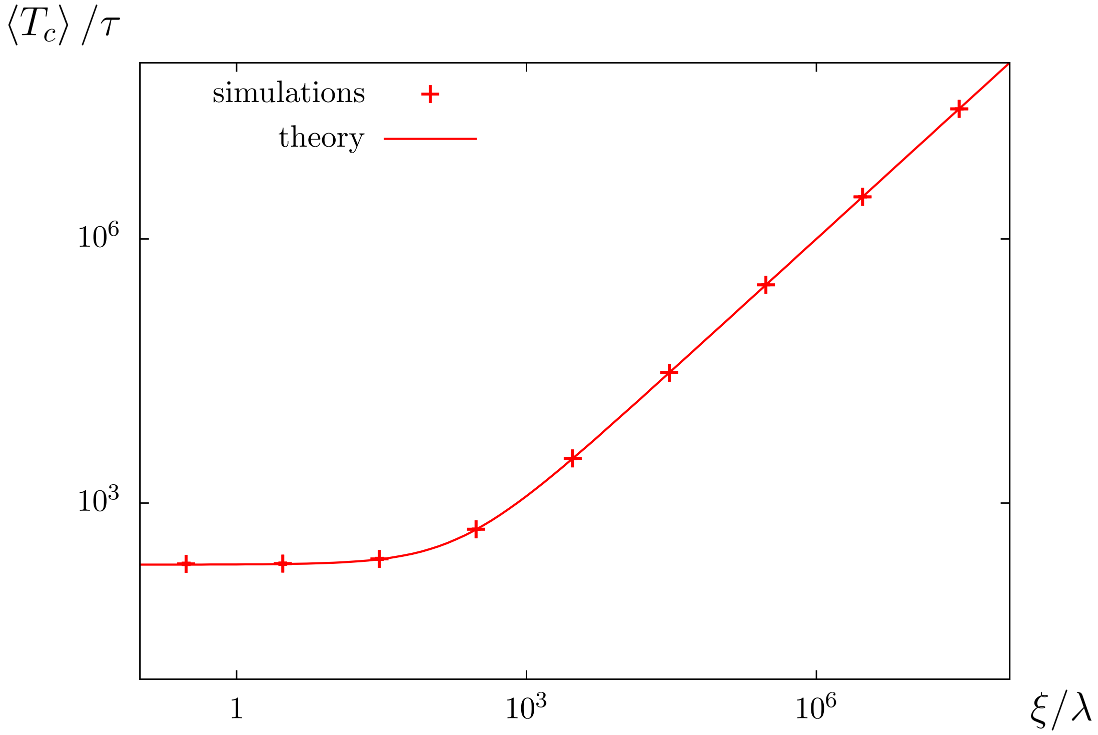


FIGURE 4.4.: **Sampling from different locations** The plot shows the mean coalescence times $\langle T_c \rangle$ of two lineages sampled at distances $\xi_1 = 0$ and ξ_2 from the front, respectively, as a function of ξ_2. The parameters of the population expansion are as in Fig. 4.1. Notice that, for small values of ξ_2, the mean coalescence times are similar to our results for within deme sampling (Figure 4.1), and we observe the same plateau height. For large values of ξ_2, no coalescence can occur at early times as lineages first have to migrate to meet one another. The solid line represents an approximation which is given by the sum of the plateau height and the waiting time $t_0 = \xi_2 / v$ for the population front to arrive at the sampling site ξ_2.

are small. Then, one can apply equation (4.2) using the appropriate time $t_0(\xi)$ for the arrival of the boundary at the sampling site.

4.2. The continuous model (planar)

To test the generality of the results derived from the linear stepping stone model, we also implemented a coalescence simulation for a spatially continuous habitat (see also Model section above). The form of the habitat is a stripe of width k, see Fig. 3.1(b), with periodic boundary conditions along the non-moving edges of the habitat. In this spatially continuous setting, individuals disperse according to a Gaussian kernel with variance σ^2 and coalesce at rate $1/K_{2D}$ when they are closer than a distance δ, which we typically choose to equal σ. The mean coalescence times as a function of sampling distance are depicted in Figure 4.5 for several parameter sets. Notice that all data for which the product of K_{2D} and habitat width k collapse onto the same curve. Furthermore, these curves are described by a corresponding one-dimensional stepping stone, for which we chose an effective deme size of $K \equiv k\ K_{2D}/(\pi\delta^2)$. This deme size was chosen to obtain the correct coalescence rates per unit time if lineage positions were uncorrelated in the direction transverse to the expansion direction. Our analytical approximation derived for the one-dimensional case are expected to break down on times shorter than the time for lineages to transverse the width of the habitat stripe. On the time scales measured in our simulations, however, the one-dimensional approximation yields excellent results.

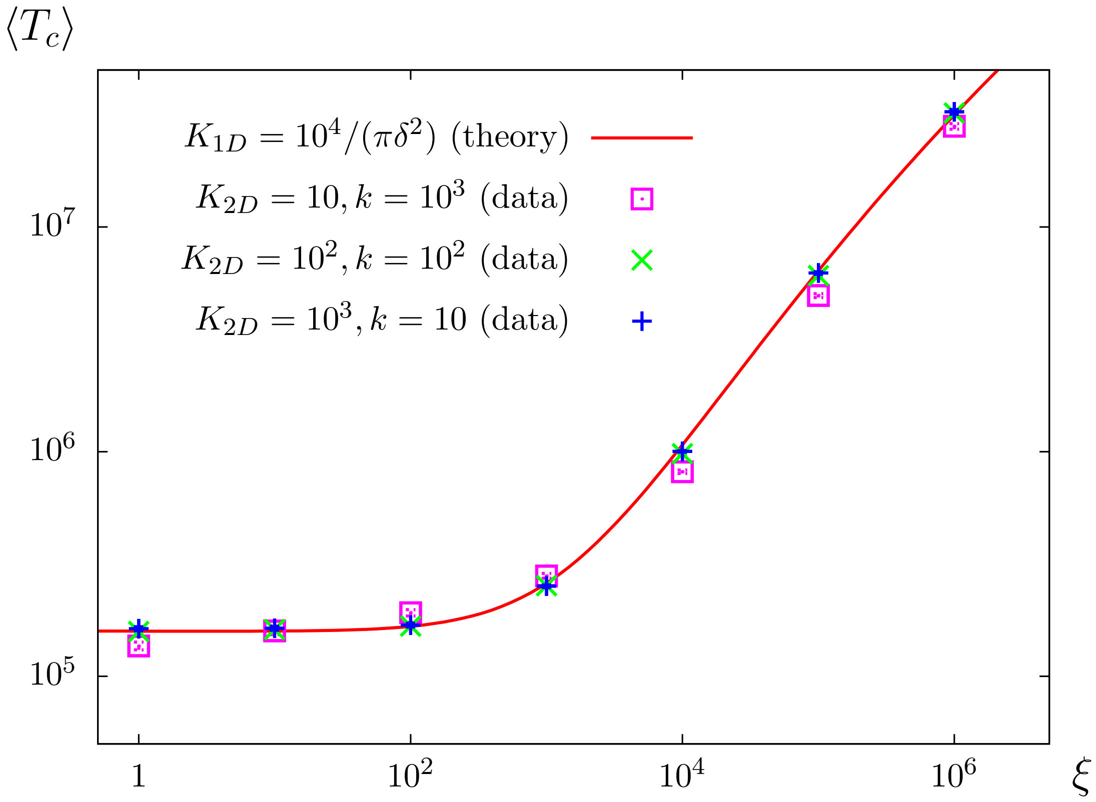


FIGURE 4.5.: **Mean coalescence times in the spatially continuous model (stripe-like habitat).** For the simulation data shown here, the product of the habitat width and the local population size was kept constant: $k \times K_{2D} = 10^4$. Notice that the data for $K_{2D} = 10^2$ and $K_{2D} = 10^3$ collapse on the same curve. Furthermore, this curve is accurately described by our approximation for the linear stepping stone model when effective parameters are used (see Text). Other simulation parameters were $v_x = 0.01$, and $\delta = \sigma = \sqrt{4/3}$. The corresponding parameters for the linear habitat model are $m = 1/3$, $v = 0.01$, and $K = 10^4/(\pi\delta^2)$.

4.3. Population density clines

As motivated in the Introduction, slow boundary-limited range expansions are expected to be at carrying capacity everywhere in the habitat because the population can keep up with the expanding boundary by migration and growth. However, the carrying capacity itself is not necessarily constant throughout the habitat, as assumed so far. Indeed, in situations where the suitability of the habitat gradually decreases towards the habitat frontier, we must expect a cline in carrying capacity.

To explore the genealogies in a moving population density cline, we modified our stepping stone model by assuming that carrying capacities gradually decline towards the edge of the boundary (see Model section). Specifically, we considered a logistic density profiles, and varied the width W of the cline. Fig. 4.6 depicts the resulting mean coalescence times as a function of sampling site. We find that if the cline is narrower than the characteristic scale $\lambda = m/v$, introduced above, the coalescence picture is nearly identical to the boundary limited case studied above. However, as soon as the width of the cline becomes comparable to λ or larger, the plateau disappears and the mean coalescence time vanishes as one samples close to the expanding frontier. These observations reminiscent of unconstrained range expansions, which are characterized by a very small population bottleneck at the front of the population and thus very small coalescence times within the founder population. This comparison is underscored by the fact that unconstrained range expansions have a typical front width of $\lambda = m/v$ [32, 61, 41], which is the precisely the length where our moving clines start to show strong founder effects. We thus conclude that boundary-limited range expansions with clines of width larger or equal to λ lead to diversity loss of similar magnitude as unconstrained range expansions. Intuitively, the cline tightens the population bottleneck of the range expansions because it isolates the few founders at the tip of the cline. Founders that are a distance λ or larger ahead of the saturated demes can propagate their genes into the new habitat with hardly any competition through gene flow from the bulk of the population.

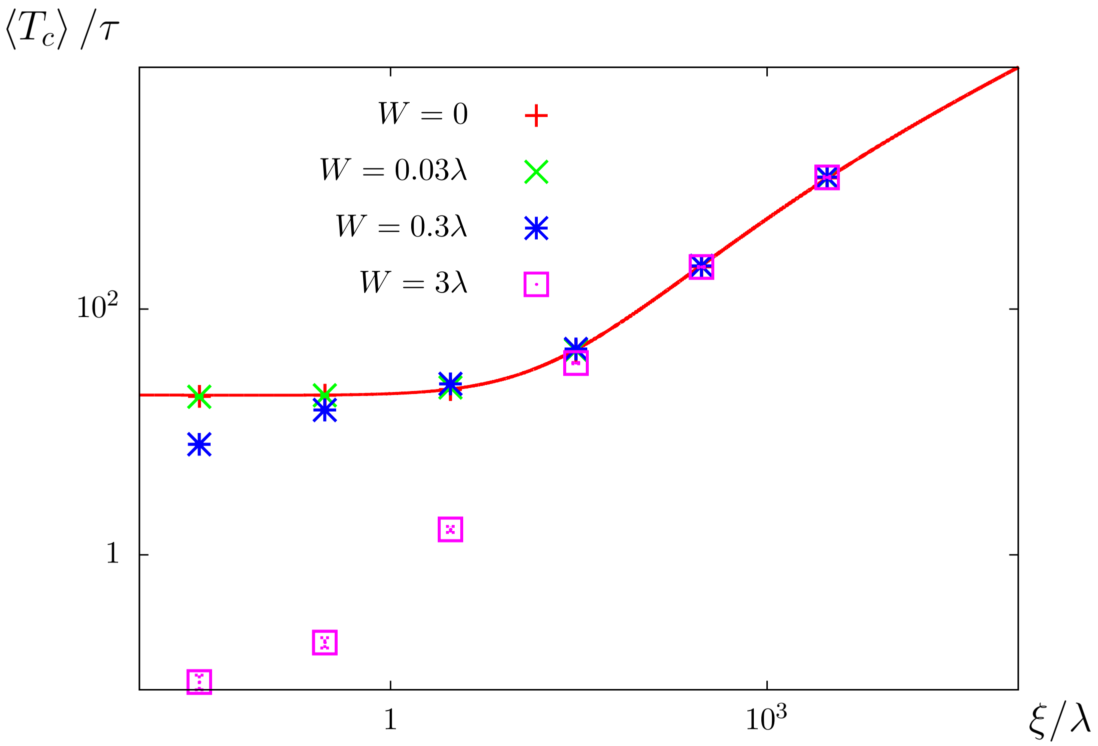


FIGURE 4.6.: **Moving clines of carrying capacity.** Gradients in environmental conditions (resources, temperature, etc.) often lead to a deterioration of the habitat quality towards the edge of a species range. This can lead to a cline in carrying capacity, or deme sizes. To investigate this effect, we simulated the coalescence process for an expanding stepping stone model with a decline in deme sizes towards the edge of the population. The form of the cline followed a logistic function, and the length of the cline varied from 0 to 3λ demes. Notice that, when the cline length is smaller than the characteristic scale $\lambda \equiv m/s$ (here 30 demes), the results are similar to the case with constant deme sizes. When, however, the cline length is larger than λ, the coalescent becomes similar to an unconstrained range expansion, with very small coalescence times close to the expanding edge. Further parameter were $v = 10^{-3}$ and $K = 10$ and $m = 0.33$.

5. Analytical approach

In this appendix, we describe how the coalescence process of two lineages in a boundary-limited range expansion (one dimensional) can be described mathematically. We begin with an exact continuum description in terms of diffusion processes. The resulting equations of motion cannot be solved analytically, but we derive very accurate approximations in the following paragraphs. Finally, we re-express the dynamics in characteristic units of space and time, which shows that the dynamics is controlled by one parameter, namely the product Kv.

5.1. Equations of motion

We consider the coalescence process of two lineages, sampled at locations ξ_1 and ξ_2, in a continuous model of a boundary-limited range expansion. Assuming spatially constant population density in the habitat (see main text), we know that the lineage positions perform an unbiased random walk backwards in time. We further assume that we can approximate the random walk by a diffusion process with diffusivity $\sigma^2/2$. Here, σ^2 is the variance in dispersal distance per generation, which is equal to the migration rate m in our stepping stone model. Both lengths are measured in units of deme separations. In the reference frame co-moving with the boundary, lineages acquire a bias towards the front of velocity. This amounts to an average drift term of velocity v. Let $f = f_{\xi_1,\xi_2}(x_1, x_2, t)$ be the probability density to find the lineages at x_1 and x_2 at time t conditional of having not coalesced. Then, f satisfies the diffusion equation

$$\begin{aligned}
\partial_t f &= \frac{\sigma^2}{2}(\partial_{x_1}^2 + \partial_{x_2}^2) f - v(\partial_{x_1} + \partial_{x_2}) f - \frac{1}{K}\delta(x_1 - x_2) f && (5.1)\\
f \restriction_{t=0} &= \delta(\xi_1 - x_1)\ \delta(\xi_2 - x_2), && (5.2)\\
0 &= \lim_{x_i \to \infty} f \quad \text{for } i \in \{1,2\}, && (5.3)\\
0 &= \left(\frac{\sigma^2}{2}\partial_{x_i} - v\right) f \restriction_{x_i=0} \quad \text{for } i \in \{1,2\}. && (5.4)
\end{aligned}$$

The term $-\frac{1}{K}\delta(x_1 - x_2) f$ in Eq. (5.1) accounts for coalescence events when the two lineages meet, at a rate proportional to the inverse carrying capacity. Notice that coalescence is represented by a *loss* term because f is defined to be conditional on non-coalescence. The initial condition (5.2) fixes the initial sampling location of both lineages at ξ_1 and ξ_2, Eq. (5.3) ensures that the probability of finding lineages decays to 0 at large distances. Finally, the reflecting boundary conditions in Eqs. (5.4) ensure that there is no diffusion current through the moving boundary.

5.2. Coalescence time distribution far from the boundary for vanishing convection speeds

The above system of equations cannot generally be solved in a closed form. For our approximations, however, it is merely necessary to know the solution for large distances from the boundary up to the time the boundary arrives at the sampling sites. For this purpose, we can assume the habitat is infinite. Then the probability of non-coalescence becomes a function only of the separation $\zeta = |\xi_1 - \xi_2|$ of the sampling distances.

Specifically, let $g_\zeta(z,t)$ be the probability density that a pair of lineages separated by a distance ζ at time 0 reaches separation $z = |x_1 - x_2|$ at time t without coalescence. $g_\zeta(z,t)$ then satisfies

$$\begin{aligned} \partial_t g &= \sigma^2 \partial_z^2 g, && (5.5)\\ g \restriction_{t=0} &= \delta(z-\zeta), && (5.6)\\ \lim_{z\to\pm\infty} g &= 0, && (5.7)\\ \sigma^2 \partial_z g \restriction_{z=0} &= \tfrac{1}{2K} g \restriction_{z=0} . && (5.8) \end{aligned}$$

The Laplace transform $G_\zeta(z,s) \equiv \int_t \exp(-st) g_\zeta(z,t)$ of g reads:

$$G_\zeta(z,s) = \frac{e^{-\sqrt{s}|\zeta - z|/\sigma}}{2\sigma\sqrt{s}} + \frac{2K\sigma\sqrt{s} - 1}{2K\sigma\sqrt{s} + 1} \frac{e^{-\sqrt{s}(z+\zeta)/\sigma}}{2\sigma\sqrt{s}} \tag{5.9}$$

Upon integrating over z, we obtain the Laplace transform of the total probability of non-coalescence up to a given time,

$$\int_0^\infty dz\, G_\zeta(z,s) = \frac{1}{s}\left(1 - \frac{e^{-\sqrt{s}\zeta/\sigma}}{2K\sigma\sqrt{s}+1}\right) \tag{5.10}$$

For simplicity, we focused in the main text mostly on the case of sampling from the same deme. We therefore choose $\zeta = 0$, and carry out the inverse Laplace transform. This yields the probability of non-coalescence at time t for sampling from the same deme in an unbounded habitat,

$$p_{\text{free}}(t, \zeta = 0) = \exp\left(\frac{t}{4K^2\sigma^2}\right) \operatorname{erfc}\left(\sqrt{\frac{t}{4K^2\sigma^2}}\right) , \tag{5.11}$$

where $\operatorname{erfc}(\cdot)$ denotes the complementary error function. Upon identifying $\sigma^2/2$ with the parameter m of our stepping stone simulations, we thus obtain Eq. (4.1). The result is compatible with our simulation data up to the time where the boundary reaches the sampling sites (c.f. Figure 4.2). For large values of t, the above expression asymptotes towards

$$p_{\text{free}}\,(t, \zeta = 0) \sim \sqrt{\frac{4K^2\sigma^2}{t}} \ . \tag{5.12}$$

5.3. Effectively well-mixed front population

In the main text, we have argued that the front population is well-mixed in boundary-limited range expansions, if $Kv \geq 1$. Here, we determine the associated effective population size N_e.

Our argument is based on a time-scale separation between mixing and coalescence: We assume that lineages explore their cage in front of the moving boundary more quickly than the time it takes for them to coalesce. Under this strong migration assumption [74], lineages coalesce at a rate

$$N_e^{-1} \equiv \int_0^\infty dx \psi(x)^2 K^{-1}(x) \tag{5.13}$$

in the continuum approximation of the stepping stone model. In equation (5.13), $K(x)$ is the carrying capacity at distance x from the moving boundary, and $\psi(x)$ is the probability density that a lineage visits location x at equilibrium. The rational behind (5.13) is that lineages meet with probability density $\psi(x)^2$ in deme x and coalesce there at rate $1/K(x)$.

The equilibrium distribution $\psi(x)$, on the other hand, is given by [41]

$$\psi(x) \propto \exp(-2vx/\sigma^2) K(x)^2 \ , \tag{5.14}$$

the pre-factor follows from the normalization condition $\int_x \psi(x) = 1$. Equation (5.13) can be derived from the master-equation of the jump process of lineages backward in time.

Equations (5.13) and (5.14) can easily be combined to calculate the effective population size for any given profile of carrying capacities. The easiest case of constant carrying capacities yields

$$N_e = K\sigma^2/v \;, \tag{5.15}$$

or $N_e = 2Km/v$ in units of our discrete stepping stone model.

5.4. Characteristic scales

By reexpressing the diffusion equation description of the coalescence process in terms of characteristic time and length scales, $\lambda \equiv \sigma^2/(2v)$ and $\tau \equiv \sigma^2/(2v^2)$, one can easily check that the rescaled problem merely depends on one parameter, Kv.

To this end, define the new rescaled function $F \equiv \lambda^2 f$, and new space and time variables, $X_i \equiv x_i/\lambda$, $T \equiv t/\tau$. Upon substituting these new variables into equations (5.1) to (5.4), we obtain

$$\begin{aligned}
\partial_T F &= (\partial^2_{X_1} + \partial^2_{X_2})F - (\partial_{X_1} + \partial_{X_2})F - \; Kv\delta(X_1 - X_2)F && (5.16)\\
F \restriction_{T=0} &= \delta(\chi_1 - X_1)\;\delta(\chi_2 - X_2), && (5.17)\\
0 &= \lim_{X_i \to \infty} F \quad \text{for } i \in \{1,2\}, && (5.18)\\
0 &= (\partial_{X_i} - 1)\, F \restriction_{X_i=0} \quad \text{for } i \in \{1,2\}\;. && (5.19)
\end{aligned}$$

Here, the sampling sites are denoted by $\chi_1 \equiv \xi_1/\lambda$ and $\chi_2 \equiv \xi_2/\lambda$. Notice that the only parameter other than the sampling positions is given by Kv, which obviously controls the behavior of solution.

6. Supplementary information

6.1. Range shifts

Until now, we considered a habitat of infinite size, assuming that the second boundary of the habitat has no significant influence. Especially in the context of climate change, however, an expansion front at one side of the habitat accompanies a retreating front at the other side [81, 82, 8, 48]. Therefore, we have also simulated our model with a second boundary that is moving at the same speed v as the expanding front. As a consequence the total habitat size remains constants, but shifts at a steady speed. This scenario is also called a *range shift* [1].

We compared the mean coalescence times generated by the range expansion model with mean coalescence times from the range shift model. The velocities of the two population fronts were set to the same value, such that the size of the habitat remained constant. As can be seen in Figure 1 in the SI , the results of both models differ only within a small range close to the contraction front. The length of that region is on the order of the characteristic length scale $\lambda = m/v$.

From our results it is clear that the retracting front has a much weaker influence on the coalescence process than the expanding front. This difference can be understood by considering their differential action on the lineages backward in time. While expanding fronts collect lineages, thereby forcing their coalescence, retracting fronts can hardly be reached by the lineages except if they are sampled from very close to the retracting boundary. Only then is it possible for a lineage to collide with the retracting front. The range of influence of the retracting front is given by the characteristic length scale $\lambda = m/v$, which results from random migration competing with the deterministic boundary motion. If the habitat is smaller than this characteristic length scale, both population fronts influence the coalescence process for all sampling positions. Figure 2 in the SI illustrates how the mean coalescence time becomes independent on the sampling position when the habitat size is smaller than $2Km$ demes. In summary, retracting fronts hardly leave any genetic signature in the genetic diversity of the population, quite in contrast to expanding fronts, if the habitat is larger than the characteristic size λ.

6.2. Model Details

6.2.1. Density gradients at the expansion front

In the first part of our study, we assumed that deme sizes are spatially constant. In the second part, we considered the case where deme sizes gradually decay to 0 near the edge of the moving boundary, to model a cline in the suitability of the environment. For these simulations, we chose a logistic density profile at the expansion front that was defined as follows. For a front width W, the linear density profile in the co-moving frame was defined as

$$N_{\text{local, logistic}}(x) = K \begin{cases} 0, & \text{if } x < 0, \\ \frac{1}{1+\exp\left(-\left(\frac{x}{W}-5\right)\right)}, & \text{if } 0 < x < 10W, \\ 1, & \text{if } 10W < x. \end{cases} \tag{6.1}$$

$N_{\text{local, logistic}}$ is defined such that logistic growth is realized in the interval $[0, 10W]$. For distances larger than $10W$, we assumed that the deme size was constant, equal to K.

7. Simulation design

In this section, we describe the simulations that were used to simulate the coalescence process during boundary-limited range expansions.

7.1. The stepping stone model simulations

In the stepping stone model, we define two variables for the lineages, each of which is assigned a positive integer: its sampling distance $x = \xi_i$ to the population front at $x = 0$. The time in generations is initialized as $t_c = 0$.

A generation consists of three parts: the (random) movement of the lineages due to migration, the (deterministic) movement of the lineages in the co-moving frame, and a possible coalescence event. After the boundary movement, the generation counter gets increased: $t_c + = 1$.

7.1.1. The random walk

The migration is modeled by a random walk. Every generation, each lineage moves to the left ($x_i \to x_i - 1$), to the right ($x_i \to x_i + 1$), or stays in place ($x_i \to x_i$) with equal probability $p = 1/3$. If a lineage is at the boundary $x_i = 0$ it stays with probability $p = 2/3$ and moves to the right with $p = 1/3$.

7.1.2. The boundary movement

The simulation is started with an integer parameter ex, the inverse expansion velocity. Consequently, the boundary moves if $t_c \% ex == 0$. As we consider a co-moving frame, a boundary move corresponds to $x_i \to x_i - 1$ ($x_i \to x_i$ if $x_i == 0$).

Random front movement can be introduced by moving the boundary if $q < 1.0/ex$ for a random number q drawn from $[0, 1]_{\mathbb{Q}}$.

7.1.3. Coalescence

If the integers x_1 and x_2 are identical after the boundary movement, coalescence occurs with probability K^{-1}. In that case, the simulation run ends and the current value of t_c is returned.

7.1.4. The contraction front

When we consider a range shift, a second boundary with a second boundary velocity is introduced at a positive integer position. The second boundary acts like the one at $x = 0$.

7.2. The planar, continuous simulations

In the planar model, we define two float vectors x_1 and x_2 as $x_i = [x_i^1, x_i^2]$. We impose periodic boundaries in the second coordinate and a reflecting boundary at $x^1 = 0$. The generation counter is defined as in the stepping stone model.

7.2.1. The Gaussian random walk

For the displacement in the second coordinate, we draw a float from a Gaussian with mean 0 and standard deviation σ_x. The float is added to x_i^2 and the result is shifted according to the reflecting boundaries if necessary.

For the displacement in the first coordinate (along the expansion axis), we draw floats from a Gaussian, again with mean 0 and standard deviation σ_x, until we obtain a float $gauss$ such that $x_i^1 + gauss > 0$. Then, the value of x_i^1 is set to $x_i^1 + gauss$.

7.2.2. Coalescence

If the lineage positions are at a distance smaller than the coalescence distance δ mentioned in the main text, the simulation run ends with probability K_{2D} and the current value of t_c is returned. The distance along the second coordinate is measured with respect to the periodic boundary conditions.

8. Discussion

Most models of range expansions consider the scenario of species colonizing pristine environments. The speed of such invasions is primarily governed by how fast individuals migrate and reproduce in the new environment where space and resources are abundant. By contrast, many range expansions occur due to a gradual environmental change such as the recent global warming or the past glacial cycles [46]. In these scenarios, the speed of the invasion is governed by how fast the environmental change shifts the habitat boundaries, which are set e.g. by retreating glaciers, minimum winter temperatures, or precipitation levels. To study the associated patterns of genetic diversity, we have analyzed genealogies in such boundary-limited range expansions using coalescence simulations that trace lineages backward in time. Two types of population structures were studied, an expanding stepping stone model and a continuous planar model with a stripe-like habitat. Such habitats can be found for instance along coastlines [45, 102] or valleys [83].

We found that, in all cases, the coalescence process follows the caricature that was hypothesized in Fig. 2.1: When lineages are sampled from the bulk of the population and traced backward in time, they coalesce either prior to the arrival of the moving boundary or they are collected by the boundary and forced to coalesce on a short time scale. Lineages have a good chance to avoid coalescence in the first stage, the *free phase of coalescence*, simply by randomly migrating through the habitable area. The rapid coalescence process in the second stage, the *enforced phase of coalescence*, was found to be characterized by an effectively well-mixed population of size N_e if $Kv \leq 1$. The value of $N_e = 2Km/v$ shows that the loss of diversity depends on comparison of the number of migrants per generation, Km, and the expansion velocity v. This result indicates that, the faster the environmental change is that drives the range expansion, the more rapid will be the loss of genetic diversity. However, considering a space-time relationship in range expansion, our result for the effective population size has the following remarkable consequence: If we assume that the habitat increases by length L through a boundary limited range expansion, and we ask how large is the loss in genetic diversity relative to the ancestral diversity (assuming no additional mutations). Then, we would estimate that the loss in heterozygosity in the front population depends on the ratio of the duration $T = L/v$ of the range expansion and the coalescence time in the front population, given by N_e. As a consequence, the diversity depends on the expression $L/(2Km)$, and is most notably *independent* of the speed v of the range expansion. Thus, the *total* loss of diversity through a boundary-limited range expansion, actually does not depend on the speed of the range expansion, but merely on the deme size and the migration rate. The larger the deme size and the migration rate, the smaller the loss of diversity.

To quantify genetic diversity, we studied the distribution of the time to common ancestry, which is proportional to the expected number of pairwise nucleotide differences. We find that the mean coalescence time $\langle T_c \rangle$ sensitively depends on the location of sampling relative to the moving boundary. If both lineages are sampled from a distance less than $2Km$ demes from to the current expansion front, the patterns of diversity resemble that of a well-mixed population of size N_e [41, 91]. In particular a gradient in diversity, one of the hallmarks of a range expansion, is absent for such a sampling scheme. This phenomenon could also obscure the footprint of the shifting of species with narrow habitat zones. To test this hypothesis, we carried out additional simulations of "range shifts" [1], where the habitat does not change in size but slowly shifts in one direction. Such range shifts are for instance driven by the gradual global warming [81, 82, 8, 48]. As detailed in the Supplementary Material, we find that the retracting front affects the coalescence dynamics only up to a distance of $\lambda \equiv m/v$ demes, quite in contrast to the expanding front. Thus, our results from the range expansion model apply directly to range shift except for samples taken close to the retracting front. Importantly, in cases where the habitat has shorter than $L_{\text{plateau}} = 2Km$ demes, the population genetics resembles a population with stationary demography [102], and the patterns of diversity show no signature of a range expansion [81, 82, 8, 48].

In view of this practical result that, for a range expansion to be detectable, the colonized region has to be larger than the length L_{plateau} of the genetically homogeneous front region, one may wonder whether this condition is usually fulfilled or not. A general rule cannot be given because population densities are notoriously difficult to estimate and highly variable between species similar to dispersal rates [55, 4, 14]. Nevertheless, it is revealing to estimate for a range of specific cases the population density that would be needed to blur the range expansion history. The habitat of the black-tailed deer (*Odocoileus hemionus columbianus*), for instance, is situated along the west coast of north America over a length of $L \approx 1500km$ and a width of $k \approx 300km$. Given a mean natal dispersal distance of $\sigma \approx 3km$ [13, 97], a range shift is only detectable if the population density is not larger than $4L/(k\sigma^2) \approx 2.2km^{-2}$. The breeding range of the Piping Plover (*Charadrius melodus*) along the north American east coast has a length of $L \approx 1400km$. The natal dispersal distance of $\sigma \approx 12km$ leads to a linear density threshold of $4L/\sigma^2 \approx 40km^{-1}$ [39, 97] (to be interpreted as individuals per km along the coastline). While alpine pioneer tree species like *Larix decidua* do not experience serial founder effects due to colonies of sexually immature individuals at the expansion front blocking short range dispersal [83], small mammals such as the various *Microtus* species colonizing areas previously occupied by glaciers fit the assumptions of our model. The area described in [83] has a length of $L \approx 4km$ and a width of $k \approx 1km$. Thus, the density threshold for *Microtus arvalis* or *agrestis* ($\sigma \approx 0.03km$ for both [97]) is $4L/(k\sigma^2) \approx 1.8 \cdot 10^4 km^{-2}$. Note that, while these threshold estimates are very rough, they still indicate that population densities do not have to be extremely large for the plateau region to blur the genetic footprint of boundary limited range expansion.

If the colonized region is larger than the plateau length L_{plateau}, it is possible to sample in regions where the genetic diversity increases strongly with distance to the frontier. Within our model, we were able to accurately describe the spatial dependence of the genetic diversity as a function of sampling distance both in a linear stepping stone model and, to a large extent, in a continuous stripe-like habitat. Deviations occurred in the latter case only at intermediate times and only for sufficiently large habitat widths, when the increase in coalescence time due to transverse migration is palpable. We expect that a truly two-dimensional model will be required primarily to accurately describe the coalescence statistics of samples larger than 2, in order to capture the so-called sectoring phenomenon [43]. Binary samples, however, follow quite well our framework, which may therefore be used to reconstruct the speed of a past range expansion from the mean number of pairwise nucleotide differences, $\boldsymbol{\pi}$. For large sampling distances, we found a square-root dependence of mean coalescence time on the sampling distance, which suggests $v \propto m(K/\boldsymbol{\pi})^2$. We expect that this relation also applies to unconstrained range expansion because it only rests on the time scale separation between a free and an enforced phase of coalescence. It is however sensitive to the migration patterns that occur in the bulk of population. For instance, our results differ from the linear dependence observed in “serial founder models” that neglect the back-migration of individuals within the bulk of the population [85, 21]. It might be interesting to explore the effect of density-dependent migration, as has been observed in many species (see e.g. [70]). While such density dependence will clearly affect the loss of diversity in the front population, it will probably not drastically change the coalescence process far from the frontier, which is controlled by migration and coalescence in the bulk and a rapid coalescence upon arrival of the frontier.

After having exposed the specific genetic footprint of a boundary-limited range expansion, one may ask how much data is needed to detect this footprint. The specific practical problem is: How many pairs of lineages need to be sampled to actually detect the difference between a habitat with fixed range and one with moving boundaries. Our description of the coalescent process suggests that this detection will only be possible if lineages that were sampled coalesce in the enforced phase. Only then will the coalescence process be influenced by the boundary that has been moving in the past. The probability that two lineages coalesce in the enforced phase is a central result of our manuscript, stated in Eq. (4.2). Note that for the parameters chosen for the simulations reported in Fig. 4.2, the probability of enforced coalescence is still 0.1 even if lineages are sampled a distance 10^7 away from the boundary. This means that a sample of 100 should be enough to sample enough lineages that are informative about the moving boundary. The long-distance sensitivity to boundary motion of the coalescence process is, ultimately, routed in the recurrence properties of random walks in one dimension. Of course, the lineages that "detect" the boundary will only be informative, if they are able to accumulate a significant number of mutations on their way to the boundary. This simply requires that since the sampling location was colonized, enough time has passed to accumulate many mutations.

Our results above on the coalescent in boundary-limited range expansions contrast with the established picture of unconstrained range expansions that are only dependent on the phenotype of the invading species. For the latter case, one finds that coalescence is always very fast when lineages have arrived at the boundary. The coalescence time typically depends on the logarithm of the deme sizes [41], and thus only grows very slowly as deme sizes are increased. In the boundary-limited case, coalescence in the collection zone can take very long as it increases linearly with deme size. Furthermore, the coalescence process in unconstrained range expansions has been shown to be characterized by multiple mergers, whereby more than two lineages coalesce simultaneously [12]. Again, this is in contrast to boundary-limited range expansions, where a truly well-mixed population at the front arises in the regime $Kv > 1$, in which multiple mergers can be neglected. Based on recent work on traveling waves conditioned on a fixed speed [11], we expect the coalescence process to cross over from the standard Kingman coalescent at small speeds to a multiple merger coalescent at the innate speed. Overall, boundary-limited range expansions thus are able to maintain a higher level diversity throughout the expansion process, and are characterized by a different coalescence process than unconstrained range expansions.

Importantly, however, a genetic footprint similar to an unconstrained range-expansion was reproduced by our simulations when we considered a declining carrying capacity towards the expanding edge. This scenario of a moving density cline was considered to account for range expansion that are driven by a moving gradient in environmental conditions, such as resources or temperature. In this cases, it can be easily imagined that carrying capacities decline gradually towards the edge of the habitat boundary. Then, a "wave-like" population density profile does not arise due to the competition of growth and dispersal, as in a species invasion. Instead, it arises due to a pre-existing cline in environmental conditions.

Our coalescent of such *moving clines* revealed a striking dependence on the length of the cline. If the cline is narrower than the characteristic length scale of $\lambda \equiv m/v$ demes, the coalescent is virtually identical to boundary-limited range expansions with fixed carrying capacities described above. However, when the width of the cline exceeds λ, the patterns appear to be more similar to unconstrained range expansions. The similarity arises ultimately because the width of the population front in unconstrained range expansions is precisely given by λ.

In summary, boundary-limited range expansions retain higher levels of genetic diversity if high carrying capacities can be maintained up to the habitat frontier. Notably we find that the speed of the environmental change (and thus of the range expansion) has no influence on the total loss of diversity. If carrying capacities decline towards the edge of the habitat, the loss of diversity depends on the rapidity of this decline. The loss of diversity will be low if the decline in carrying capacity is abrupt compared to the length scale λ. If the decline is shallow, it will be a small founder population at the tip of the population front that will primarily contribute to the gene pool of the newly colonized territory.

Part II.

The spatial distribution of alleles in expanding populations

Abstract

In the following part of the thesis we pursue two different approaches in the context of expanding populations.

Inspired by observations in microbial experiments, we first ask whether the detailed colonization paths have a major impact on the evolution of an expanding population. If so, under what conditions has this effect to be acknowledged for in population models?

The second approach is much closer to experimental work: we aim at developing tools that can actually be used in experiments to characterize range expansions in two–dimensional habitats based on genetic data.

We will address both approaches on the basis of observables that describe the shape of *mutation patches*, that is spatially grouped individuals with the same mutation at a specific locus. Our results confirm drastic differences between the predictions of models with and without attention to the colonization paths. Furthermore, we develop a flexible method for the analysis of neutral genetic data.

9. Introduction

In part I of this thesis, we introduced the distinction between two types of range expansions. We analyzed the impact of boundary–limited range expansions on the neutral genetic diversity in linear habitats and emphasized the contrast to phenotype–limited range expansions. The latter concept is the standard approach and is used in most preceding publications. We focussed on effectively linear habitats such as coastlines, rives, and valleys and described the distributions of pairwise coalescence times. Therefore, our analysis of linear habitats covers the changes of neutral genetic diversity along, but misses possible patterns perpendicular to the expansion axis.

In this second part of the thesis, we fill this gap by analyzing the consequences of range expansions in two–dimensional habitats. We will focus on the spatial distribution of mutant alleles rather than on coalescence times. To this end, we assume the infinite sites model of mutation [27, 101] according to which each mutation occurs at a previously non polymorphic locus. Consequently, each polymorphic locus bisects the population at the mutant locus into mutant and wild–type alleles. We develop observables that can be measured directly in experimental data sets and take advantage of their increasing abundance.

The spatial distribution of alleles with the same mutation depends on various parameters, notably on genetic drift, spatial heterogeneities and large scale population dynamics such as range changes. However, under the infinite sites model of mutation, the mutant alleles at a given locus all share the ancestor in which the mutation occurred. Therefore, mutant alleles tend to be spatially clustered (*e.g.* [77]).

While this observation holds also for stationary scenarios, the clustering of mutant alleles during range expansions is particularly perspicuous: mutations can fixate locally at the expansion front and form sectors, and *surf* on the expansion wave [40, 42, 43, 62]. In many cases, the two sector boundaries coalesce after a while and leave a well delimited patch in the colony (Figure 9.1).

In the microbial experiments, migration is heavily limited and reproduction halts behind the expansion front. Therefore, the sectors remain genetically homogeneous and well delimited. Note that, in scenarios with continuing reproduction and migration behind the front, the sectors are expected to dissolve due to diffusion.

Based on these observations, we will refer to a group of individuals with the same mutation as *mutant patche*. For the purpose of clear distinction between different scenarios, we will further differentiate between *surfer patches* (patches that profited from a surfing event) and *non–surfer patches* (patches that emerged without surfing) whenever needed.

In the following, we will present two approaches based on simulation data: first, we will show that the spatial distribution of mutant alleles allows the detection of ongoing and possibly past range expansions. Second, we analyze how the detailed colonization history of a expanding population impacts the genetic diversity and under what conditions the colonization paths must be acknowledged for in the coalescent. Interestingly, the colonization does not only impact the coalescent at individual loci, it can also provoke severe linkage disequilibrium between genetically unlinked loci.

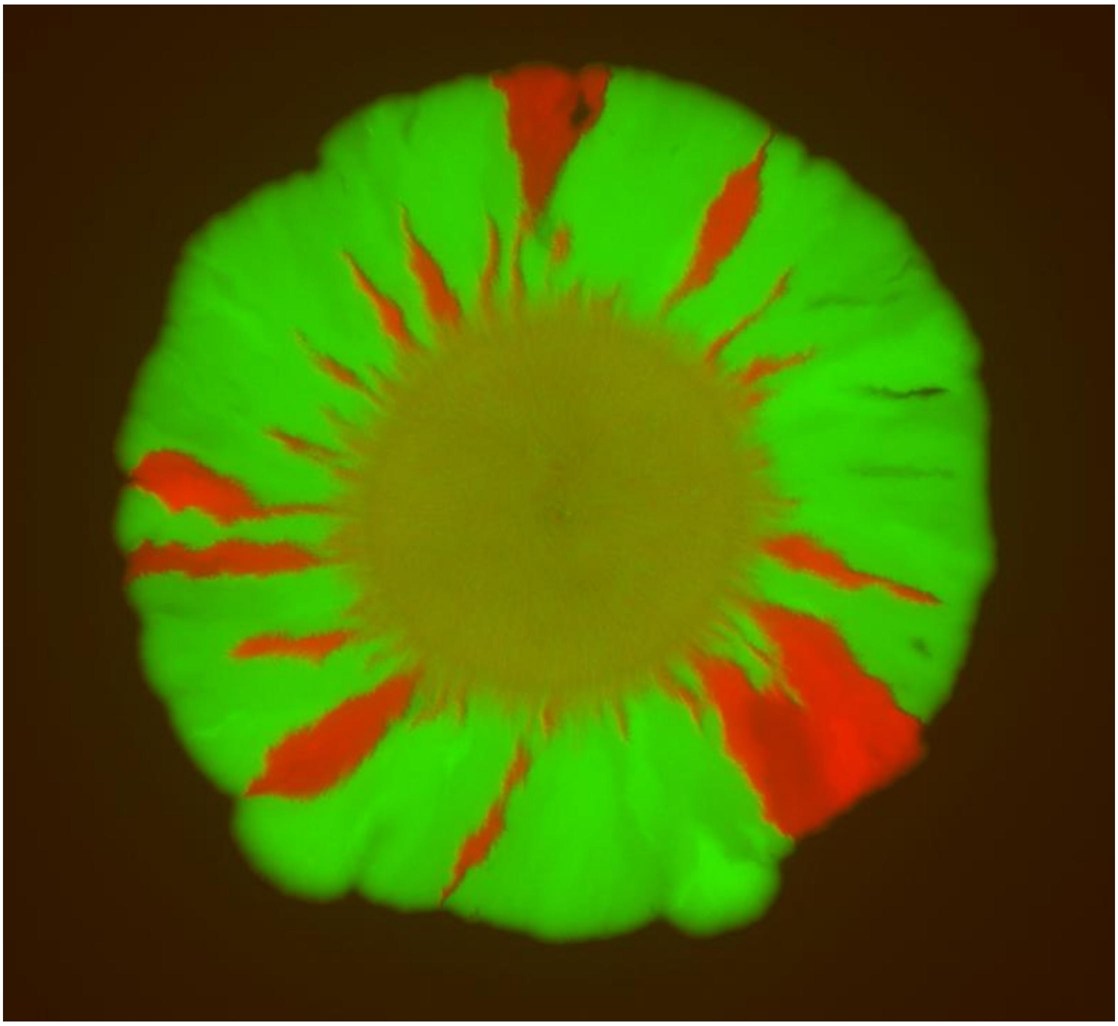


FIGURE 9.1.: **Mutation patches in a agar plated colony of *E. coli.*** In competition experiments between neutral strains of *E. coli* labeled with two different fluorescent protein markers, sectors emerge from the initial colony in the middle. The sectors correspond to surfing events. When a sector loses the contact to the expansion front, a patch like structure that we call *mutant patch* remains. The microscopy image was made by Fabian Stiewe in the lab of Oskar Hallatschek. Equivalent images can be found for instance in [40, 42, 43, 62].

9.1. The spatial distribution of alleles in stationary habitats

In spatially structured populations, an individual's genome can reveal its geographic origin [86, 77]. On the other hand, the spatial distribution of alleles can be used to infer the geographic origin of a mutation [89, 20].

When a mutation appears in a population it is subject to genetic drift: it may go extinct or increase in frequency as explained in the general introduction. The spatial dynamics of a mutation, that is the distribution of individuals with that mutation over time, can be modeled by a branching random walk. Individuals with the mutation reproduce and their offspring is placed in proximity of the parent's place of birth.

Taking a sample from a population implies a retrospective and selective view of the process: only successful mutations are present in the data and (almost always) only a subset of the mutants is sampled. The spatial distribution of the mutants in the sample is, thus, an approximation of the mutant distribution in the population. We discuss the accuracy of the approximation in the context of sparse sampling.

The shape of random objects like the mutant distribution in spatial habitats has been addressed by many authors in the context of random walks and different applications. The shape of the path of a single random walker, for instance, is characterized in [88]. J. Rudnick uses the radii of gyration to point out that the typical shape of a random walker's path is not round but rather elongated. With the standard deviation of the mutant positions along the habitat axes, we use a similar method to describe the shape of the mutant patches.

In [88] only single random walkers are addressed. Groups of correlated or uncorrelated random walkers are discussed in [67] using the convex hull of the walker positions. Using a similar approach, E. Dumonteil and colleagues present an analysis of the outbreak of epidemics [22]. As the propagation of a mutation in our model follows almost the same mechanism as described in [22] and shares many features with the models in [88] and [67], we expect elongated mutation patches in stationary habitats.

Note that the maximal elongations of the mutation patches are not aligned along any fixed axis in scenarios described in [88, 67, 22].

9.2. The spatial distribution of alleles in expanding habitats

In expanding populations, (even slightly deleterious) mutations can 'surf on the wave of advance' [60, 98, 28] and form a sector of characteristic shape. These surfing events can drastically affect the shape (and size) of mutant patches. A surfing event will typically stretch the involved mutation patches along the expansion axis — note that it is the *same* axis for all surfer patches.

The impact of colonization paths Spatial models are often constructed such that the migration is isotropic to ensure mathematical tractability (for instance in circular or toroidal stepping stone models, *e.g.* [94, 90], infinite habitat models, *e.g.* [56], continuous, finite, linear models [102]). Most coalescent models are based on the averaged migration rates of a forward model, *e.g.* [28], the lineage movement can then be modeled as a random walk.

The sectoring pattern observed in experiments with microbes and simulations [40, 42, 43, 62] indicates that such assumptions are not always legitimate. These microbial colonies grow only within a thin layer close to the expansion front. Individuals behind the front do not longer benefit from suitable growing conditions, stop reproducing, and remain as a frozen record of the initial colonization process. A coalescent for such a scenario must therefore virtually mirror the colonization paths: the lineage of a sampled individual from inside the colony 'waits' for the expansion front before it moves backwards along the path of colonization.

If the statistical properties of the inverse colonization paths differ from the standard assumptions of lineage movement, the colonization paths must be acknowledged for in the coalescent.

Linkage and hitchhiking Genetic *linkage* between loci describes the tendency of loci to be inherited together due to the reproduction mechanism. Consider a diploid species (such as humans) and two loci $\mathcal{A}$ and $\mathcal{B}$ located close to each other on the same autosome. Let both loci be segregating sites with alleles A_1, A_2 and B_1, B_2, respectively. If no recombination occurs between the two loci, they will be copied together and found in the same configuration in offspring individuals that inherit the autosome.

The classical observable for linkage between loci, denoted D, compares the frequency of alleles (here $p(A_1)$ and $p(B_1)$) in a population with the frequency of finding the alleles together ($p\,(A_1 \text{ and } B_1)$):

$$D := p\,(A_1)\,p\,(B_1) - p\,(A_1 \text{ and } B_1) \tag{9.1}$$

If the loci are inherited independently and the population is in equilibrium, $D \approx 0$ is expected. Deviations from $D \approx 0$ are called *linkage disequilibrium* and indicate that at least one of the above assumptions is violated.

The classical example for linkage disequilibrium is called *hitchhiking*. Consider a newly arising beneficial mutation that overcomes genetic drift and fixates in the population. This scenario is called hard selective sweep. Neutral and even slightly deleterious mutations at loci genetically linked to the beneficial mutation can 'hitchhike' to high frequencies just because the beneficial mutation initially occurred on their genetic background. After the fixation, the beneficial mutation and the hitchhikers occur together $p\,(A_1 \text{ and } B_1)$ more often than expected under the assumption of independence from their frequencies $p\,(A_1)$ and $p\,(B_1)$ in the population. Hitchhiking in spatially structured populations is of particular importance in the context of range expansions [5].

Note that linkage disequilibrium can also occur due to demography [96], spatial structure, and range expansions. See *e.g.* [99] for a detailed discussion.

Linkage through colonization history As mentioned in the general introduction, the coalescent is a stochastic model that accounts for our lack of knowledge of the true pedigree of a sample. Considered as random variables, the coalescents of two genetically unlinked, neutral loci from the same sample are assumed to be independent and identically distributed (i.i.d.).

The microbial colonies, however, show a completely different behavior. The lineages at all loci move backwards along the branches of the same colonization tree. Due to optional sexual reproduction in yeast and horizontal gene transfer in *E. coli*, lineages of different loci may jump occasionally to neighboring branches of this tree, but these branches are likely to re–coalesce soon. In short, all loci in the microbial experiments are forced onto almost the same coalescent — independent of their genetic linkage. Heavy linkage disequilibrium is to be expected.

The colonization represents a form of *quenched randomness* for the coalescent: if the colonization paths influence the lineage movement, the coalescents at genetically unlinked loci are *not independent*. In most range expansions though, the bond between the colonization paths and the coalescent will be much weaker than in the agar–plated microbe colonies. Nevertheless, even a tendency of the lineages to follow the colonization paths could increase linkage disequilibrium.

10. Model and methods

We model the colonization process within scenarios of boundary–limited and phenotype–limited range expansions using the Eden growth model [24]. Based on these colonies, we employ three versions of a spatial coalescence process to access the spatial distribution of alleles.

The Eden growth model covers only the primary colonization as a branching process. (In fact, we do not aim to model the complete dynamics of a population in the forward in time model.) Recolonization events are incorporated in the coalescent.

10.1. Colony growth

The model habitat is a Cartesian lattice of width W and length L with periodic boundaries at $y = 0$ and $y = W$. Absorbing boundaries are imposed at $x = 0$ and $x_{\mathrm{max}}(t)$, the latter representing the habitat boundary moving at velocity v. Each lattice site has carrying capacity 1. Time is measured in generations.

The colony growth starts at time $t = 0$ from a initial population of W individuals positioned along $x = 0$, the boundary position is given by

$$x_{\mathrm{max}}(t) = 1 + vt. \tag{10.1}$$

The expansion front has a maximal possible velocity, the *phenotype–limited expansion velocity* v_{pheno} introduced in part I. In our version of the Eden growth model, the boundary limitation holds for $v_{\mathrm{boundary}} < v_{\mathrm{pheno}}$. For $v_{\mathrm{boundary}} > v_{\mathrm{pheno}}$, there is effectively no boundary towards $x = \infty$ and the colonization occurs under the scenario of a phenotype–limited range expansion.

Each lattice site (x, y) (called *deme* in the following) is assigned to one of four categories: active, passive, open (to colonization), or blocked. The active demes are colonized and at least one of the four neighboring demes is not colonized. The passive demes are colonized and all their neighbors are also colonized. The open demes are not occupied and within the current habitat boundaries ($0 \leq x < x_{\mathrm{max}}(t)$). The blocked demes are beyond the moving boundary ($x > x_{\mathrm{max}}(t)$).

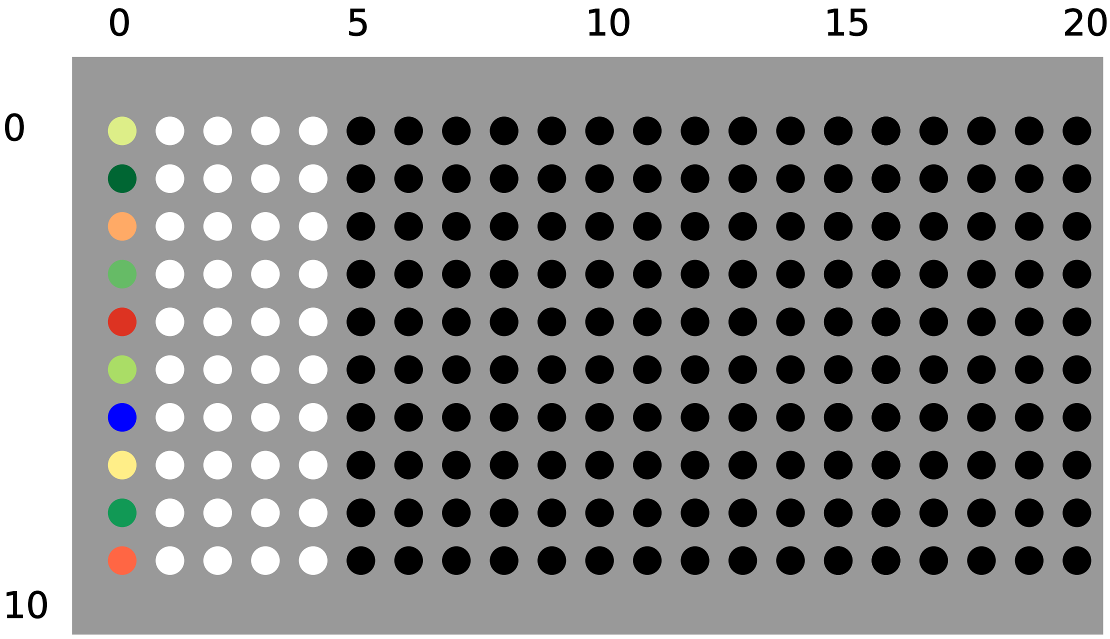


FIGURE 10.1.: **Start configuration of the colony.** The colored circles on the left ($x = 0$) represent the initial colony. The white circles are demes open to colonization. The black circles are blocked (by the boundary at $x = 5$ for $t = 0$. On the vertical axis, the habitat ranges from $y = 0$ to $y = 9$. Via periodic boundary conditions $y = 10$ is identified with $y = 0$.

The reproduction attempts of the demes are modeled independent and identically distributed. Passive demes cannot reproduce successfully, thus, only the reproduction attempts of the active demes are considered in the simulation. We assume that the reproduction is a Poisson process with rate 1. Then, the reproduction of all active demes is a Poisson process with rate equal to the current number of active demes. The mean time to the next reproduction attempt of an active deme is $1/\#$(active demes).

Each simulation step encompasses the following parts. The parent is drawn from the pool of active demes at random. The model time increases by $1/\#$(active demes). One of the four neighboring demes is drawn as a offspring candidate. If the candidate deme is open to colonization it is colonized by a copy of the parent, the status of the new deme and its neighbors is updated. If the candidate deme is occupied (*i.e.* active or passive) or blocked, the reproduction attempt fails. The position $x_{\max}(t)$ of the front and the status of the demes along the expansion front are updated.

For each deme (x, y), we record the *colonization time* $t_{x,y}$ and the parent deme.

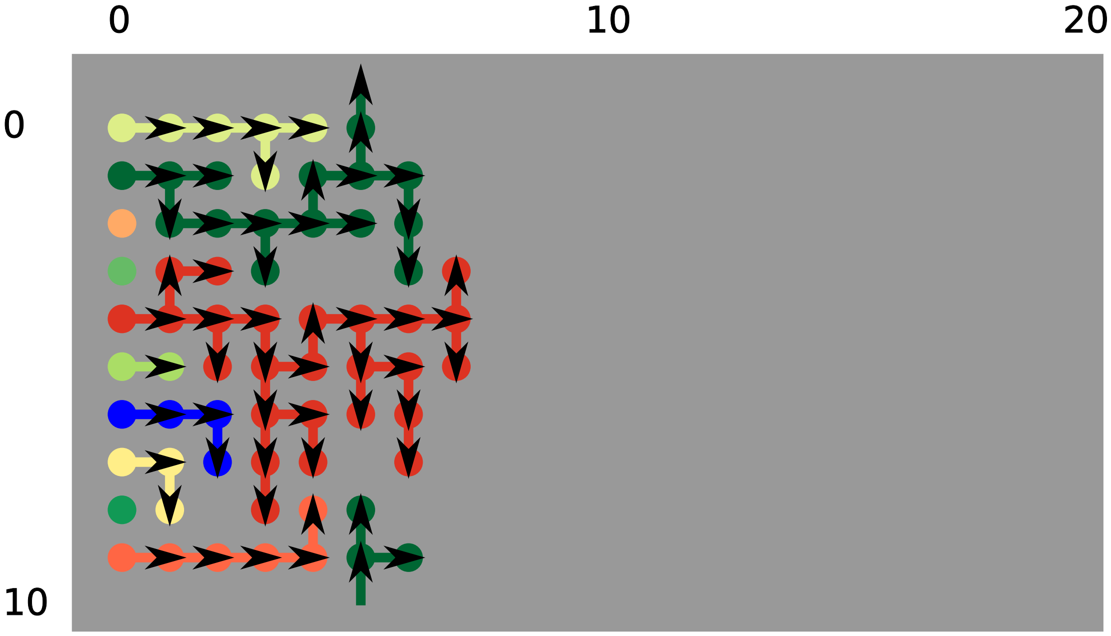


FIGURE 10.2.: **Colony after 10 generations.** In a simulation time step, an active deme is drawn at random and attempts to reproduce. One of the four directions is drawn, if the corresponding deme is open the reproduction is successful. Note that each deme from the initial colony gives rise to a tree. All but the genotypes from $(0, 1)$ and $(0, 4)$ have lost contact to the expansion front.

The colony growth stops when the first deme at $x = L$ is colonized.

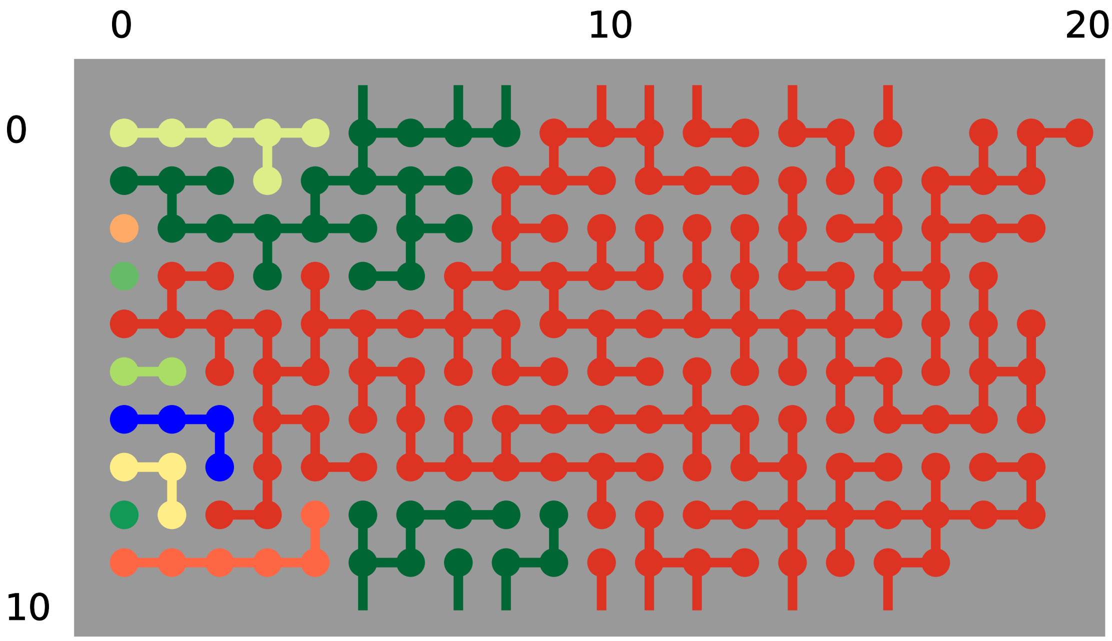


FIGURE 10.3.: **End of the colonization.** The simulation stops when the first deme at $x = x_{\max}$ (here at $(20, 0)$) is colonized. Note that in this example all individuals at the expansion front trace back to a single individual at $(5, 4)$. The genotype from $(0, 4)$ successfully surfed on the wave of advance and fixated at the front.

10.2. The coalescence process

As noted earlier, the coalescent accounts for our inevitable lack of information about the actual pedigree of a population. The randomness of the population dynamics forward in time is transferred to the randomness of the lineage movement backwards in time. The modeling of the lineage movement, however, is not obvious but depends on the details of the population dynamics.

In the following sections, we present coalescent models that cover different scenarios of the forward dynamics.

10.2.1. Lineage movement strictly along the colonization paths: the fully quenched coalescent

Bacterial colonies on agar plates grow almost exclusively within a thin active layer at the expansion front with sufficient access to nutrients. The inside of the colony is deprived of nutrients and keeps a 'frozen record' of the colonization process [40].

Scenarios with such a drastic advantage for the initial colonizers can also occur due to high density blocking [100].

Note that each site is only colonized once and from a single parent site. Therefore, each lineage in such a scenario waits in place until this colonization event. Then, it must move to the site it was colonized from.

We model the coalescent in such situations by forcing the sampled lineages to move strictly along the colonization path and only at colonization time. For a given colony, each sample has a unique coalescent completely determined by the colonization paths (Figure 10.3). We will refer to this version of the coalescent as the *fully quenched coalescent*.

In the simulation we draw a sample of size n either at random from the entire habitat or along the front line and set the current time to the end of the colonization process. The simulation steps consist of the following parts. Move each lineage along the colonization path while its current time is smaller than the colonization time of the deme the lineage is located in ($t < t_{x,y}$). Coalesce lineages that are in the same deme. Decrease the model time by one. Figure 10.4 shows a fully quenched coalescent on the background of a phenotype–limited range expansion.

The coalescence process ends when either only one lineage remains or when the model time is 0. In the latter case, the coalescence may not be complete.

10.2.2. Diffusive lineage movement behind the expansion front: the front–quenched coalescent

When migration and reproduction continue in the whole habitat, demes remote from the expansion front are likely to be recolonized repeatedly and from all directions. In such situations, we assume that lineage movement behind the front can be modeled as independent from the initial colonization paths. Close to the front in contrast, the colonization paths determine the lineage movement. In the following, we refer to this version of the coalescent as the *front–quenched coalescent*.

Just as in the previous version, we draw a sample of size n and set the current time to the end of the colonization process. The simulation steps consist of the following parts. Move each lineages along the colonization path while its current time is smaller than the colonization time of the deme the lineage is located in ($t < t_{x,y}$). If a lineage was not moved in the current generation, move it to one of the neighboring colonized demes. Coalesce lineages that are in the same deme. Decrease the model time by one. Figure 10.5 shows a coalescent with diffusive lineage movement behind the expansion front on the background of a phenotype–limited range expansion.

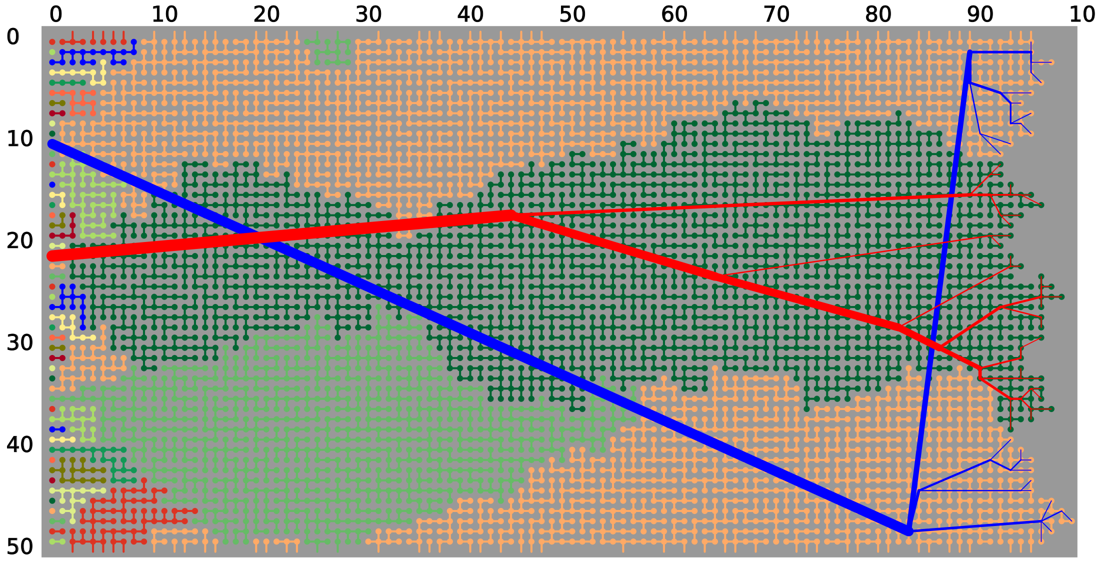


FIGURE 10.4.: **The fully quenched coalescent.** A sample of size $n = 50$ is drawn along the expansion front. The nodes of the tree denote the coalescence events, the width of the lines encodes the number of samples that have coalesced into that lineage. Note that all lineages stay in the sector they were sampled in and that they do not cross each other. Coalescences of more than two lineages in a single event are possible.

In contrast to the fully quenched coalescent, the colonization process does not completely determine the front–quenched coalescent. Realizations of the process result in different trees slightly linked by the common colonization pattern.

10.2.3. Diffusive lineage movement within the whole habitat: the unquenched coalescent

The impact of the quenched randomness due to the colonizations paths on the coalescent is unknown. Therefore, as a null model within the framework of range expansions, we introduce a third version of the coalescent for our model.

For this version, we assume that only the position of the habitat boundaries impact the lineage movements. Each lineage moves to one of the neighboring demes inside the current boundaries once per generation. Lineages in the same deme coalesce. The coalescence process ends when only one lineage remains. In the following, we refer to this version of the coalescent as the *unquenched coalescent*.

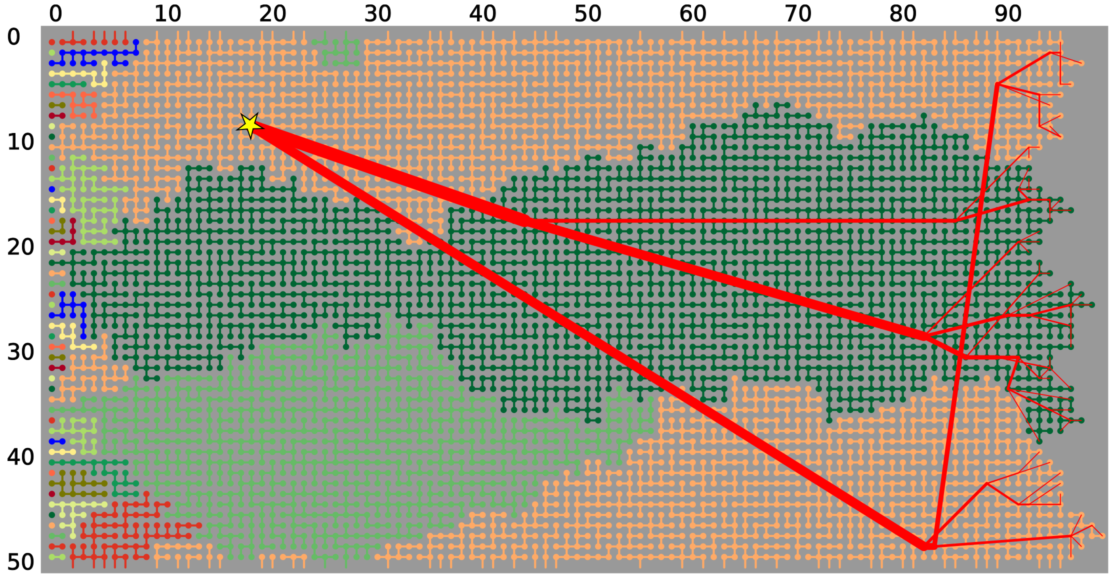


FIGURE 10.5.: **A realization of the coalescent with diffusive lineage movement behind the expansion front.** A sample of size $n = 50$ is drawn along the expansion front. The nodes of the tree denote the coalescence events, the width of the lines encodes the number of samples that have coalesced into that lineage. The yellow star denotes the most recent common ancestor of the sample. Note that the lineages can cross each other and the sector boundaries. Coalescences of more than two lineages in a single event are possible.

10.3. Observables

As announced in section 9, we will describe and analyze the spatial distribution of mutant alleles. We focus on the shape of the mutant patches rather than to the patch locations in the habitat. The *patch shape* will generally be influenced by a number of factors such as the length and width of the habitat, the expansion velocity, spatial heterogeneities, and the roughness of the front. The observables introduced in this section are designed to disentangle these factors and to allow the identification and analysis of range expansions based on genetic data.

Our model habitat does not include spatial heterogeneities but it has, just as the natural habitats considered in experiments, finite ranges. Before we can analyze the impact of the range expansion and specifically the colonization paths, we filter the influence of the habitat size.

The properties of the mutation patches we describe are: the *patch size*, *i.e.*, the number of individuals in the patch, the *patch position*, *i.e.*, the mean over the positions of the individuals in the patch, and the patch dimensions (roughly: length and width).

Note that in a real sample not every possible patch occurs. As explained earlier (see Figure 1.1), each edge of the coalescent tree corresponds to a possible mutant wild–type bipartition of the sample. The mutant subset of that bipartition is the patch. The probability of having a patch in the sample is the probability of at least one mutation along the corresponding edge of the coalescent tree and, thus, proportional to the length of the edge in generations.

For the analysis of the coalescent models we will consider *all possible patches*. The occurrence of the patches in real samples is addressed in the discussion of this part.

10.3.1. The critical patch size

When we want to analyze patch shapes independent of the habitat size and shape, we must restrict our analysis on patches that were not influenced by the non-moving habitat boundaries. At the same time, the filter must leave the statistical properties of the remaining patches unchanged.

It is, for instance, not correct to only exclude patches that ‘touch’ one of the boundaries. By doing this, we would introduce a bias towards the habitat shape.

In our model setup, the finite habitat will prove to be the critical dimension. The boundary at $x = 0$ has no significant impact as we will see on the basis of simulation data. The impact of the moving boundary will, of course, not be filtered as it is the expansion process we are interested in.

In order to solve the problem of the finite habitat width, we define the *critical patch size* B_{critical}. For each patch size, we determine the fraction of patches that cover the full habitat width or get close to this. B_{critical} will serve as a safety margin to patches that are skewed by the habitat dimensions: patches of size $B > B_{\text{critical}}$ are excluded from the analysis.

The patches that remain in the analysis are unaffected by finite habitat width as they did not grow large enough to encounter a non-moving habitat boundary. Depending on the choice of B_{critical}, a few patches might actually violate this last statement. This issue will be discussed based on the simulation data.

Note that we do not claim to describe the general statistical properties of mutation patches. Instead, we describe the properties of the patches with $B < B_{\text{critical}}$ and determine whether they contain enough information to analyze the range expansion.

The actual choice of B_{critical} will depend on the details of the habitat and the threshold chosen in the particular experiment. In experiments the periodic boundaries must certainly be replaced by reflecting ones and a filter must account for the patch position perpendicular to the expansion axis. The critical patch size in the context of our simulation is presented and determined in section 11.2.1.1.

The effort we made here to compensate for the finite habitat width is not motivated by a lack of computing power and the resultant need to limit the habitat dimensions. Much rather, we aim at developing tools for the analysis of real data sets: natural populations populate finite ranges.

10.3.2. Bubble shapes in expanding populations

The patch size filter will prove to be effective and we will now proceed and discuss the patch shapes in expanding populations.

When a mutation surfs on the expansion wave, we expect the formation of a sector [40, 42, 43, 62]. (Figure 10.4 and 10.5 show the sectoring effect in the Eden colonization model.) As long as the mutation is present at the expansion front, the elongation of the sector grows linearly in time. The movement of sector boundaries is a stochastic process and can often be modeled as a simple random walk [40]. As the expected distance of a simple random walker grows with the square root of time, the ratio of width and length of such surfer patches should form a power–law with exponent 0.5. If the sector boundary movement is sub–diffusive or super–diffusive the power–law exponent is expected to be smaller or larger than 0.5, respectively.

Mutations that are not influenced by the range expansion can nevertheless reach large frequencies due to genetic drift. In contrast to the surfer patches, the average length to width ratio for these patches is expected to be approximately 1. These patches can be used to account for non–isotropic migration (see section 10.3.2.3).

Length and width can be defined in different ways. Here, we follow [93] and [88] who consider the *principal radii of gyration* of the trails of random walks and flights, respectively. For each patch, we calculate the standard deviations of the individual mutant positions in the patch along the expansion axis and perpendicular to it. If the expansion direction is the first principal component of the patch, this approach is equivalent to the one of [93] and [88].

Here, we are interested in the average radii of gyration along a potential expansion axis of many patches rather than in the gyration radii of individual patches. As we will see in section 10.3.3 (methods) and section 11.2.1.2 (results), the expansion direction is in fact close the first principal component of the ensemble of the mutation patches. Consequently, we measure the radii of gyration along the expansion axis and perpendicular to it.

For a patch of size B and the expansion direction x we define

$$\mathrm{SD}_x = \sqrt{\frac{1}{B}\sum_{i=1}^{B} (\overline{x} - x)^2}. \tag{10.2}$$

SD_y is defined accordingly. These standard deviations are equivalent to the radii of gyration along the coordinate axes.

We call

$$c := \frac{\log \mathrm{SD}_y}{\log \mathrm{SD}_x}, \tag{10.3}$$

the *(logarithmic) compression factor* of the mutation patch.

10.3.2.1. Extreme SD–values

As all habitats have finite ranges, the patches of successful mutations may cover the full habitat width or length. Clearly, in such a situation, the compression factor is not a genuine result of the range expansion but of the dimension of the habitat. To be able to estimate unbiased compression factors, we briefly discuss extreme distributions of individuals in a patch in the context of our model habitat. Recall that we consider a expanding habitat of maximal length L and width W with reflecting boundary conditions at $x = 0$ and $x_{\max}(t)$ and periodic boundary conditions at $y = 0 \equiv W + 1$.

If all individuals in a patch share the same x–coordinate, we have $\mathrm{SD}_x = 0$ and the compression factor is not defined. The corresponding result for the y–coordinate gives $\mathrm{SD}_y = 0$, of course. These are pathologic cases that occur for patches of size 1 and (in the simulation with its regular lattice) occasionally for slightly larger patches. Such small patches do not confer much information about patch shapes. We exclude patches of size smaller than $B_{\min} = 5$ from the analysis.

The theoretical maximum SD–value along an axis occurs, for instance, in a patch of size 2 with individuals at maximal distance within the habitat. For SD_x, that would be $x_1 = 0$ and $x_2 = L$. Thus,

$$\mathrm{SD}_{x,\max} = \sqrt{\left(\frac{L}{2} - 0\right)^2 + \left(\frac{L}{2} - L\right)^2} = \frac{L}{\sqrt{2}}. \tag{10.4}$$

For SD_y, we could have $x_1 = 0$ and $x_2 = W/2$, for instance. A moment's reflection suffices to see that for these values, both, $\overline{y} = W/4$ and $\overline{y} = 3W/4$ are possible mean values. For both choices, we obtain

$$\begin{aligned} \mathrm{SD}_{y,\mathrm{max}} &= \sqrt{\left(\frac{W}{4}-0\right)^2+\left(\frac{W}{4}-\frac{W}{2}\right)^2} \\ &= \sqrt{\left(\frac{3W}{4}-W\right)^2+\left(\frac{3W}{4}-\frac{W}{2}\right)^2} \\ &= \sqrt{\left(\frac{W}{4}\right)^2+\left(\frac{W}{4}\right)^2}=\frac{W}{2\sqrt{2}}. \end{aligned} \tag{10.5}$$

For the second line, note that $0 \equiv W (\mathrm{mod}\ W)$.

Nevertheless, such values are unlikely as the mutants in a patch tend to be grouped. It is therefore more helpful to calculate the SD–values for a patch that covers all lattice sites in the habitat. Each row in the habitat gives the same result. For a single row and $\overline{x} = L/2$, we obtain

$$\mathrm{SD}_{x,\mathrm{full}} = \sqrt{\frac{1}{L}\sum_{x=0}^{L}\left(\frac{L}{2}-x\right)^2}. \tag{10.6}$$

The sum can be easily calculated:

$$\begin{aligned} \sum_{x=0}^{L}\left(\frac{L}{2}-x\right)^2 &= \sum_{x=0}^{L}\left(\frac{L^2}{4}-Lx+x^2\right) \\ &= \frac{1}{12}L^3-\frac{1}{6}L. \end{aligned} \tag{10.7}$$

For $L \gg 1$ we approximate to

$$\mathrm{SD}_{x,\mathrm{critical}} := \frac{1}{\sqrt{12}}L. \tag{10.8}$$

For the full habitat patch, along the periodic y–axis all choices of $\overline{y}$ are equivalent. But as we have the boundary at $y = 0 \equiv W$, $\overline{y} = W/2$ is the correct choice (it minimizes SD_y). Consequently, we obtain

$$\mathrm{SD}_{y,\mathrm{critical}} := \frac{1}{\sqrt{12}}W \tag{10.9}$$

as above.

10.3.2.2. Expected SD–values

Assume that a mutation surfs on the expansion front and drops from the front before the expansion ends and let l and w be the length and width of the corresponding patch. For such a patch, the SD–values can not be calculated as above.

When the mutation occurs, the patch 'opens' and starts growing in length and width. At some point, the patch will have reached its maximal width and begin to collapse until, finally, the last mutant individual has lost contact to the front.

The standard deviations SD_x, SD_y in both directions can be written in a continuum approximation as

$$\begin{aligned} SD_x &= \frac{\int_B dA\, \left(x - l/2\right)^2}{\int_B dA}\ , && (10.10) \\ SD_y &= \frac{\int_B dA\, y^2}{\int_B dA} && (10.11) \end{aligned}$$

with $\int_B dA$ the integral over the area of the patch B.

If the boundary is sufficiently slow, the width of the population front is mostly 1. The sector boundaries in our model will then perform a simple random walk along the y–axis with probabilities $p_{\text{up}} = 0.25$ and $p_{\text{down}} = 0.25$. The diffusion constant (in units of the lattice constant and generations) of the random walk depends on the boundary velocity.

For a simple approximation, we assume that the patch is symmetric with respect to both habitat axes. Then, it is sufficient to calculate the standard deviation of the upper left quarter of the patch from the center. Integrating yields the area $A^{(1/4)}$ of a quarter of the patch, $B^{(1/4)}$. In terms of x and y we can use

$$\begin{aligned} \int_{B^{(1/4)}} dA &= \int_0^{l/2} dx \int_0^{\sqrt{2Dx}} dy && (10.12) \\ &= \int_0^{w/2} dy \int_0^{l/2 - y^2/(2D)} dx\ . \end{aligned}$$

For the area $A^{(1/4)}$ we obtain

$$\begin{aligned} A^{(1/4)} = \int\limits_{B^{(1/4)}} dA &= \int\limits_0^{l/2} dx \int\limits_0^{\sqrt{2Dx}} dy = \int\limits_0^{l/2} dx \, \sqrt{2Dx} = \frac{1}{3}\sqrt{D} l^{3/2} \\ &= \int\limits_0^{w/2} dy \int\limits_0^{l/2-y^2/(2D)} dx = \int\limits_0^{w/2} dy \, \big(l/2 - y^2/(2D)\big) \\ &= \frac{lw}{4} - \frac{1}{48D} w^3, \end{aligned} \tag{10.13}$$

and the SD–values are

$$\begin{aligned} SD_x &= \frac{1}{A^{(1/4)}} \int\limits_{B^{(1/4)}} dA \, \big(x - l/2\big)^2 \\ &= \frac{1}{A^{(1/4)}} \int\limits_0^{l/2} dx \int\limits_0^{\sqrt{2Dx}} dy \, \big(x - l/2\big)^2 \\ &= \frac{1}{A^{(1/4)}} \left(\frac{2D^{1/2} l^{7/2}}{105} \right) = \frac{2}{35} l^2 \end{aligned} \tag{10.14}$$

and

$$\begin{aligned} SD_y &= \frac{1}{A^{(1/4)}} \int\limits_{B^{(1/4)}} dA \, y^2 = \frac{1}{A^{(1/4)}} \int\limits_0^{w/2} dy \int\limits_0^{l/2-y^2/(2D)} dx \, y^2 \\ &= \frac{1}{A^{(1/4)}} \left(\frac{D^{3/2} l^{5/2}}{15} \right) = \frac{1}{5} D l \end{aligned} \tag{10.15}$$

The expected logarithmic compression factor under the above assumptions is, thus,

$$\begin{aligned} c(l) &= \frac{\log SD_y}{\log SD_x} = \frac{\log(\frac{1}{5} D l)}{\log(\frac{2}{35} l^2)} \\ &= \frac{1}{2} \left(\frac{\log l + \log \frac{D}{5}}{\log l + \log \sqrt{\frac{2}{35}}} \right) \\ &\overset{l \to \infty}{\longrightarrow} \frac{1}{2} \end{aligned} \tag{10.16}$$

For large patches ($l >> 1$) we recover the expected logarithmic scale factor of 1/2.

We will not pursue this approach in more detail. However, note that for $D < 1$ (this is the case for all expansion velocities) we have $D/5 < \sqrt{2/35}$. Thus, for $\log l + \log D/5 > 0$, $c(l)$ in equation (10.16) grows with the patch length l.

Note that in our two versions of the coalescent with diffusive lineage movement, the patches continue to change their shape after the front has passed.

10.3.2.3. Isotropy of migration

If migration is isotropic, the expected compression factor for patches that are neither affected by the range expansion nor by the habitat boundaries is 1. Thus, a potential anisotropy can be identified and corrected for using the compression factor of such patches. Note that the expected shape of a single unbiased patch is not round but rather oval (*cf.* [88, 67, 22]).

If a mutation surfs on the expansion wave, we expect the formation of a sector [40, 42, 43]. As long as the mutation surfs, the elongation of the sector grows linearly in time, while the sector boundaries perform random walks perpendicular to the expansion axis. Mutation patches corresponding to a surfing event are therefore expected to deviate from $\langle c \rangle = 1$. It depends on the model details whether the boundary movement is sub–diffusive ($c < 0.5$), diffusive ($c = 0.5$), or super–diffusive ($c > 0.5$).

10.3.3. Estimation of the expansion direction

For the analysis of the radii of gyration, we assume that the expansion direction is known. Therefore, we implemented a simple method to estimate the expansion direction. The method is based on the observation that, on average, surfer patches extend more along the expansion axis than perpendicular to it.

We assume that the sample coordinates of an experiment are given in an arbitrary but common Cartesian coordinate system. For each patch b, we calculate the center $(\overline{x}_b, \overline{y}_b)$ and define the relative mutant positions as

$$(x_b, y_b) := (x - \overline{x}_b, y - \overline{y}_b). \tag{10.17}$$

All relative mutant positions are pooled into a single data set. Note that, each mutant is represented several times, once for each patch it belongs to. As we will see, the largest part of the variance in the pooled data set is typically found along the expansion axis. Then, the expansion direction can be estimated by calculating the first principal component of the pooled centered patch data.

11. Results

In this section, we briefly describe the colonization process and the transition between boundary–limited and phenotype–limited range expansions. We will then present the results of the coalescence simulations for different expansion velocities and recolonization scenarios.

11.1. Colonization

The the expansion velocity cannot exceed the boundary velocity, by construction. Moreover, we expect that the expansion velocity reaches a finite, maximal value if the expansion is not boundary–limited. Figure 11.1 shows the expansion velocity as a function of the boundary velocity for colonizations of the (500×10000)–lattice.

Accordingly, we define the *phenotypical expansion velocity* v_{pheno}. Its exact value will depend on the details and parameters of the colonization process but for our objective it is sufficient to assert that the expansion velocity saturates and to measure the value $v_{\text{pheno}} \approx 0.618$.

For the unquenched coalescent, we assume that all demes inside the habitat boundaries are populated. Consequently, the simulations for unquenched coalescent must be restricted to velocities $v < v_{\text{pheno}}$. For the remaining two versions of the coalescent, it is sufficient to run simulations for $v < v_{\text{pheno}}$ and, additionally, for a single value significantly larger than v_{pheno}, in order to cover the phenotype–limited case.

As expected from previous results [40, 42, 43] and already seen in Figure 10.5, we observe the formation of sectors. The sectoring depends on the expansion scenario: for boundary velocity below the phenotypic expansion velocity, we see a boundary–limited range expansion and the population front mirrors the boundary shape.

Different shapes of the moving boundary can have a drastic impact on the sectoring and thereby on genetic diversity (see Figures B.1 and B.2). Variation in the boundary shape introduces a new and complex parameter to the model. We decided to subordinate this topic to the analysis of the more basic model with flat

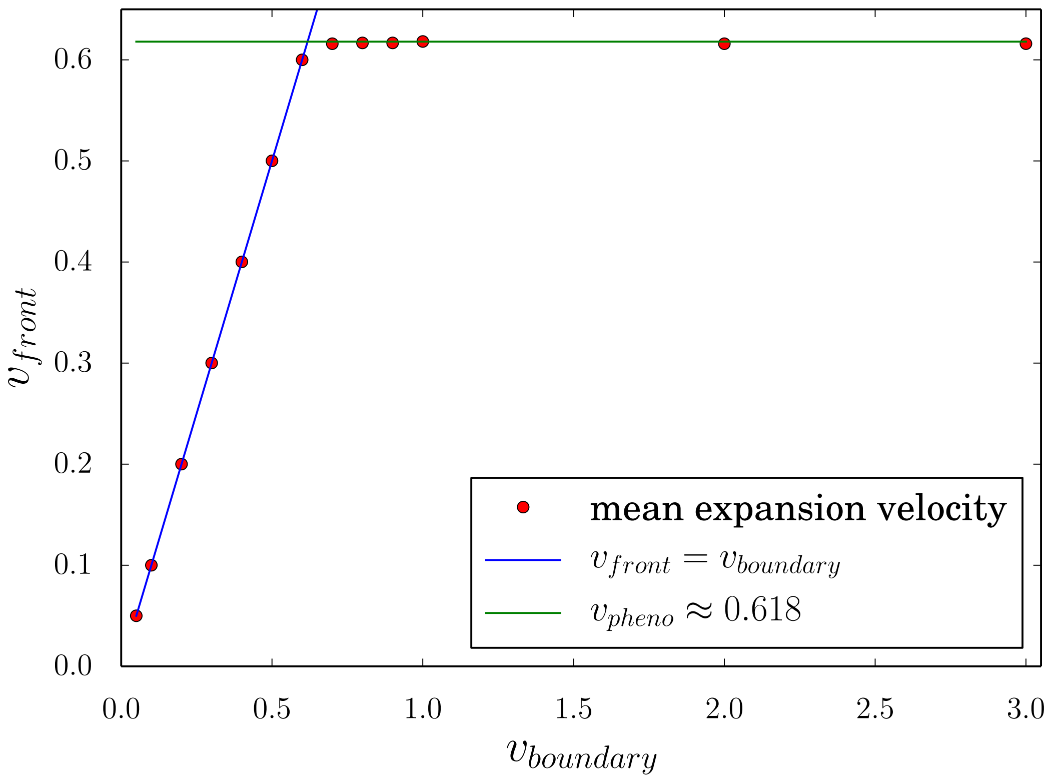


FIGURE 11.1.: **The expansion velocity as a function of the boundary velocity.** For small velocities, the expansion velocity matches the boundary velocity as expected. For boundary velocities larger than $v \approx 0.618$, the velocity of the front saturates. For the parameters of our model habitat, the phenotypical expansion velocity is, thus, $v_{\mathrm{pheno}} \approx 0.618$.

boundary and postpone the analysis of boundary shapes to the discussion and to future research.

For boundary velocities above the phenotypic expansion velocity, we are in the scenario of phenotype–limited range expansions. The population front takes a irregular and rough shape as expected in the context of the Eden growth model. Figure 10.4 shows an example of such an expansion.

11.2. Coalescence

We carried out simulations for the three versions of the coalescent described above. Within the three versions of the coalescent, the unquenched coalescent model represents a null model with respect to the possible quenched randomness due to the colonization process. Therefore, we begin our analysis with this version.

As a first step, we determine the critical patch size introduced in section 10.3.1. In the next step, we estimate the expansion direction from centered patch data (section 10.3.3). In a third step, we describe the shape of the mutation patches with respect to the expansion axis and develop predictions for realistic samples.

11.2.1. The unquenched coalescent

We assume that all demes are populated inside the habitat boundaries during the colonization, no specific colonization paths are assumed. The phenotypic expansion velocity in our simulations is $v_{\text{pheno}} \approx 0.618$ (Figure 11.1). Expansion velocities $v > v_{\text{pheno}}$ are therefore not realistic with respect to the forward model. Therefore, we simulated this version of the coalescent for velocities between $v = 0$ and $v = 0.6$.

11.2.1.1. The critical patch size

The definition of the critical patch size is based on the habitat dimensions. The compression factor of patches that span the entire width or length of the habitat is likely to be biased.

The mean values of SD_x and SD_y for a stationary habitat should have identical expectation unless the habitat dimensions limit the patch growth. Indeed, $\overline{\text{SD}_x}$ and $\overline{\text{SD}_y}$ deviate already for relatively small patch sizes (Figure 11.2). Clearly, the habitat width introduces a strong bias to the patch widths for patch sizes larger than 100 in this particular example.

To be able to quantify the impact of the limited habitat width, we determine the fraction of patches of critical width as a function of the patch size B. Figure 11.3 shows the results for $v = 0$. Figure 11.4 show the results for $v > 0$.

The actual choice of $B_{critical}$ can be made on the basis of these Figures. However, the choice is still arbitrary to a certain extend. Note that the critical patch size increases with the expansion velocity. For our parameter choice, only the largest patches must be excluded for $v \geq 0.4$.

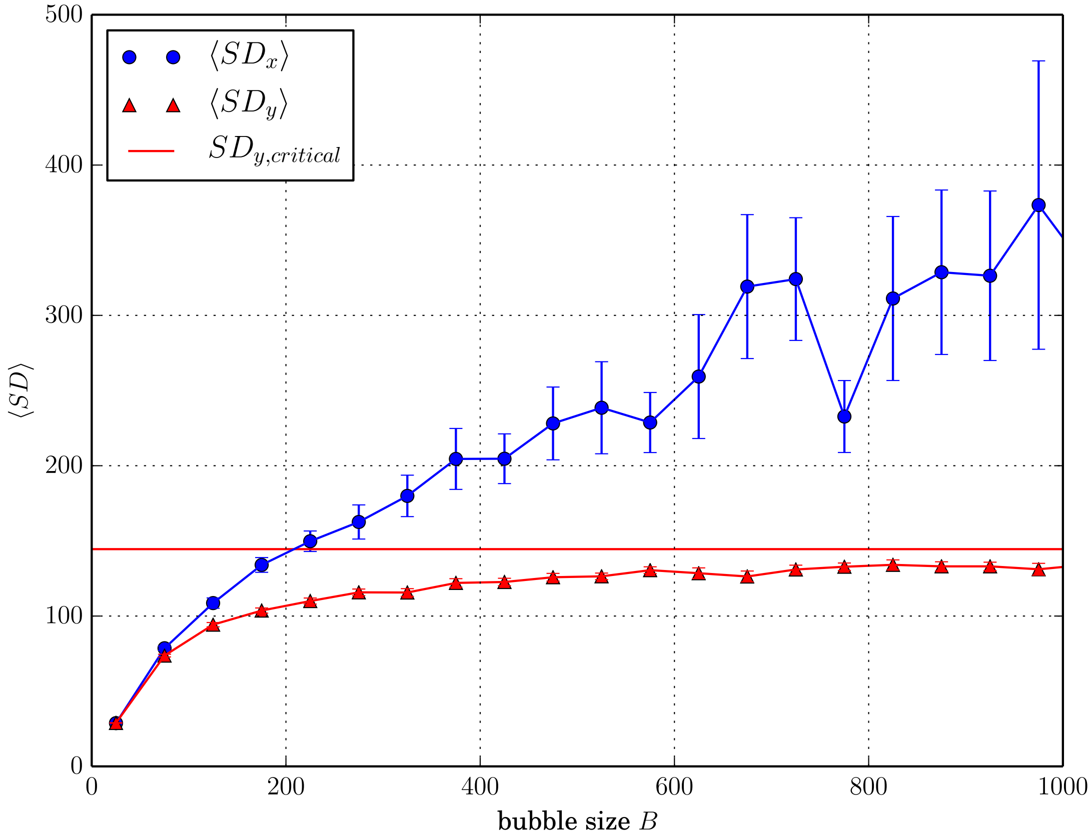


FIGURE 11.2.: **Mean patch dimensions for $v = 0$.** The mean values of SD_x and SD_y from a single coalescence with sample size $n = 50000$ in a stationary habitat are shown as a function of the patch size. The samples were drawn at random from the habitat. The patch sizes are binned. The error bars show the standard deviation of the sample. $\mathrm{SD}_{y,\text{critical}}$ refers to the critical patch width derived in equation (10.9).

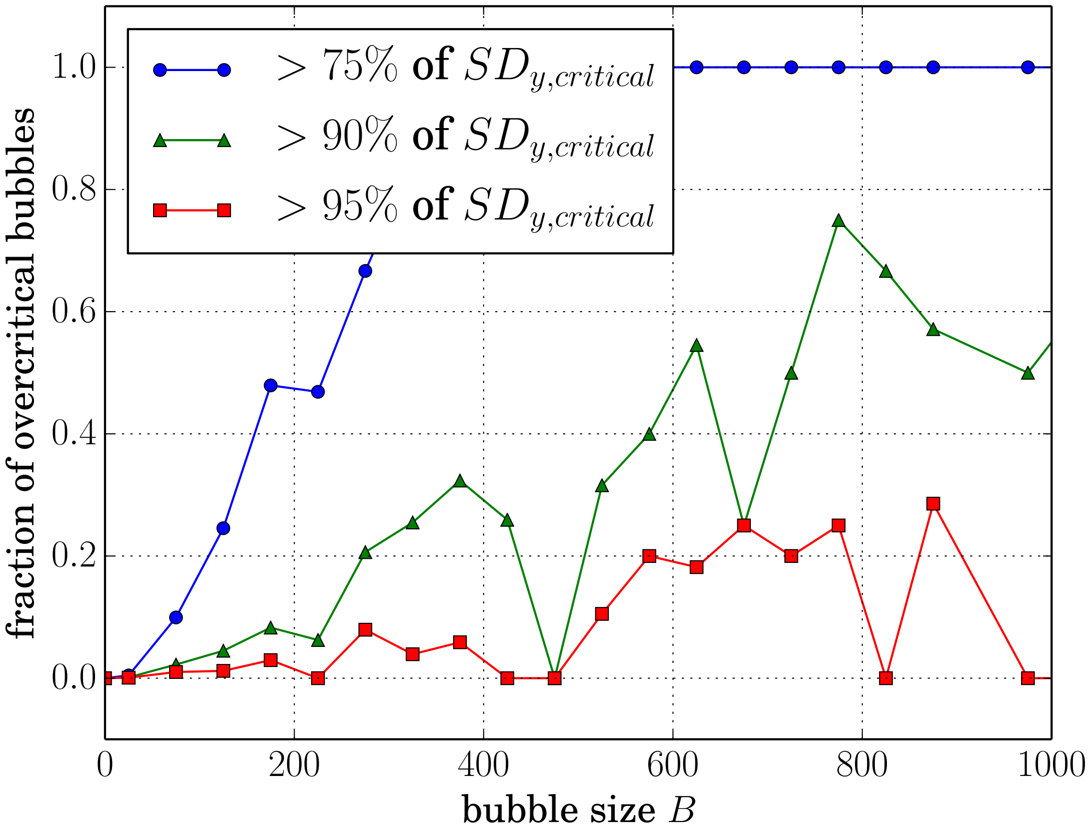


FIGURE 11.3.: **The fraction of overcritical patches for $v = 0$.** Even for small values of the critical patch size such as $B_{critical} = 100$, a significant fraction of the patches is likely to be influenced by the finite habitat width.

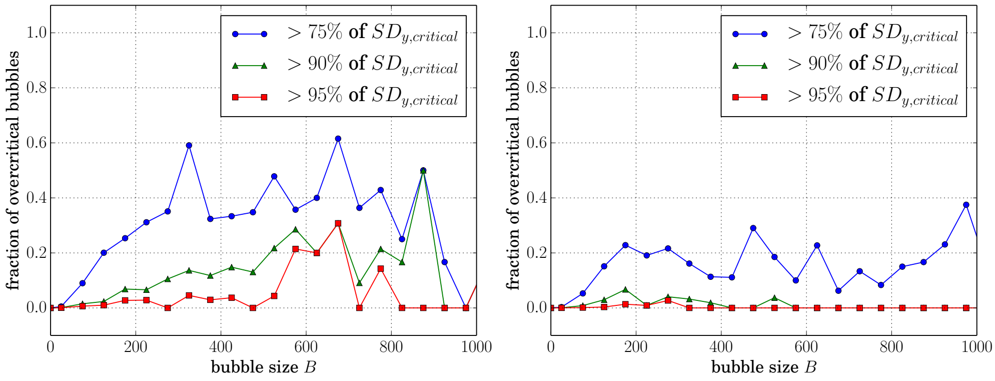
11.2.1.2. Estimation of the expansion direction

We applied the method introduced in section 10.3.3 to simulation data for different expansion velocities and all the three versions of the coalescent. For expansions of velocity close to the phenotype–limit, the first principal component is almost identical with the expansion axis of the phenotype–limited range expansions (Figure 11.5) as expected.

For the choice of $B_{critical} = 1000$, the first principal component explains 87% of the variance (Figure 11.6). For $B_{critical} = 10$ or $B_{critical} = 50$, however, the expansion direction can not be safely deduced from the PCA.

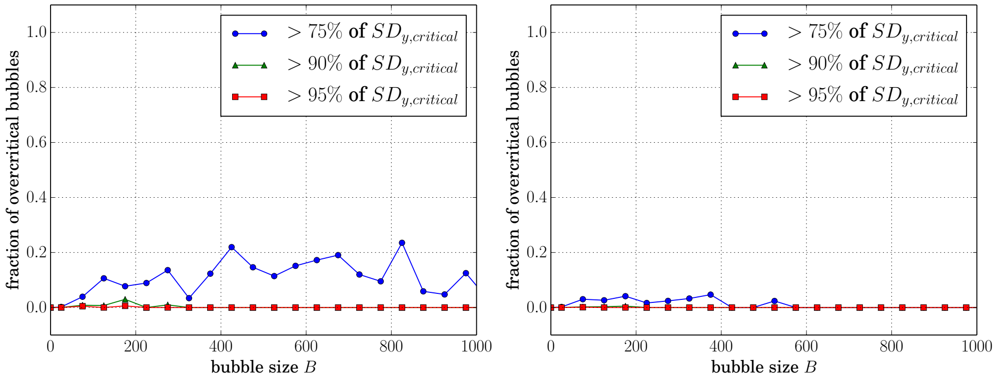
Note that without filtering, the first principal component always matches the long habitat axis (*c.f.* Figure B.4 and B.5).

For smaller boundary velocities, the first principal component explains less sample



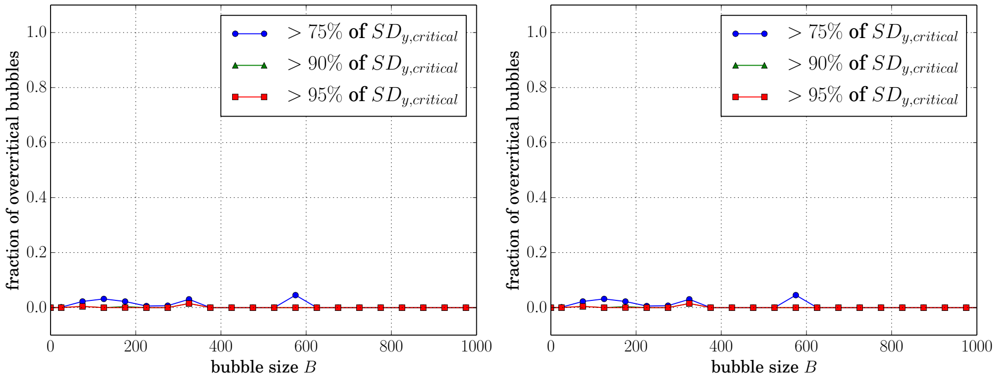
(a) $v = 0.1$.

(b) $v = 0.2$.



(c) $v = 0.3$.

(d) $v = 0.4$.



(e) $v = 0.5$.

(f) $v = 0.6$.

FIGURE 11.4.: **Fraction of overcritical patches.** For the data from Figure B.3, the fraction of patches that exceeds 75%, 90% and 95% of $\mathrm{SD}_{y,\mathrm{critical}}$ is shown. Interestingly, the fraction of overcritical patches is not increasing monotonically. In fact, in expanding habitats, large patches tend to correspond to surfing events.

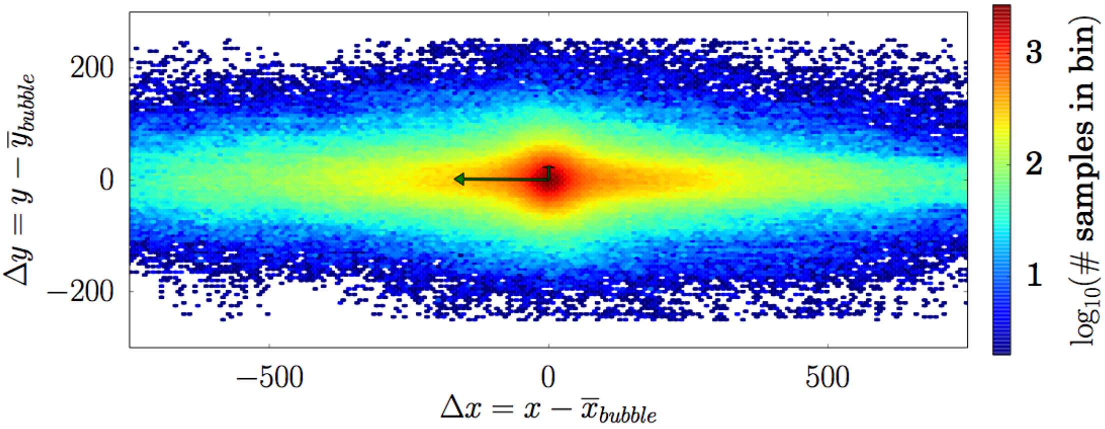


FIGURE 11.5.: **Centered patch histogram for $v = 0.6$ and $B_{critical} = 1000$.** The data shown here corresponds to a single coalescence simulation with sample size $n = 50000$. Note that the result is qualitatively the same, if more simulation runs are considered. The choice of $B_{critical}$ is justified by the observation in Figure 11.4 (f). The centered mutant positions show a clear sign of a range expansion. The arrows indicate the principal components. The first principal component explains 87% of the variation and has slope $m = 0.0072$.

variance. The identification of the expansion axis becomes more and more uncertain, see Figure 11.7 and Figures 11.8. Note that slower expansions have a weaker impact on the patch shapes and require the choice of a smaller value of $B_{critical}$. Our method's capacity of estimating the expansion direction is reduced by both factors.

For $v = 0$ and sensible filtering, the two principal components are equally important (Figure 11.9).

11.2.1.3. The shape of mutation patches

In order to be able to analyze the mutation patches in detail, we develop filters based on patch properties that can be calculated in experimental samples. First, we will describe the data for a boundary velocity of $v = 0.5$ in detail. Second, we will briefly summarize the corresponding results for the other values of v.

Experimental data sets usually provide information for only a single sample. Thus, we do not use pooled data from many independent simulations runs for our method but from only one simulation run per parameter set. Note, however, that incorporating more data sets does not lead to qualitatively different results.

We represent the mutation patches as tuples (SD_x, SD_y) in a two–dimensional

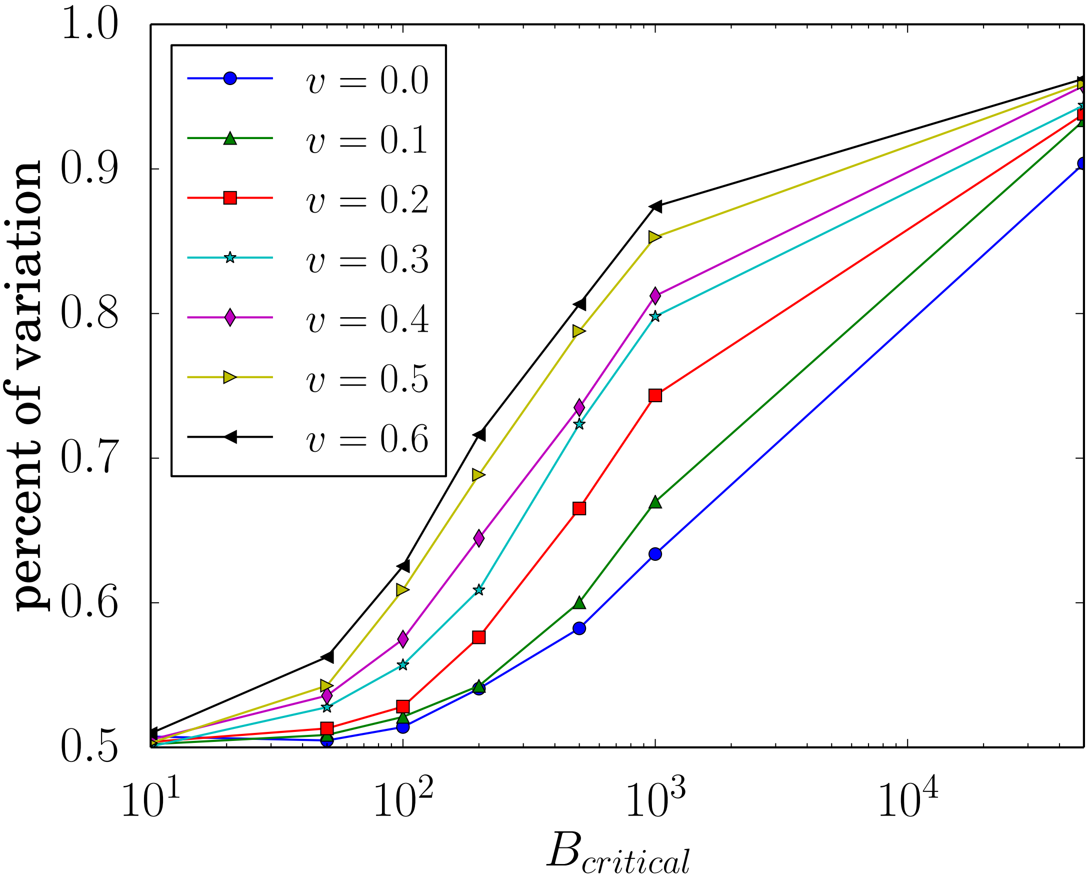


FIGURE 11.6.: **Sample variance explained by the first principal component.** Note that small patches do not provide enough information to deduce the expansion direction. The high values for $B_{critical} > 1000$ reflect the habitat shape rather than the range expansion.

histogram. The unfiltered data for $v = 0.5$ (Figure 11.10) shows no influence of the habitat boundaries as expected from our previous analysis. However, the data still contains tiny patches ($B < 5$). Such patches correspond to very rare mutations and do not contain much information about the expansion process.

For the following analysis, we will only consider patches of size $B > 5$. We will see later on that a slightly higher threshold does not influence the analysis. In the representation of the filtered data (Figure 11.11), we clearly distinguish two clusters, a third one is less apparent.

Figure 11.12 shows the data from Figure 11.11 with each datapoint colored according to the size of the corresponding mutation patch. Note that, roughly speaking, small

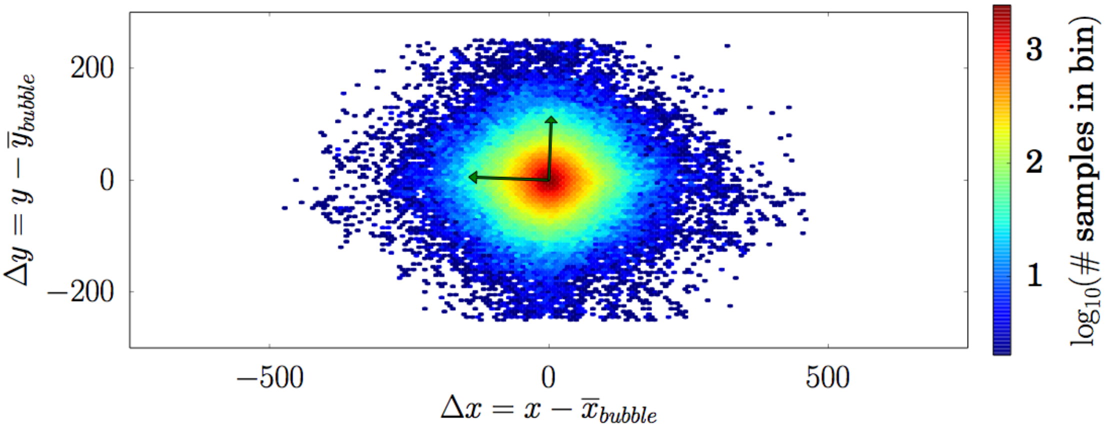


FIGURE 11.7.: **Centered patch histogram for $v = 0.3$ and $B_{critical} = 100$.** The data shown here corresponds to a single coalescence simulation with sample size $n = 50000$. The choice of $B_{critical}$ is justified by the observation in Figure 11.4. The centered mutant positions show a weak sign of a range expansion. The arrows indicate the principal components. The first principal component explains 56% of the variation and has slope $m = -0.0379$.

patches belong to the the large cluster on the left.

A total least squares regression of the small patches data (Figure 11.13) has a slope of $m \approx 1$. This pattern matches our expectation for patches that were not affected by the expansion process (see section 10.3.2).

The large patches are a little more difficult to analyze (Figure 11.14): we observe a power–law like structure as mentioned before and a less clearly defined group 'above the power law'. The slope of the apparent power–law is roughly 0.5. This matches our assumptions for the shape of surfer patches (again, see section 10.3.2).

Based on the assumption, that these patches were in fact influenced by the expansion process, we define additional filters: Mutations that surfed on the wave of advance until the end of the colonization will leave mutants close the right edge of the habitat at $x = 10000$. In fact, if we further restrict our attention to patches with at least on sampled individual at $x > 9000$, we can (almost) isolate the apparent power law (Figure 11.15). Selecting the patches according to, for instance, $\overline{x} > 8000$ gives a similar but less accurate result.

If a surfing mutation loses contact to the expansion front, its boundaries will diffuse along both dimensions. The mean increase of absolute width and length has identical expectation. According to the data we presented so far, a surfer patch's SD_x–value is typically at least one order of magnitude larger than its SD_y–value. Therefore, if a mutation loses contact to the front the relative increase in width is

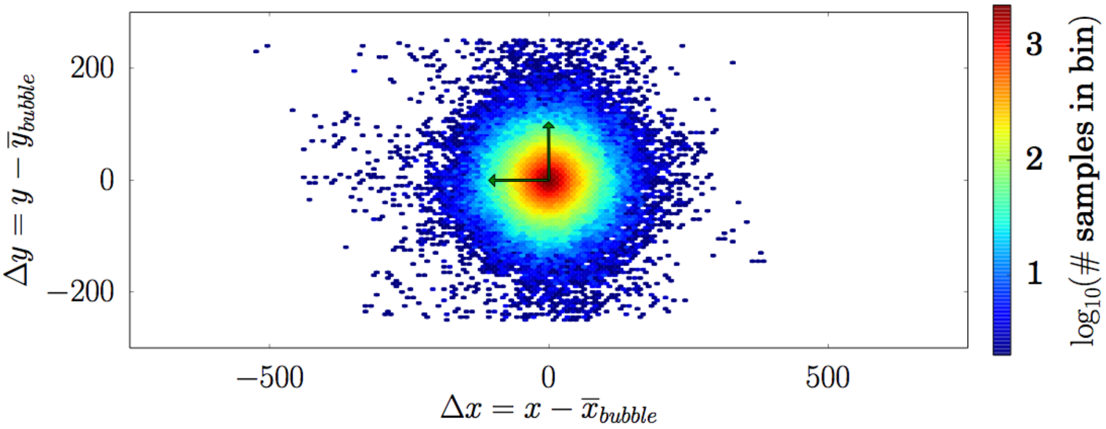


FIGURE 11.8.: **Centered patch histogram for $v = 0.1$.** If patches of size $B > 50$ are excluded, the centered mutant position show no sign of a range expansion. The first two principal components capture around 50% of the variation.

larger than the relative increase in length. Metaphorically speaking, surfer patches start moving upwards in the SD–Figures as soon as they drop from the expansion front.

The phenomenon described in the last paragraph matches the so far unexplained group of patches ‘above the power law’ We underpin that statement with yet another filter: coloring the patches according to the mutation age at the sampling time, we observe that moving upwards from the power–law bar, corresponds to a change into older mutation classes.

Note that this last observation is easily applicable in the simulation context but maybe impossible to apply with experimental data. All previous filters were applicable in both situations.

We finish the characterization of the patch shapes by analyzing the distance from the expansion front at which the mutations occurred. As we do not specify, in what individual along an edge of the coalescent tree the mutation of the corresponding patch occurs, we decided to describe the two extreme cases and mutate the oldest (Figure 11.15) and the youngest individual of the edge (Figure 11.15) and measure the distance of the respective individual to the front during its life. As expected, almost all surfer mutations occurred at the front.

Bubble shapes for different velocities The analysis presented above for $v = 0.5$ gives corresponding results for the other values of v we ran simulations with. The Figures 11.19 to 11.22 show the data. For velocities $v \leq 0.2$ our method does not detect a clear sign of the range expansion.

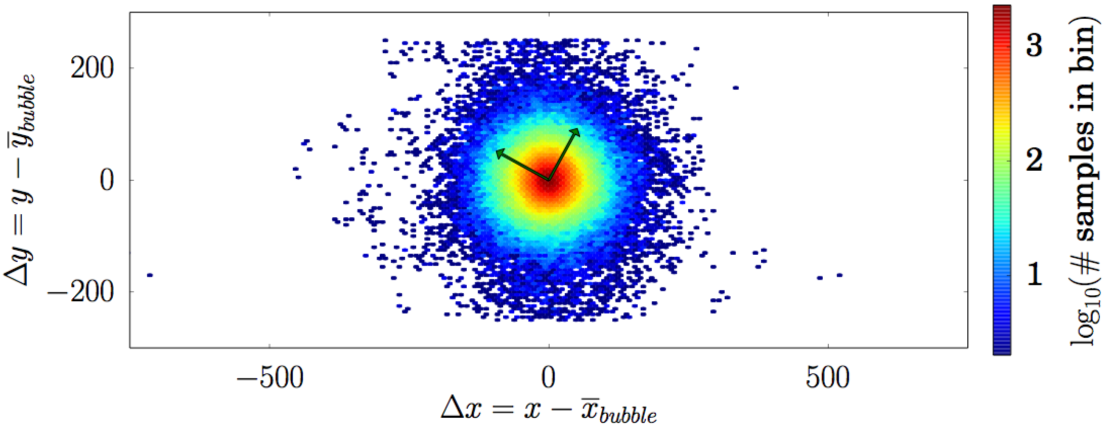


FIGURE 11.9.: **Centered patch histogram for $v = 0$.** If patches of size $B > 50$ are excluded, the centered mutant position show no sign of a range expansion. The first two principal components capture around 50% of the variation, each.

11.2.1.4. Bubble probabilities

We create the mutant patches on the base of the coalescence tree of the sample and analyzed all potential mutant patches. However, a mutant patch is present in the sample if and only if a mutation occurred long the corresponding edge of the coalescent tree.

The expected number of mutant patches of size B must therefore be proportional to the total length in generations of all edges with progeny of size B in the sample. Note that this includes patches corresponding to the same edge.

Figure 11.23 shows the number of edges and the mean edge lengths as a function of the binned patch size for the stationary habitat and $v = 0.6$, respectively. The product of both describes the relative abundance of small patches as compared to large patches. The Figures B.14 to B.16 show the results for the other values of v. The expected absolute number of patches depends on the mutation rate per locus and the number of sequenced loci.

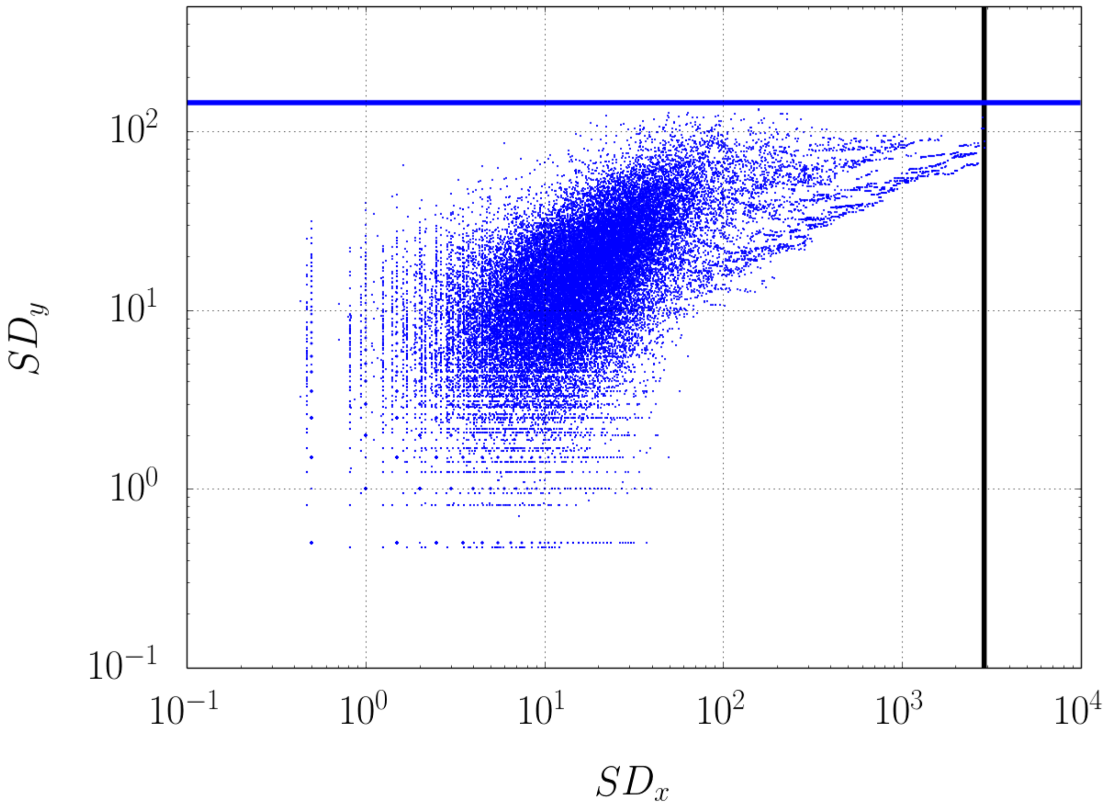


FIGURE 11.10.: **Unfiltered patch shapes for** $v = 0.5$ The thick blue horizontal and black vertical lines show the threshold values for SD_x and SD_y derived in section B.2.2.

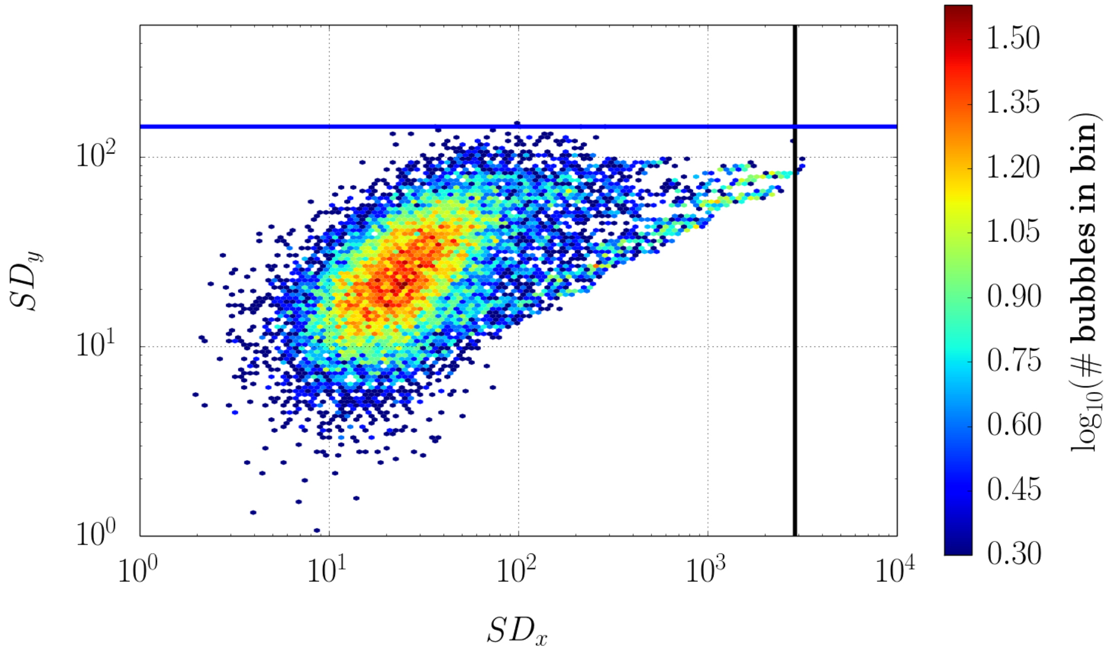


FIGURE 11.11.: **Histogram of the patch shapes for $v = 0.5$.** For sample size $n = 5 \times 10^4$, the coalescent tree has up to $2n - 1$ edges. For each edge, the variances along each of the coordinate axes for the corresponding mutation patch are shown in a histogram with hexagonal logarithmic bins. The thick blue horizontal and black vertical lines show the threshold values for SD_x and SD_y derived in section B.2.2. Two main clusters are apparent: one large group on the left and a power–law like structure on the right.

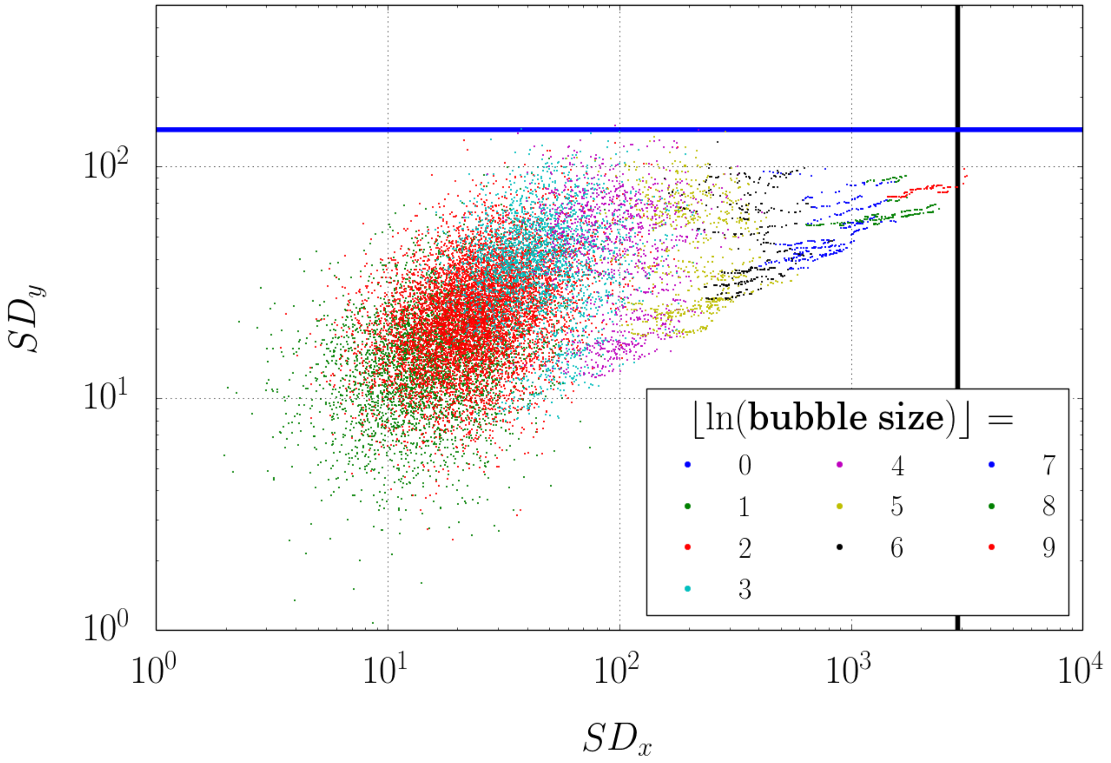


FIGURE 11.12.: **Bubble shapes colored according to the patch size** Note that smaller patches belong typically to the large cluster to the left whereas larger patches are disproportionately extended along the habitat axis and tend to fall in the second cluster. The thick blue horizontal and black vertical lines show the threshold values for SD_x and SD_y derived in section B.2.2.

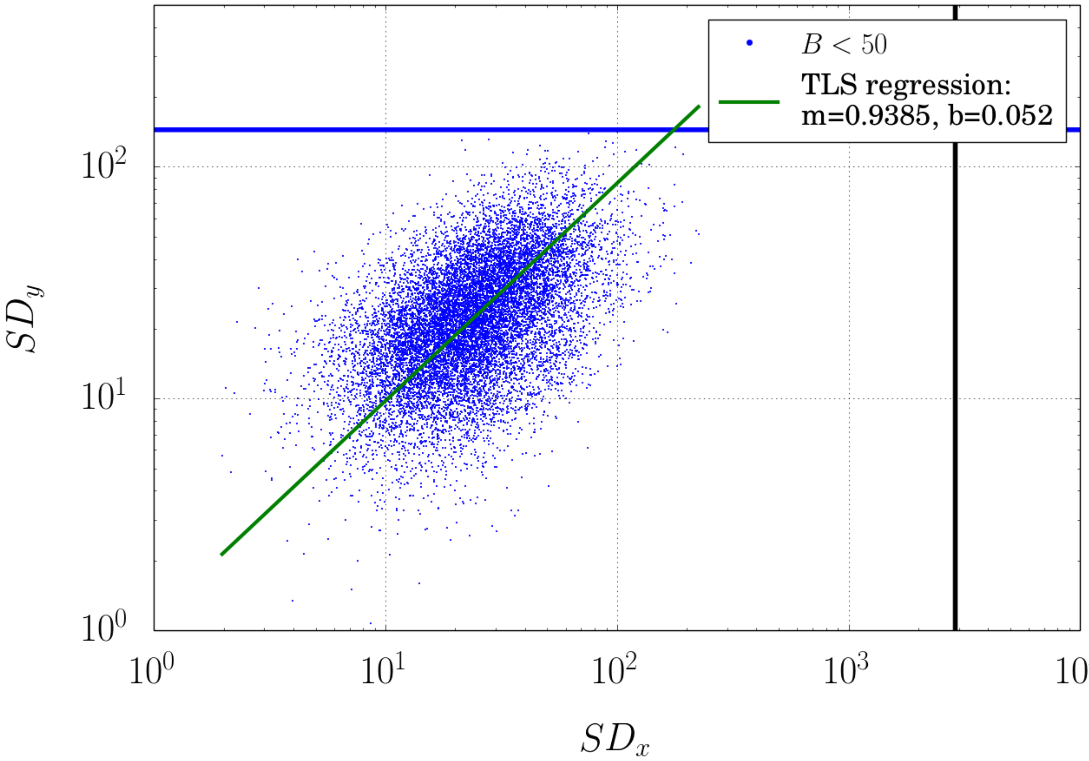


FIGURE 11.13.: **Small patch shapes for $v = 0.5$.** The patches of size $B < 50$ form a separate cluster. The total least squares regression of the logarithmic SD–values has slope of almost 1. The thick blue horizontal and black vertical lines show the threshold values for SD_x and SD_y derived in section B.2.2.

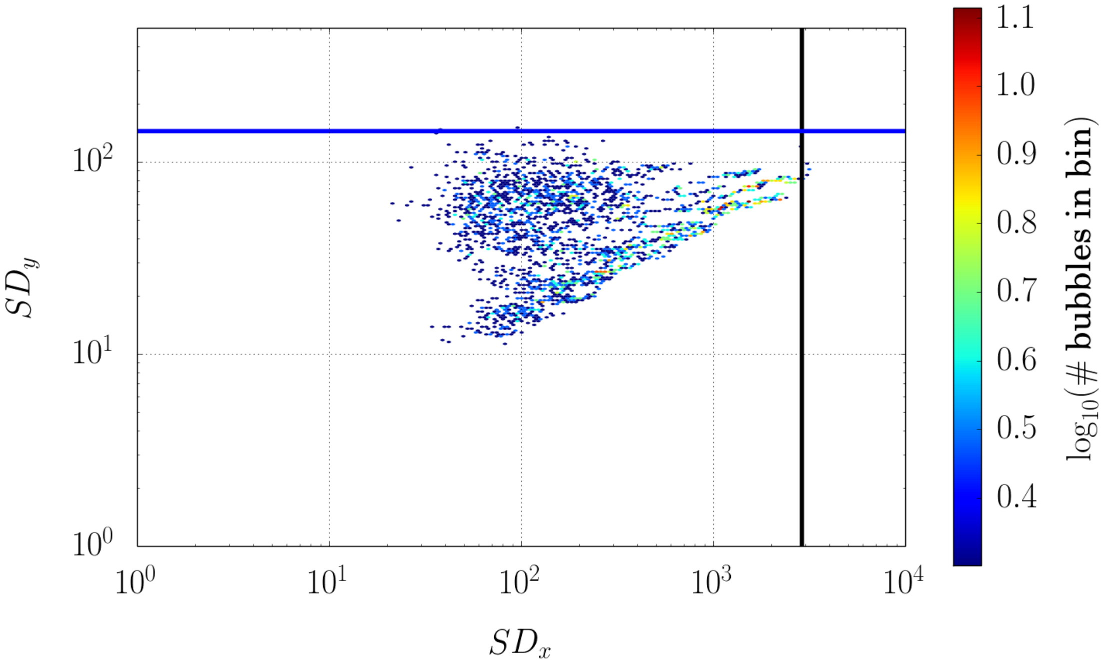


FIGURE 11.14.: **Large patch shapes for $v = 0.5$.** The patches of size $B > 50$ form a two clusters. The lower one resembles a power–law with exponent 0.5, the upper one has no apparent structure. The thick blue horizontal and black vertical lines show the threshold values for SD_x and SD_y derived in section B.2.2. The thick blue horizontal and black vertical lines show the threshold values for SD_x and SD_y derived in section B.2.2.

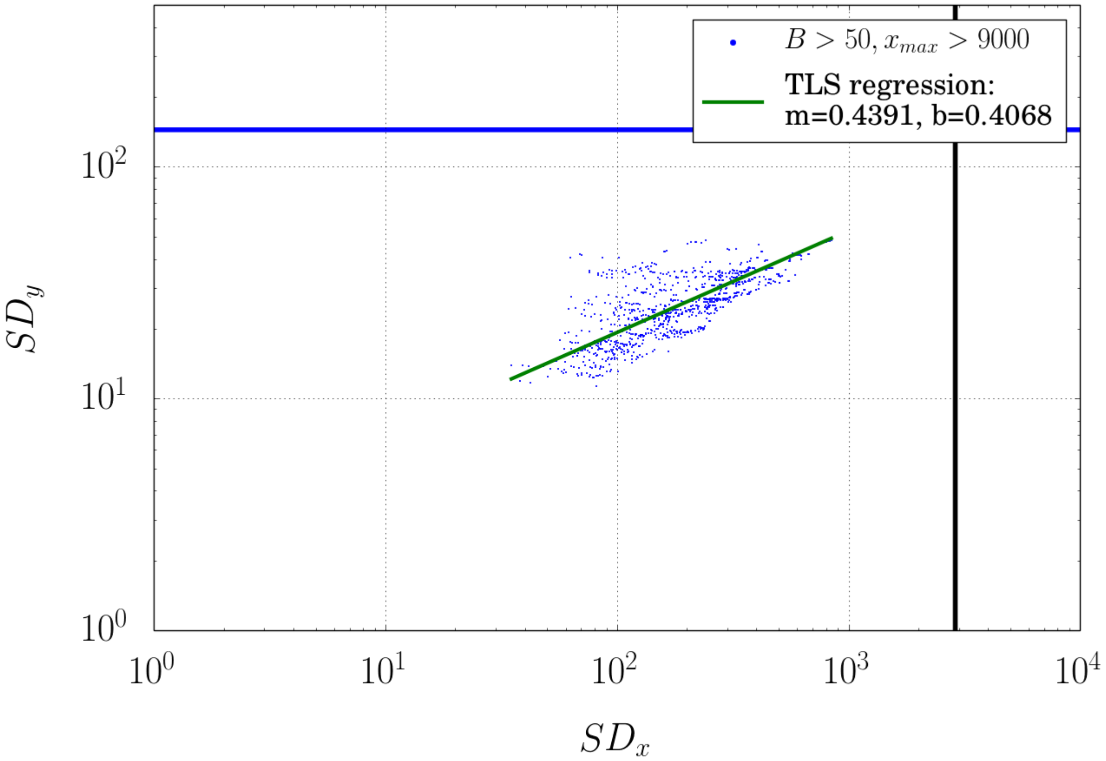


FIGURE 11.15.: **Large patch shapes for $v = 0.5$ close to the expansion front.** The restriction to patches with at least one individual with $x > 9000$ isolates the power–law like cluster. The slope ≈ 0.44 of the total least squares regression is reasonably close to the expected 0.5. Note that we filtered with a relatively simple method. We did not expect the result to be accurate. The thick blue horizontal and black vertical lines show the threshold values for SD_x and SD_y derived in section B.2.2.

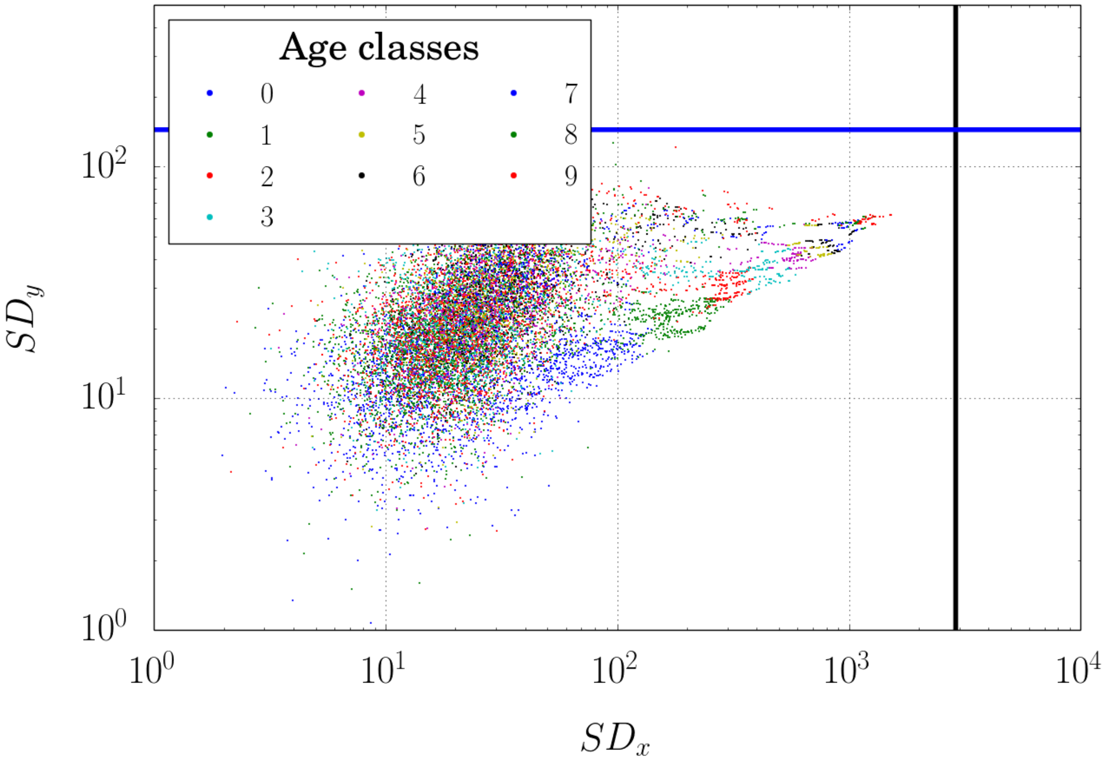


FIGURE 11.16.: **Bubble shape coloring according to their mutation age.** Each age class i corresponds to a mean mutation age between $i \times 1000$ and $(i+1) \times 1000$ generations. The cluster of the smaller patches includes patches of all ages without obvious order. The age classes in the power–law bar are well sorted. Moving to higher SD_y–values from the power–law bar corresponds to a change into older mutation classes. This behavior is in agreement with our expectation for surfer mutations that loose contact to the front after a while. The thick blue horizontal and black vertical lines show the threshold values for SD_x and SD_y derived in section B.2.2.

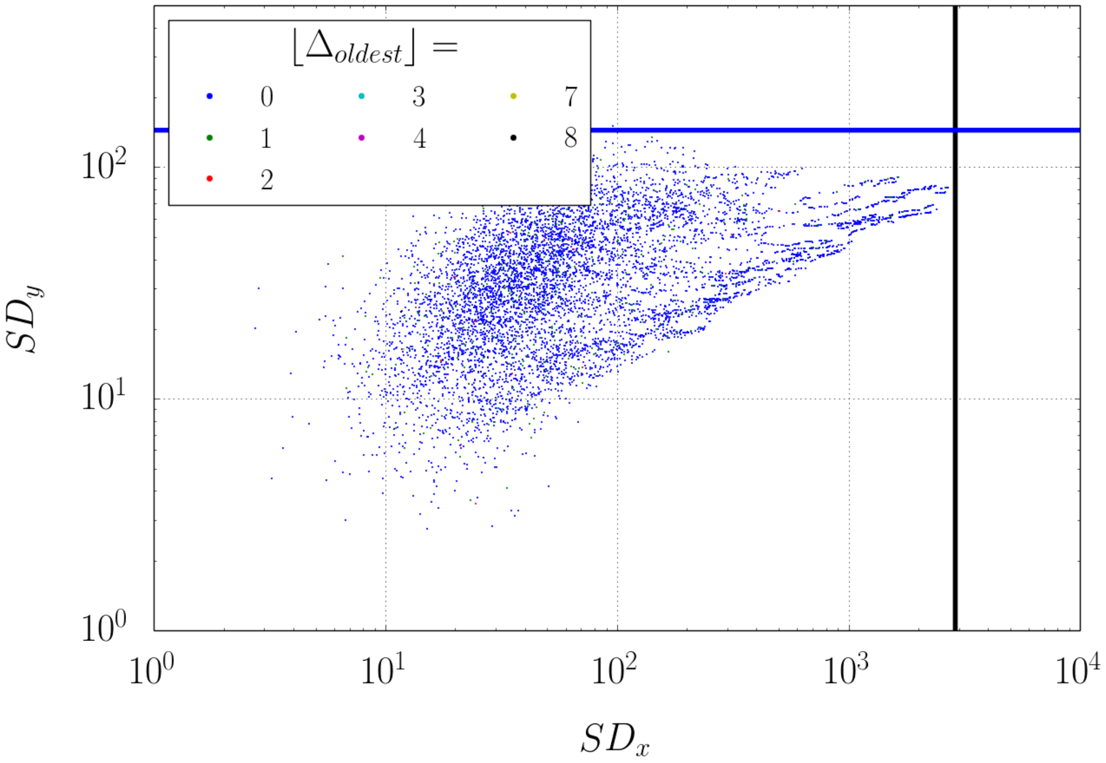


FIGURE 11.17.: **Bubble shape coloring for $B > 50$ according to the distance of the mutation event to the expansion front.** Δ_{oldest} refers to the distance of the mutation event to the population front. Here we assume that the mutation occurs in the oldest individual along the mutated edge of the coalescent (*cf.* Figure 11.17). Note that almost all successful mutations have occurred directly at the front. The thick blue horizontal and black vertical lines show the threshold values for SD_x and SD_y derived in section B.2.2.

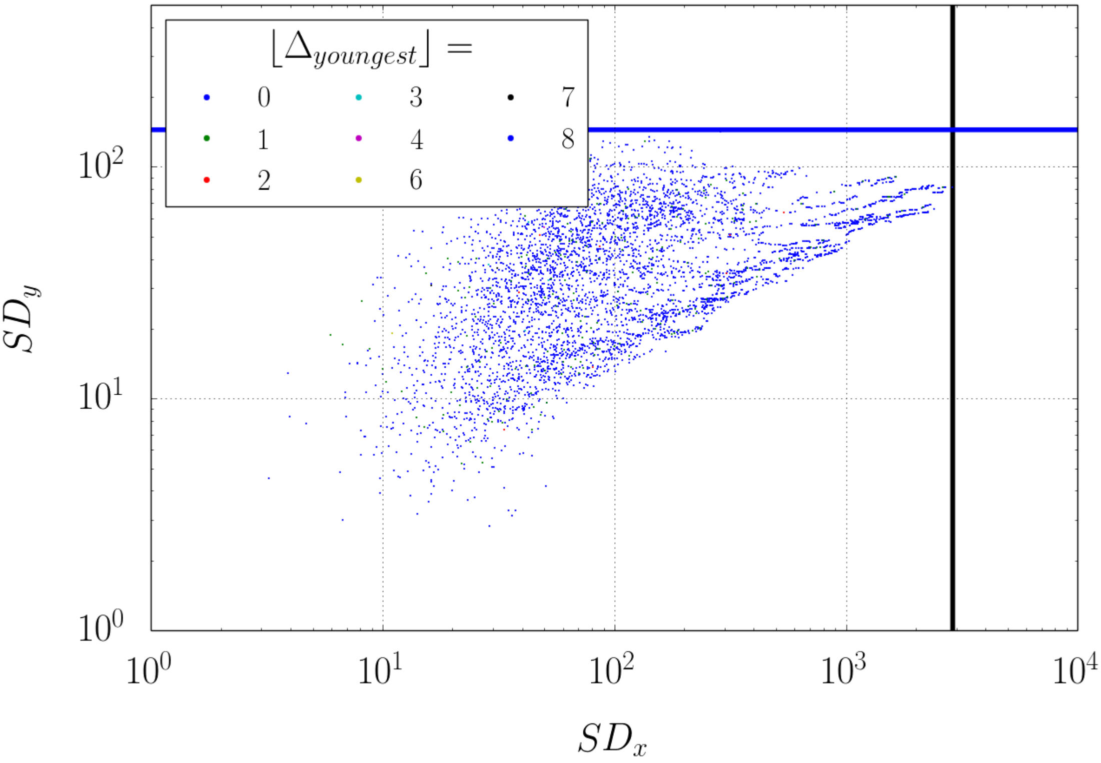
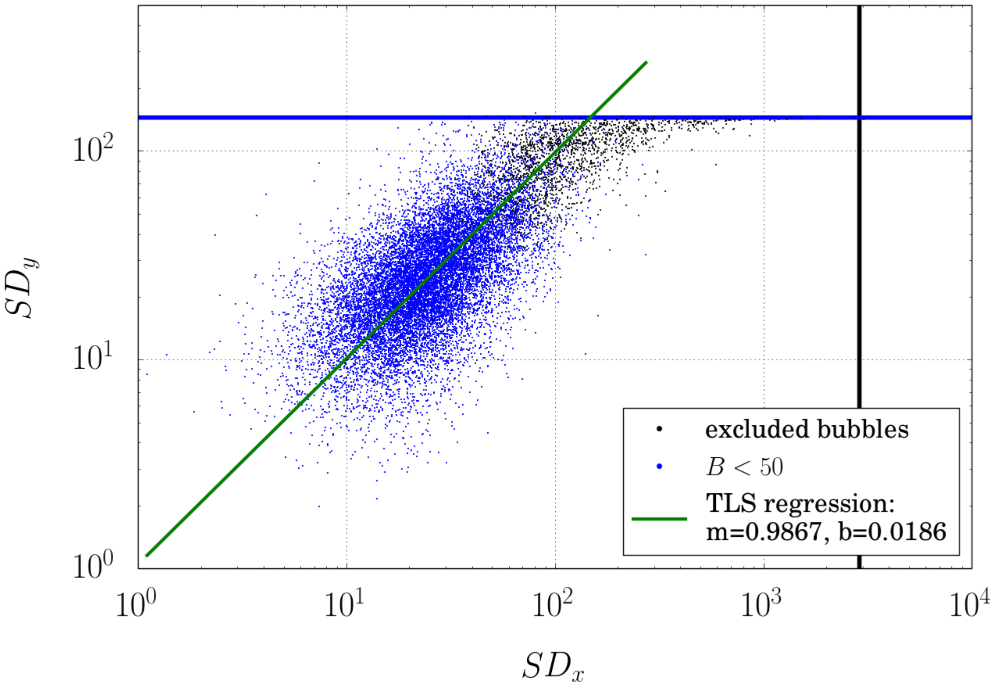
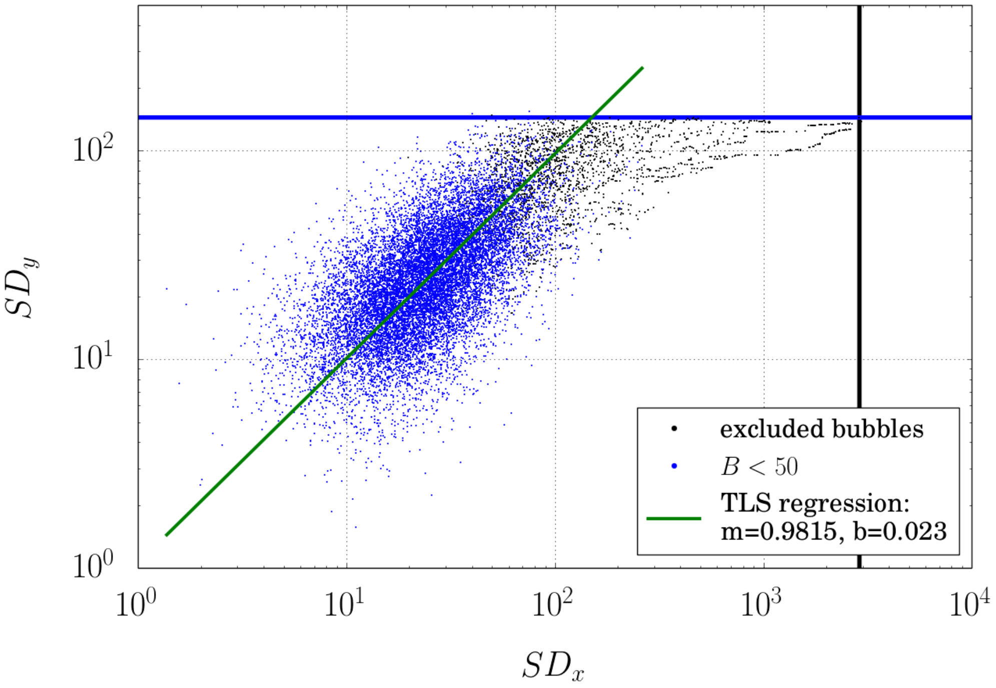


FIGURE 11.18.: **Bubble shape coloring for $B > 50$ according to the distance of the mutation event to the expansion front.** $\Delta_{youngest}$ refers to the distance of the mutation event to the population front. Here we assume that the mutation occurs in the youngest individual along the mutated edge of the coalescent (*cf.* Figure 11.18). Note that almost all successful mutations have occurred directly at the front. The thick blue horizontal and black vertical lines show the threshold values for SD_x and SD_y derived in section B.2.2.

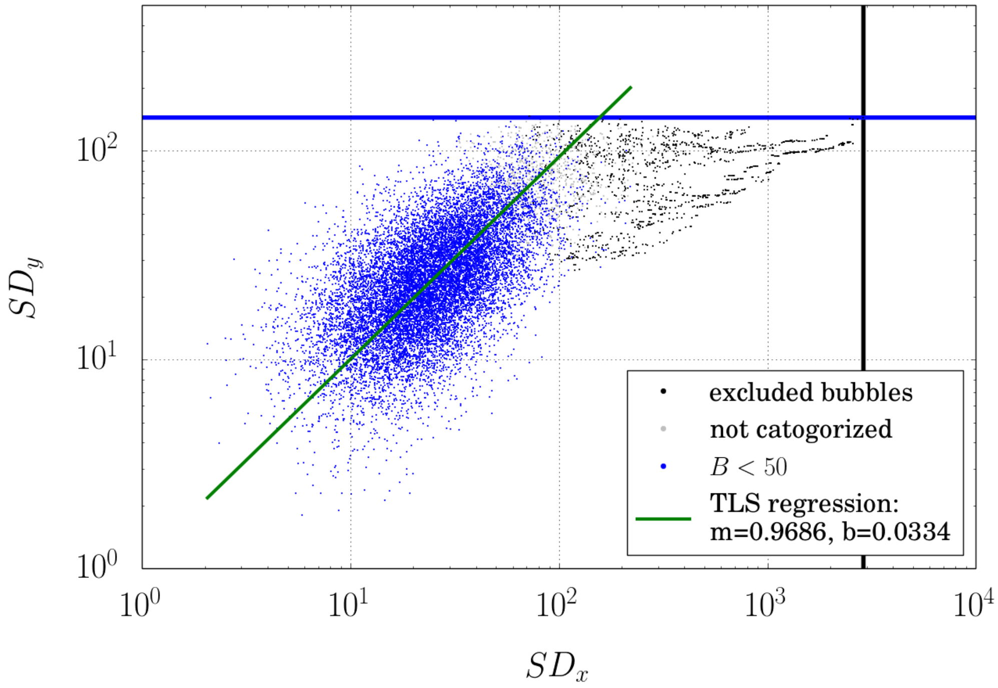


(a) $v = 0.0,\ B_{critical} = 50$.

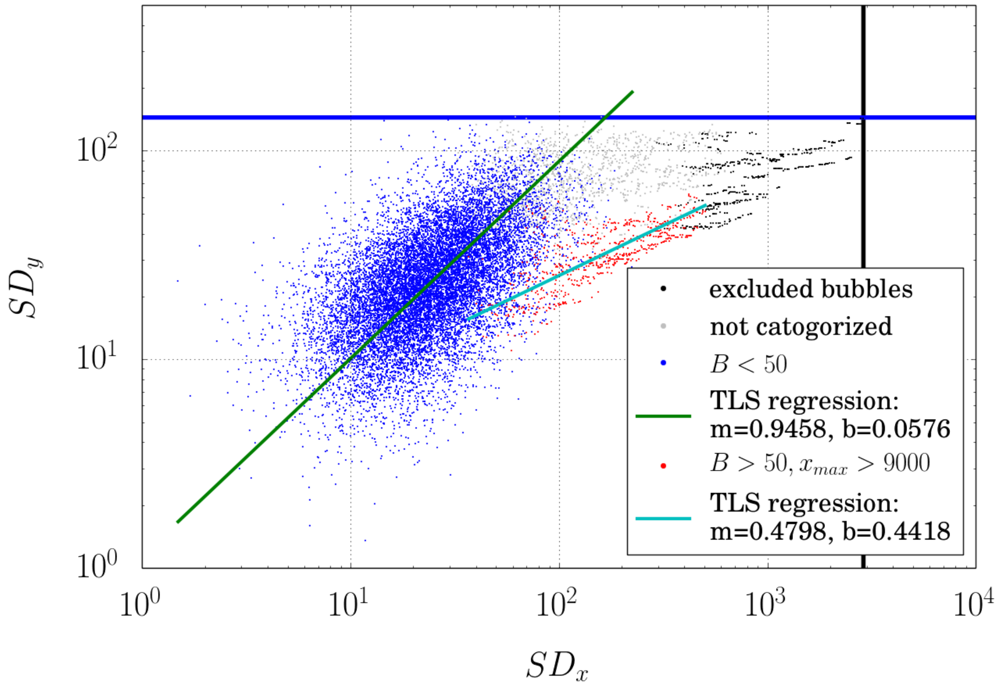


(b) $v = 0.1,\ B_{critical} = 50$.

FIGURE 11.19.: Bubble shapes for different expansion velocities. For a description see Figure 11.22.

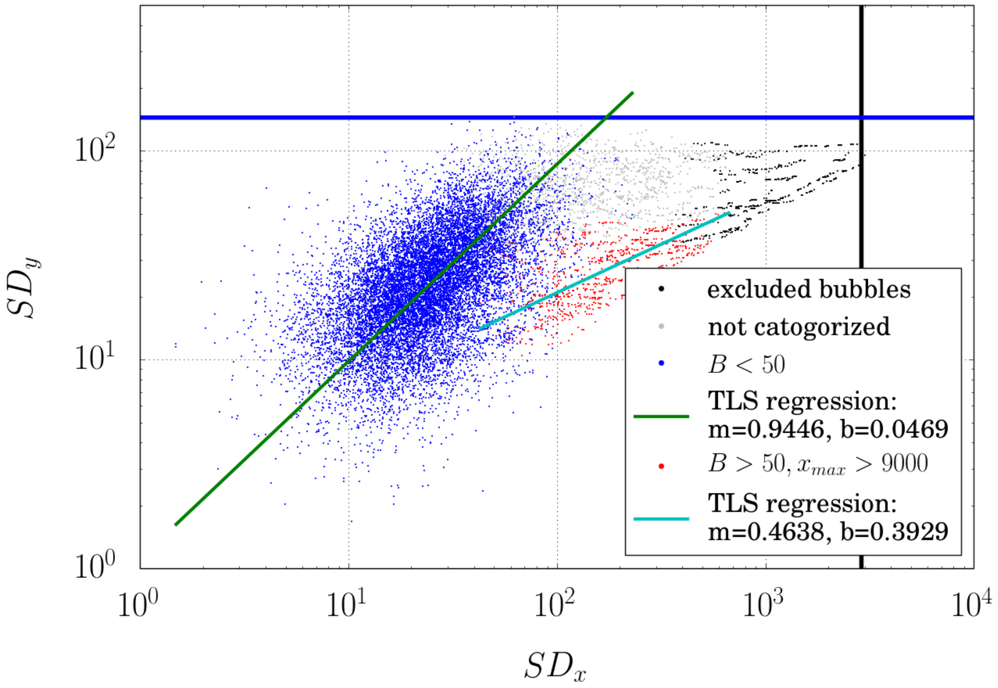


(a) $v = 0.2,\ B_{critical} = 150.$

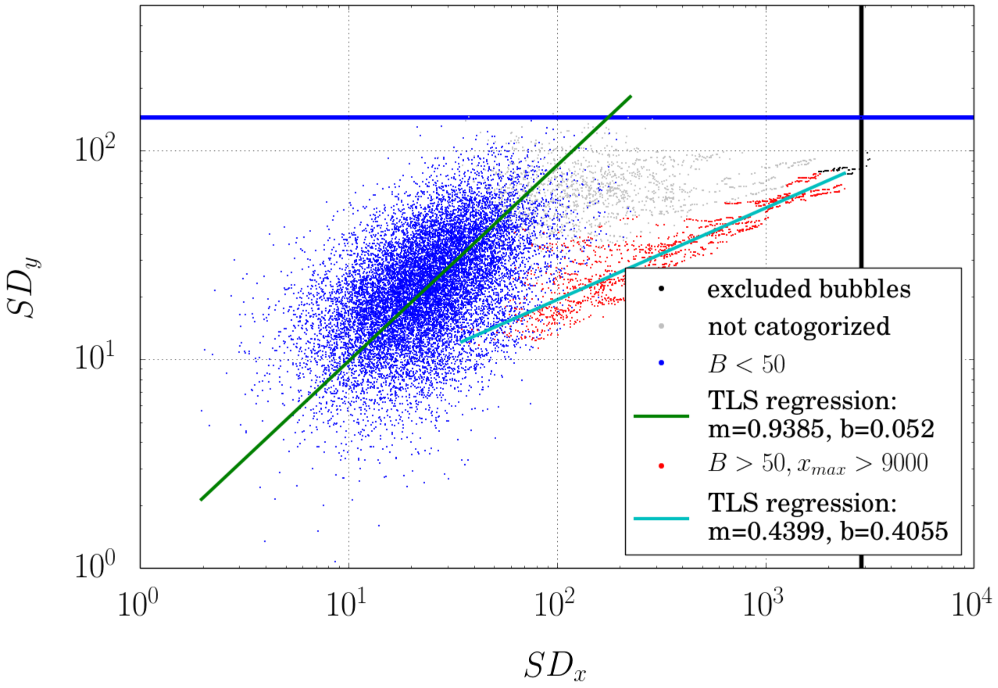


(b) $v = 0.3,\ B_{critical} = 1000.$

FIGURE 11.20.: Bubble shapes for different expansion velocities. For a description see Figure 11.22.



(a) $v = 0.4,\ B_{critical} = 1000$.



(b) $v = 0.5,\ B_{critical} = 10000$.

FIGURE 11.21.: Bubble shapes for different expansion velocities. For a description see Figure 11.22.

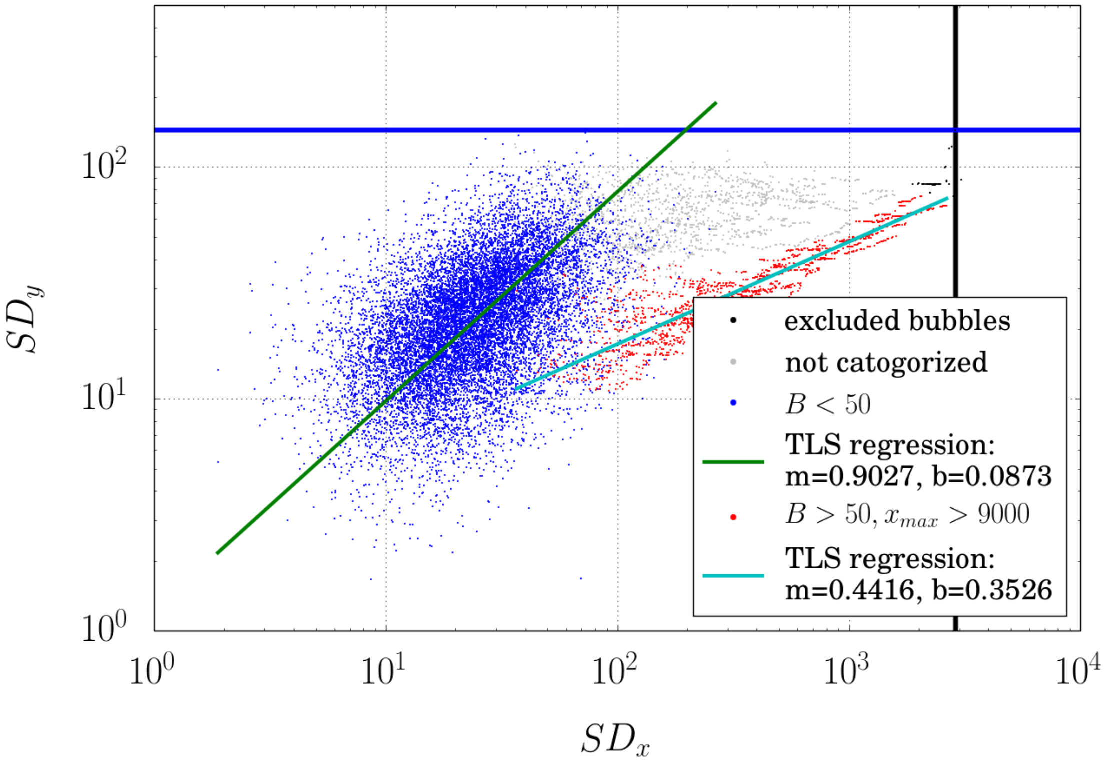
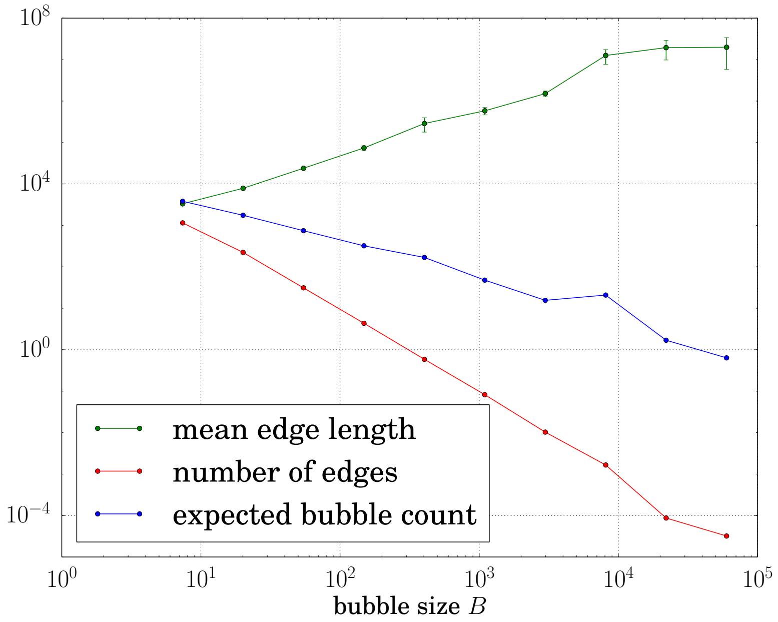
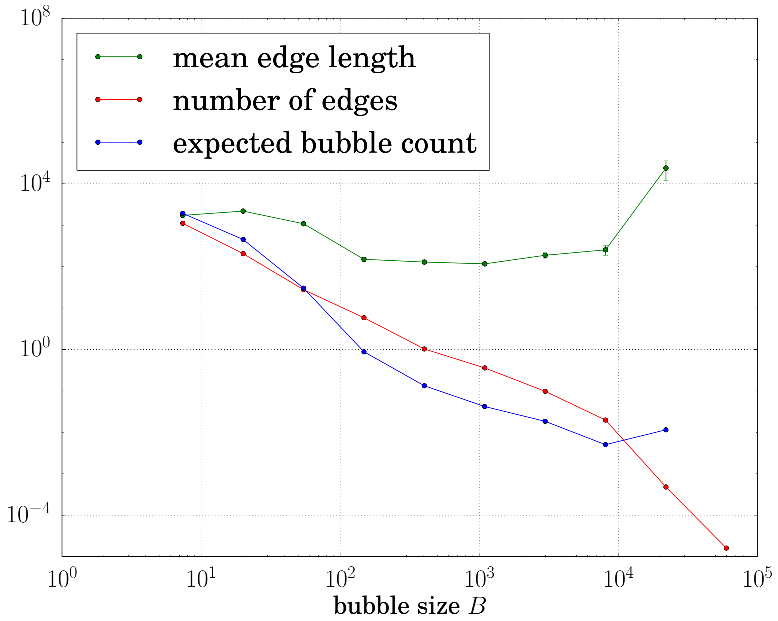


FIGURE 11.22.: $v = 0.6$, $B_{critical} = 10000$. The patch shapes for the velocities in our setup can be summarized as follows: For expansion velocities $v \geq 0.4$, we can clearly distinguish two clusters. One corresponds to small patches, the total least squares regression has a slope of appoximately 1. The second cluster corresponds to surfer patches, the total least squares regression has a slope of close to 0.5. For expansion velocities $v \leq 0.3$, the small patch cluster remains, but no other cluster can be clearly inferred. Recall that these results hold for our choice of the habitat dimensions. For other choices, the results will differ quantitatively.



(a) **Bubble occurrence for** $v = 0$.



(b) **Bubble occurrence for** $v = 0.6$.

FIGURE 11.23.: The mean number edges on the coalescent tree and the mean edge length in generations are displayed as a function of the patch size. The patch sizes are binned logarithmically. The expected patch count refers to the number of patches expected in a sample. Of course, it depends on the mutation rate. Here we display the product of the number of edges and the mean edge length (rescaled by 10^{-3} for convenience).

11.2.2. The fully quenched coalescent

After the description of the unquenched coalescent, we will now describe the other extreme: the fully quenched coalescent. Note that, as reproduction halts behind the front, all patches are surfer patches in this case.

Recall that stationary habitats cannot be modeled in this version and that it is sufficient to simulate a single boundary velocity above the phenotypical expansion velocity.

11.2.2.1. Estimation of the expansion direction

The estimation of the expansion direction can be executed with the same method as above. The results of the principal component analysis of the centered patches is shown in Figure 11.24. The signal of the range expansion is clear for all boundary velocities.

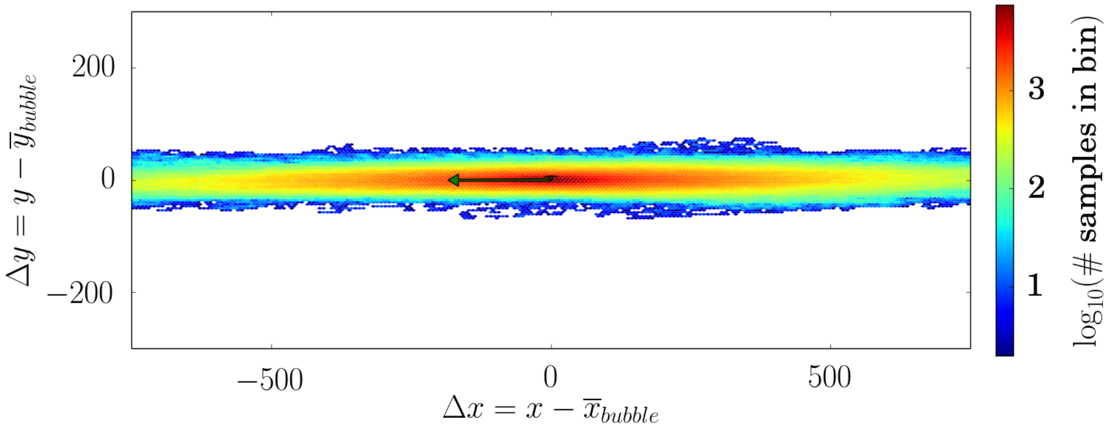
Figure 11.24 shows the results for $v = 0.1$ and $v = 0.2$, that is, for boundary–limited range expansions. Figure 11.25 shows the results for $v = 0.6$ and $v = 1.0$: $v = 0.6$ is only slightly below the phenotypical velocity limit and represents the transition between the two expansion regimes. $v = 1.0$ is clearly larger than the phenotypical velocity limit and can be considered as a phenotype–limited range expansion.

Note that the patch size limit $B_{critical} = 1000$ is sufficient for all velocities. A separate analysis of the critical patch size is not necessary. However, we can not include all patches, as the patch of size 50000 covers the full habitat, by construction.

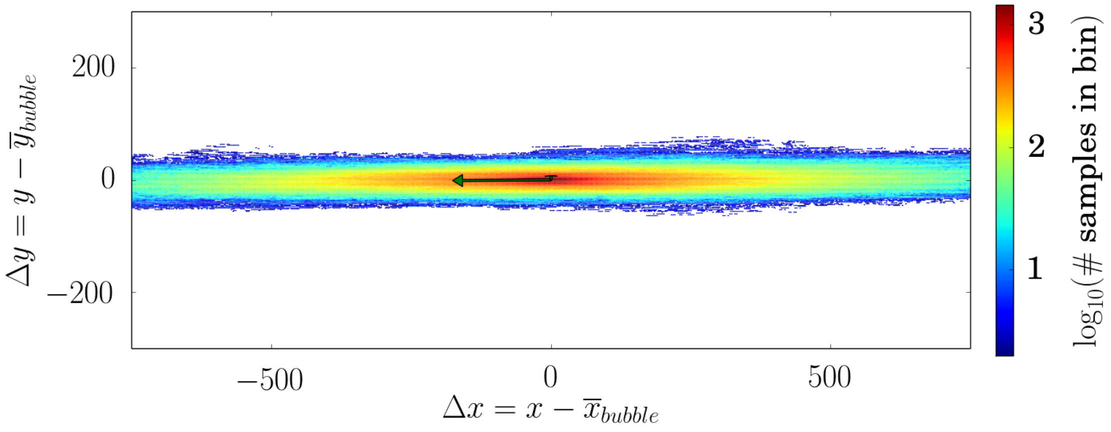
11.2.2.2. The shape of mutation patches

As all mutations surf, we see only one cluster for the full sample in the SD-plot. Figure 11.26 shows the result for $v = 0.1$, that is, for a boundary–limited range expansion. As every colonization results in different detailed colonization paths, we present data for 10 independent colonizations. The slope of the total least squares regression for the pooled data set is $m \approx 0.59$ for all individual colonizations for $v = 0.1$.

Figure 11.27 shows the patch shapes colored according to the patch size. The result is as expected. Figure 11.28 shows the patch shapes colored according to the colonization they belong to. Differences between the colonies are apparent only for the largest patches. The general pattern is the same for all colonizations.

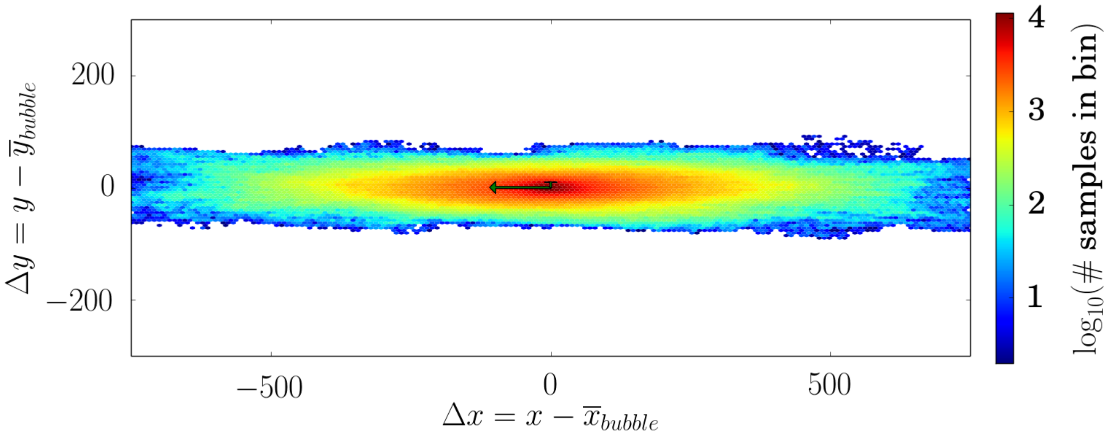


(a) $v = 0.1$.

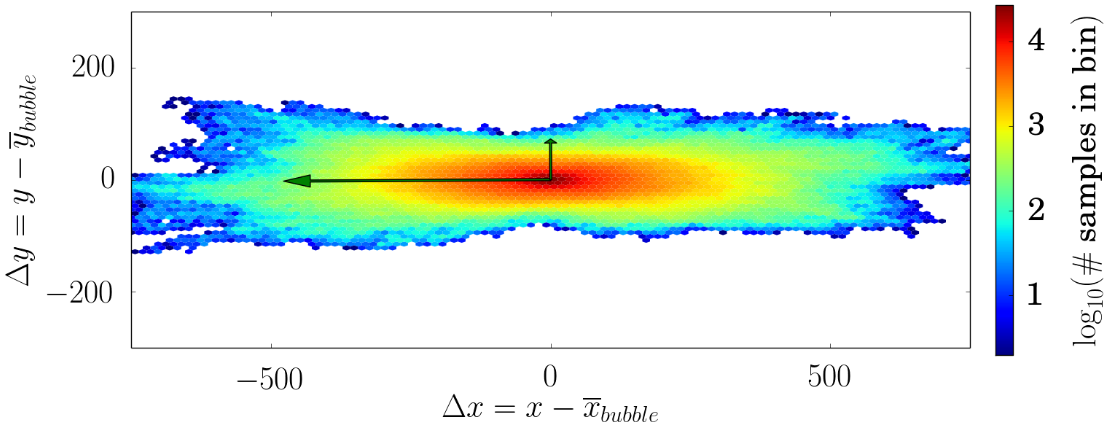


(b) $v = 0.2$.

FIGURE 11.24.: **Centered patch histogram for $v = 0.1$ and $v = 0.2$.** Both figures show patches of size $B < 1000$. As no mutant deviates by more than 200 from the corresponding patch center, the choice of $B_{critical}$ is sufficiently low. (a): The first principal component explains 96.36% of the variation and is aligned with the expansion axis ($m = 0.0038$). (b): The first principal component explains 96.05% of the variation and is aligned with the expansion axis ($m = 0.0075$).



(a) $v = 0.6$.



(b) $v = 1.0$.

FIGURE 11.25.: **Centered patch histogram for $v = 0.6$ and $v = 1.0$.** Both figures show patches of size $B < 1000$. As no mutant deviates by more than 200 from the corresponding patch center, the choice of $B_{critical}$ is sufficiently low. (a): The first principal component explains 91.17% of the variation and is aligned with the expansion axis ($m = -0.0036$). (b): The first principal component explains 86.79% of the variation and is aligned with the expansion axis ($m = 0.0056$).

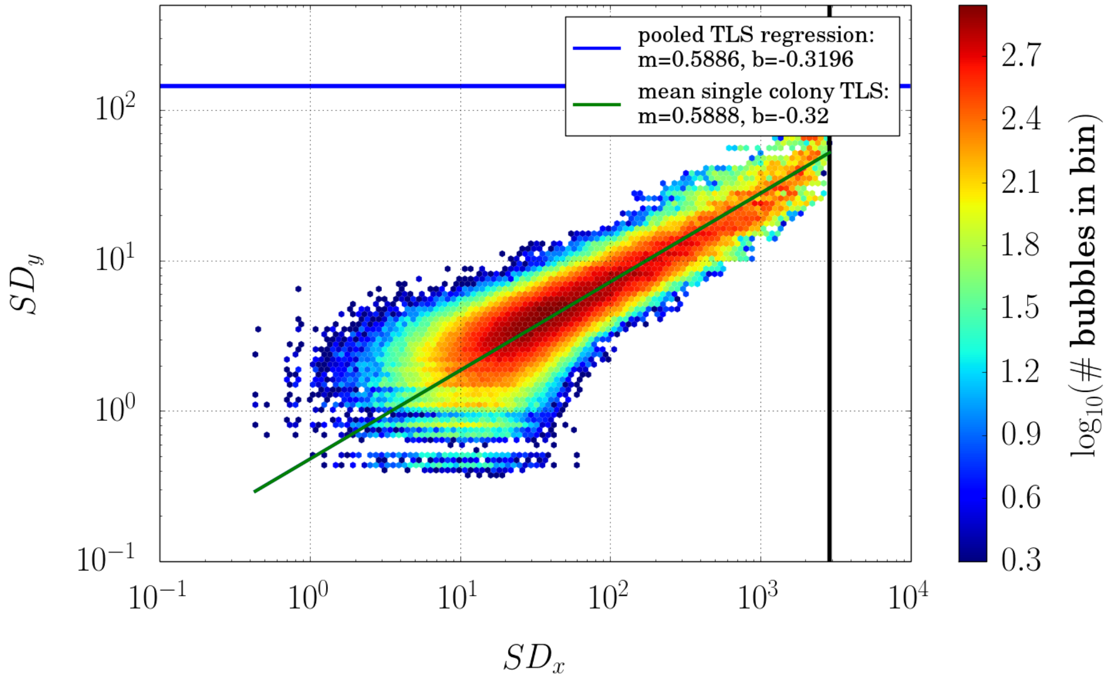


FIGURE 11.26.: **Histogram of patch shapes for $v = 0.1$.** The SD–values of 10 independent colonizations are shown. The patch shapes form a single cluster as expected. The mean over the regressions of all ten realizations is almost identical with the regression for the pooled data.

The patches shapes for other expansion velocities (see Figures B.7 to B.10) show the same general pattern, but the slopes of the total least squares regressions differ. Figure 11.29 shows the slope of the total least squares regressions.

11.2.3. The front–quenched coalescent

For boundary velocities clearly below the phenotypical limit, the data from the front–quenched coalescent shows the same general pattern as the unquenched coalescent presented in section 11.2: we observe a cluster of small patches apparently unaffected by the expansion and the characteristic power–law bar with slope $m \leq 0.5$.

The influence of the movement along the colonization paths is not apparent in the patch shapes. Figure 11.30 and Figures B.11 to B.12 show the results.

In contrast to the unquenched coalescent, we can now investigate the transition from the boundary–limitation to the phenotype–limitation. In fact, for boundary velocities $v > v_{\text{pheno}}$, we observe a cluster of patches that shows a pattern similar to one observed in the fully quenched coalescent.

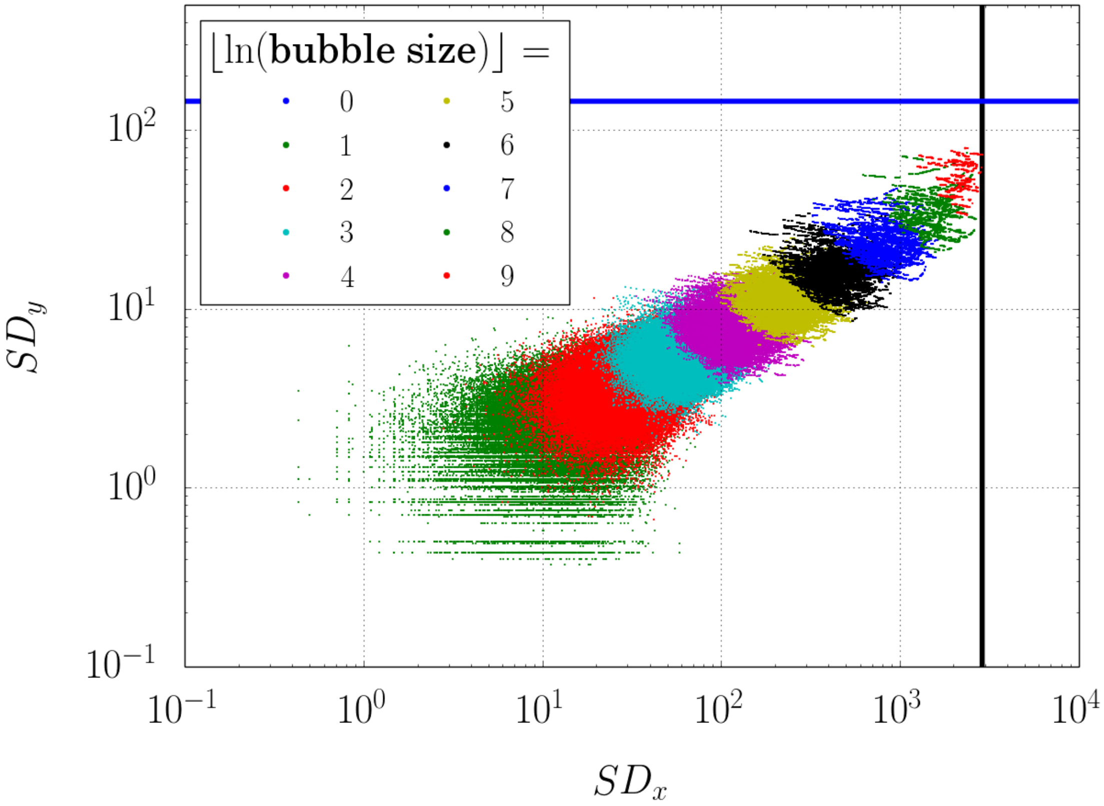


FIGURE 11.27.: **Bubble shapes for $v = 0.1$ colored according to the patch size.**

Figure 11.31 shows this pattern for a boundary velocity of $v = 1$ (clearly above v_{pheno}). In addition to the cluster of small patches, we see a power–law bar of slope $m \approx 0.66$ — an indicator for super–diffusive lineage movement. Figure B.13 shows that for a boundary velocity slightly below v_{pheno} the transition is not yet apparent.

The impact of the quenched front can be quantified by the frequency of lineage movements along the colonization paths. Figure 11.32 shows that, in fact, only a tiny fraction of the lineage moves were defined by the colonization paths. Figure 11.33, however, demonstrates that the fraction of forced moves at the front is significant especially for higher boundary velocities.

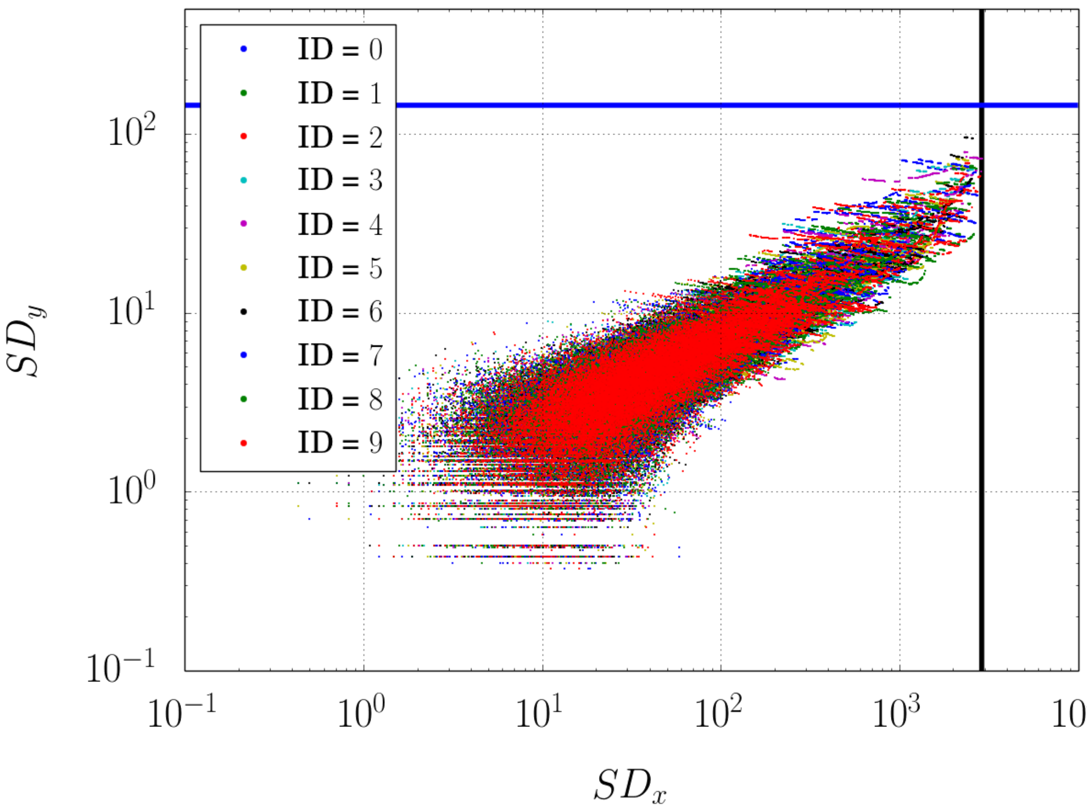


FIGURE 11.28.: **Bubble shapes for $v = 0.1$ colored according to the colonization.** The detailed colonization paths are different for every realization of the colonization process. These differences are apparent in the fine structure of the patch shapes for large patch sizes (upper right area of the cluster). However, the patch shapes do not differ significantly between the colonies.

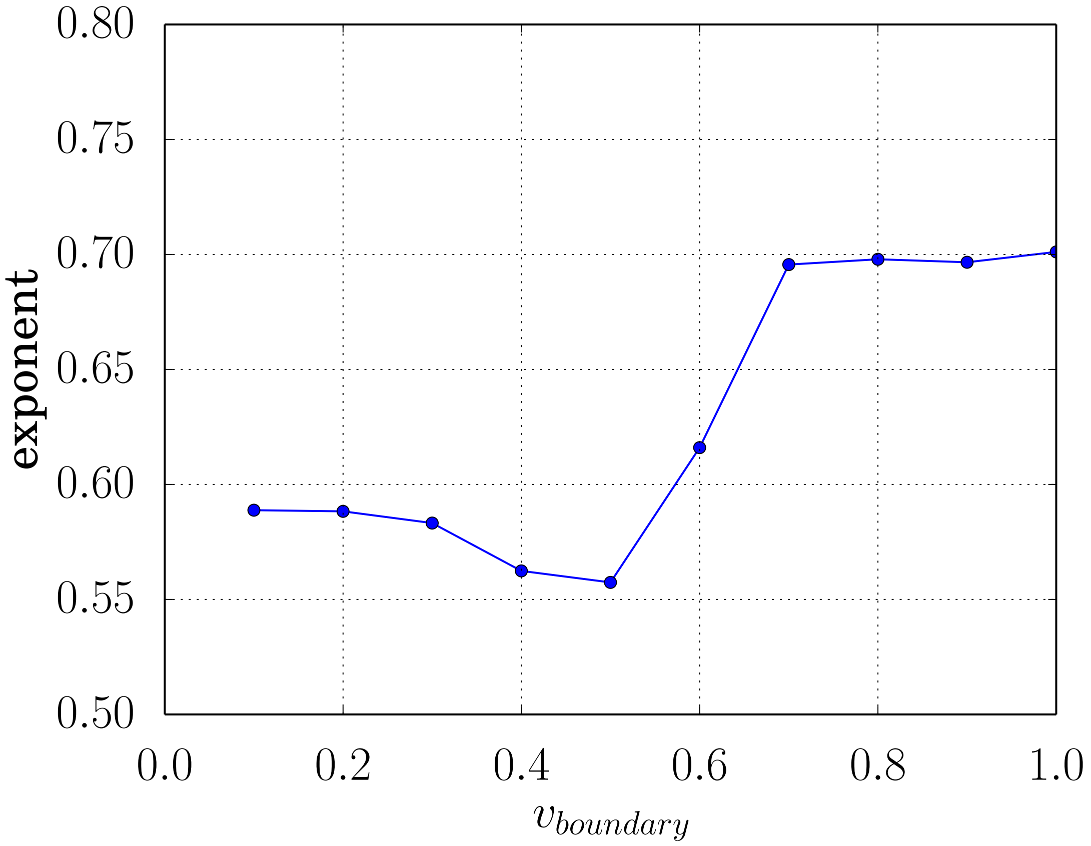


FIGURE 11.29.: **The slopes of the total least squares regressions** are shown for different boundary velocities. Each value corresponds to the mean of 10 independent colonies. The standard errors are all smaller than 0.001. Note that, as expected the function saturates above the phenotypical expansion velocity $v_{\text{pheno}} \approx 0.615$.

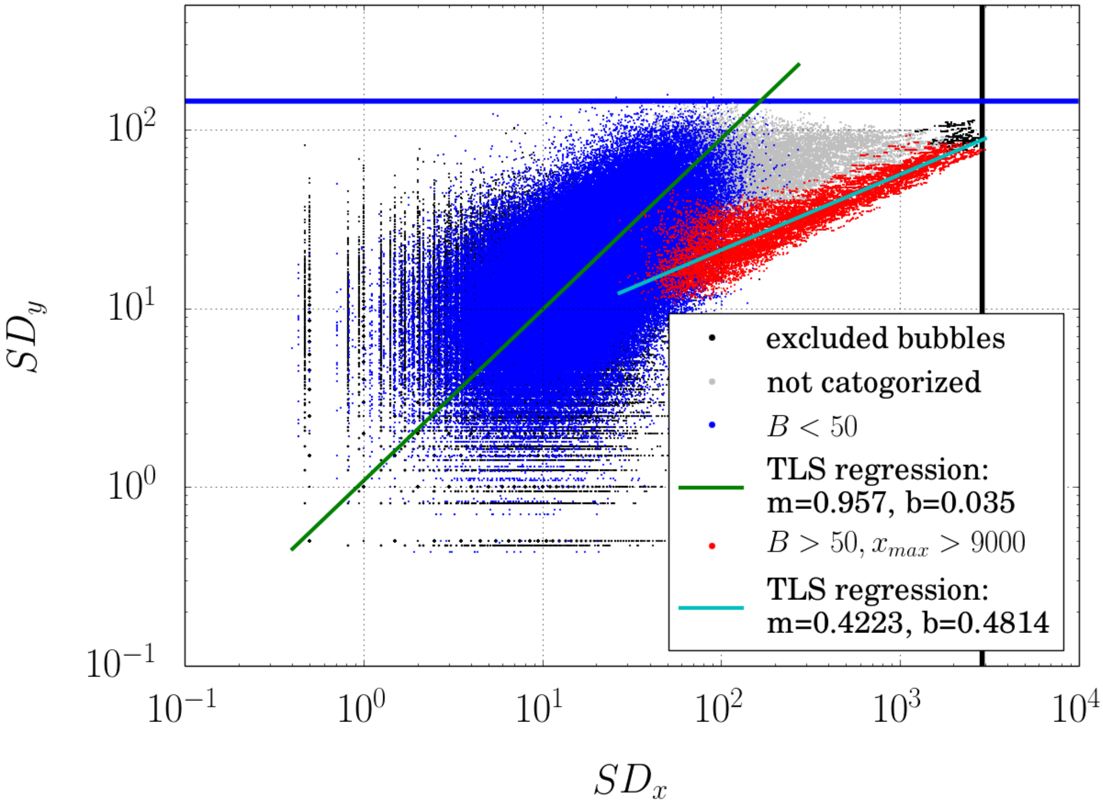


FIGURE 11.30.: **Bubble shapes for $v = 0.5$.** Pooled data from 10 independent realizations of the front-quenched coalescent is shown. Compared to the unquenched coalescent, some differences are apparent (see Figure 11.21(b)) but the general structure is identical.

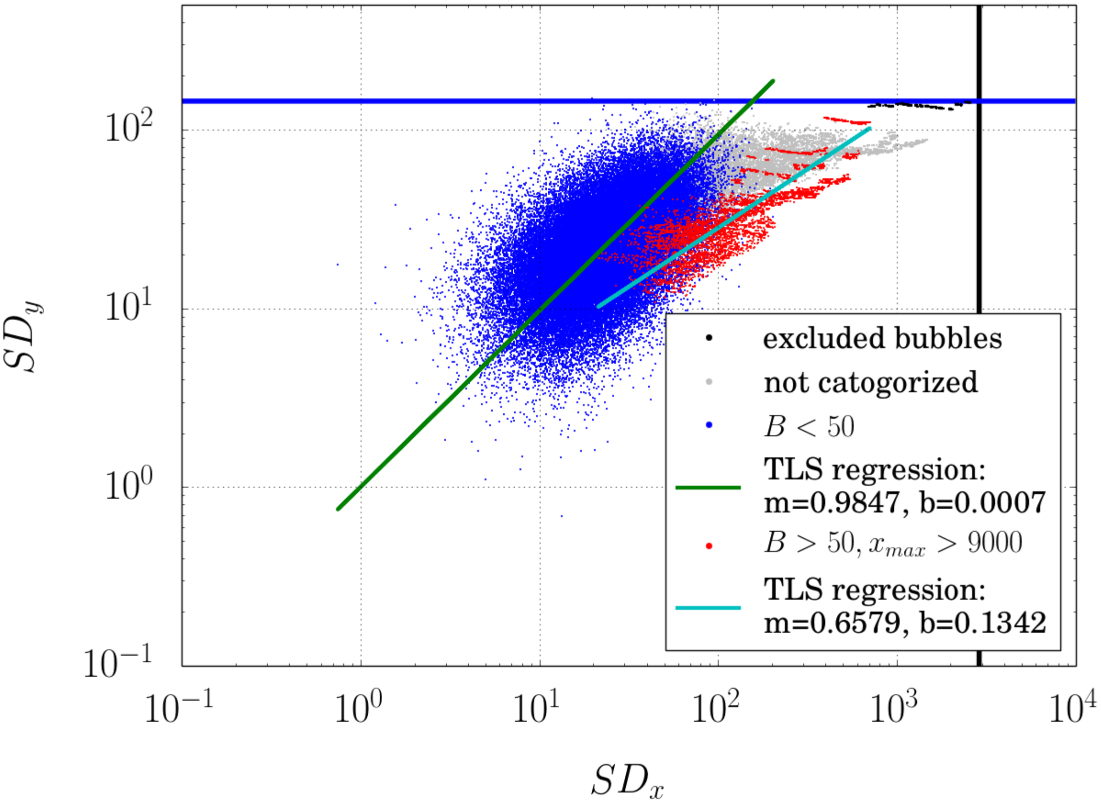


FIGURE 11.31.: **Bubble shapes for $v = 1.0$.** Pooled data from 5 independent realizations of the front-quenched coalescent is shown. Compared to the unquenched coalescent, some differences are apparent (see Figure 11.21(b)) but the general structure is identical.

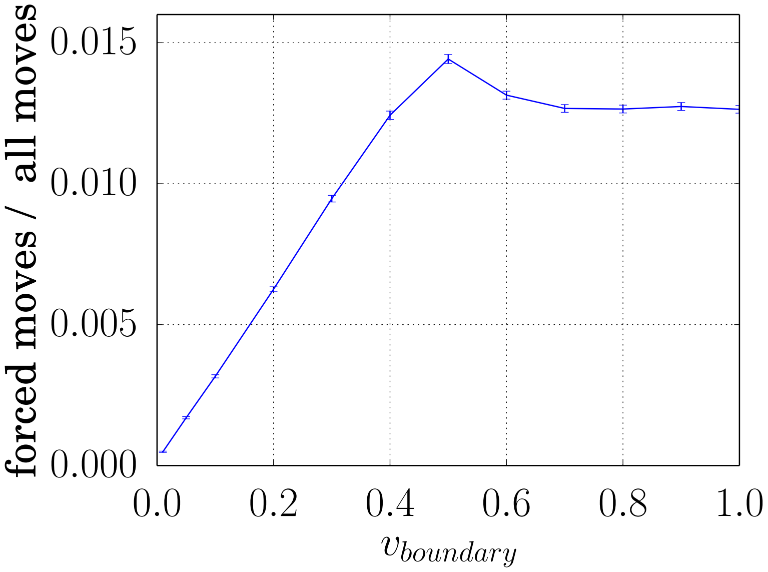


FIGURE 11.32.: **The frequency of forced moves** for lineages uniformly sampled from the entire habitat.

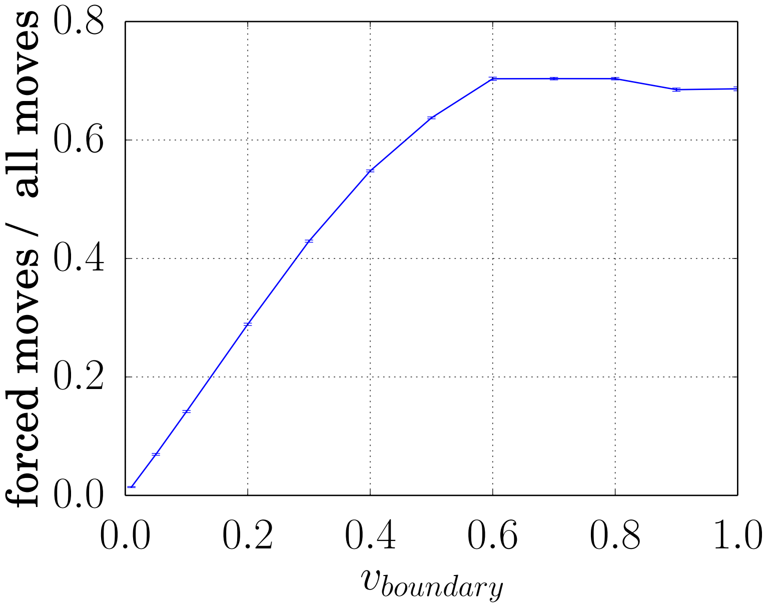


FIGURE 11.33.: **The frequency of forced moves** for lineages sampled along the population front.

12. Discussion

The identification and analysis of range expansions based on genetic data is a complex topic. Many parameters can influence the gene flow in a population, some can blur the genetic footprint of the underlying population dynamics. In this part of the thesis we developed tools that complement the existing methods and extend our capacity to read in the cornucopia of data we addressed in the introduction.

12.1. The unquenched coalescent

The *critical patch size* can be used to avoid a bias of the patch shapes due to the habitat shape and size. This step is the basis of the approach presented here. Filtering purely by means of the patch size works for the specific setup of our model but for data from real populations the filter has to be refined. For instance, natural populations almost never live on cylinders. If we replace the periodic boundary condition in our model by reflecting ones, patches can 'collide' with these boundaries irrespective of their size. In order to account for reflecting boundaries, we suggest to restrict the analysis to patches with $\overline{y}_{\text{patch}}$ sufficiently far away from the boundaries in addition to the size filter. All filters must be designed such that they do not introduce a filter bias.

The *principal component analysis* of the *centered patches* can be used to infer the expansion direction. As small patches do not confer much information and large patches tend to be biased and will therefore be excluded from the analysis, the parameters of the population define a window of patch sizes. The method presented here only works if this window provides a clear first principal component.

We have identified that *small patches* are typically *non–surfing mutations*. Note that a fraction of the small patches occurred at the expansion front, and an even smaller fraction probably surfed for a limited time. Such short time surfer mutation are likely to be impossible to identify and they are one reason why the slope of the total least squares regression is expected to be slightly smaller than 1. The second source of a potential downward bias of the slope is the limited habitat size. Even with a conservative choice, we can not fully exclude that no patches were biased by the limited habitat width.

The analysis of the *radii of gyration* (that is, the standard deviations of the mutants from the patch center) along the expansion axis and perpendicular to it proved to be a rich source of information. The radii of gyration can be used to qualify the impact of the colonization paths, to classify the patches according to their surfer history (see Figure 12.1). Note that the existence of a surfer group is a strong indicator of a range expansion. The *logarithmic compression factor* corresponds to the exponent of the apparent power–law of the group of surfing patches.

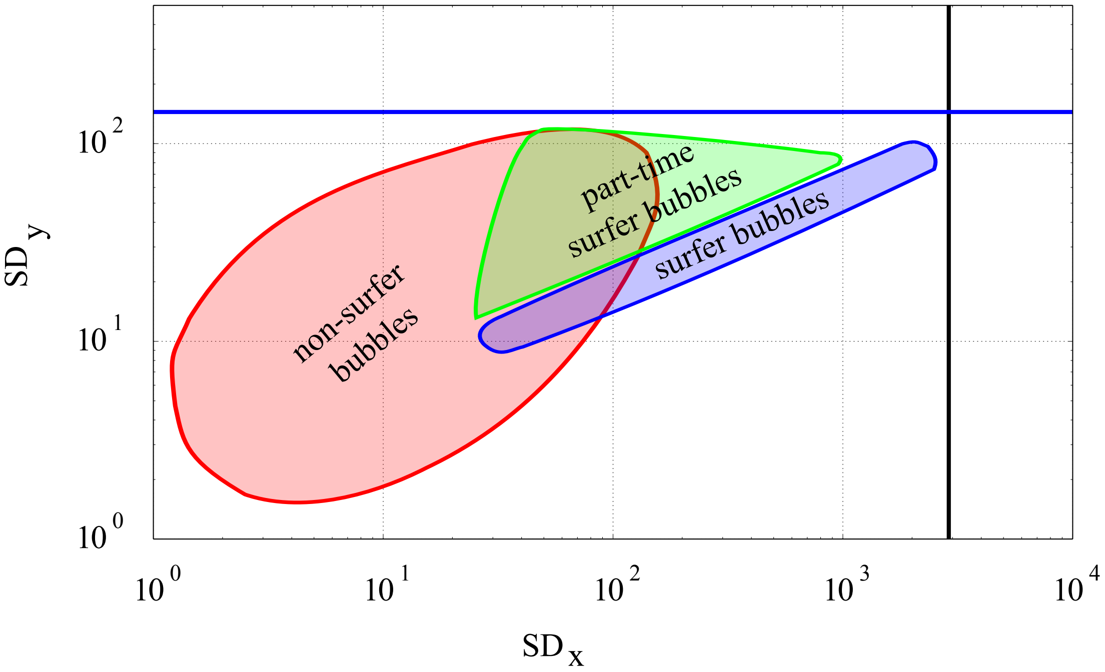


FIGURE 12.1.: **Sketch of the patch categories.** In the unquenched coalescent, each mutation falls into one of the following three categories: surfer patches correspond to mutation that surfed on the wave of advance until the end of the colonization. Part–time surfer patches correspond to mutations that surfed for some time and then lost contact to the front. Non–surfer patches correspond to mutations that did not surf. Although the categories are not mutually disjoint, they can be clearly distinguished in the gyration plots.

The analysis of the non–logarithmic *compression factor* of non–surfing patches can be used to identify spatial heterogeneities leading to non–isotropic migration. Non–surfing patches exist independent of the expansion velocity and the slope of the regression is largely unaffected by the expansion velocity. A strong deviation from a slope of 1 is a clear indicator of non–isotropic migration. We did not carry out an analysis of isotropic migration here, but the calculation of the compression factor is straightforward.

12.2. The fully quenched coalescent

The patch shapes in the fully quenched coalescent are strikingly different as compared to the two other versions. As all mutation surf, there is only one cluster of patches.

The slope of the cluster in the loglog–plot indicates that the lineage movement during fully–quenched phenotype–limited range expansion is super–diffusive. For boundary–limited range expansion, we expected diffusive lineage movement but the data does not allow a clear call.

The fully–quenched coalescent is expected in biofilms such as the microbial colonies mentioned in the introduction. Furthermore, high–density blocking is a potential reason for a fully–quenched scenario. In [100] several examples for high–density blocking between species are given: the first colonizers densely occupy a habitat and hinder the establishment and reproduction of secondary colonizers.

High–density blocking is based on the impossibility of sexual reproduction between the primary and the secondary colonizers. Thus, for asexual species an equivalent mechanism applies within the species. Plant species with vegetative propagation are a promising model for a study of fully quenched coalescence: vegetative propagation is a paradigmatic example of limited migration.

12.3. The front–quenched coalescent

The patch shapes of the front–quenched coalescent for boundary–limited range expansions are largely identical to the patch shapes of the unquenched coalescent. This is as expected, as only a tiny fraction of the lineage moves is given by the colonization paths (Figure 11.32).

For phenotype–limited range expansion the unquenched coalescent cannot provide corresponding data, by construction. The patch shapes (Figure 11.31), however, show striking differences as compared to the scenarios slightly below the phenotypical limit: the surfing patches form a cluster with compression factor of $c \approx 0.66$ that we already observed in the fully-quenched coalescent.

We did not address the coalescent of samples along the expansion front, as they cannot be treated in the framework of patch-shapes. The large fraction of forced moves for lineages sampled at the front (Figure 11.33) indicates similarities to the fully–quenched coalescent.

12.4. Smaller samples

The sample size chosen for our analysis is quite large, albeit not unrealistic. For a habitat with 5×10^6 demes we sampled at 5×10^4 randomly distributed positions (1% of all individuals were sampled). Especially in older experimental data sets, the sampling scheme will be usually less dense.

In order to estimate the impact of the sample size on the patch shape measurements, we have to briefly discuss the statistical framework of the mutant patches. For simplicity we assume that sampling mutants is equivalent to evaluating a random variable X with finite mean μ and variance $\mathrm{Var}(X)$.

The *uncorrected standard deviation of the sample*

$$SD = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x - \overline{x})^2} \tag{12.1}$$

underestimates the standard deviation of X as, by construction, $\overline{x}$ minimizes equation (12.1). Replacing $\overline{x}$ by μ in equation (12.1) would, thus, lead to a larger SD–value.

When we compare the standard deviations in the context of the mutant patches, we must expect a similar result: in average, the larger sample mean will give a better approximation of the true patch mean than the smaller sample. Consequently, the standard deviation of the smaller sample will exhibit a stronger bias.

This problem can be easily solved by replacing the uncorrected standard deviation by the usual corrected standard deviation

$$SD_{\mathrm{corr}} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x - \overline{x})^2}. \tag{12.2}$$

Note that, the bias is only significant for small patches and the larger patches contain the most valuable information. Furthermore, the two components of the tuples (SD_x, SD_y), our analysis is primarily based on, are both biased in the same way. The impact on the compression factor is, thus, even less relevant.

We conclude that, with the corrected standard deviation, smaller samples will give less accurate results with the same expectations.

12.5. Conclusions

Based on the results from our coalescent models, we propose to analyze the spatial distribution of neutral mutations in two–dimensional habitats with limited migration in data sets from natural populations. We developed methods on the basis of simulation data and are aware of the fact that they cannot be adopted directly in experiments.

The general workflow for the analysis is the following: first, determine the critical patch size by creating plots such as in Figures 11.4. (B_{critical} is not sharply defined and must be chosen according to the quality of the data and especially according to the habitat details.) Second, estimate the expansion direction by performing a principal component analysis on the centered patches such as presented in Figure 11.5. Third, plot the patch shapes and analyze apparent clusters: identify possible non-isotropic migration based on the cluster of non–surfing patches, identify clusters of surfer patches and deduce the impact of the colonization history.

We restricted our attention to a single habitat size, namely $[10000 \times 500]$. It might seem natural to test the impact of the habitat size by comparing the results from different habitat configurations and we will consider this in a future publication. However, we are convinced that the filters developed here effectively purged the impact of the habitat size and no significant differences will be observed in larger habitats. Still, a potential impact resides in the grain size of the lattice and we will quantify its impact for the publication of the project.

As Figure B.1 and B.2 indicate, the shape of the moving boundary remains a major topic. The profile used in Figure B.2 could be found in very similar form in population that expand, or instance, along a valley with higher temperature or better nutrient supply in the center. In such a situation, we expect the fixation of neutral alleles that happen to start in the center and profit from a spatial advantage.

Many boundary–limited range expansions will have irregular boundary shapes but the shape will not be stationary — due to spatial heterogeneities, for instance. The introduction to a noisy moving habitat boundary is therefore a promising idea that we did not yet incorporate.

12.5.1. Hitch–hiker patches

So far, we considered only neutral loci and assumed the adaptation processes in the population do not influence the patch shapes. However, recurrent selective sweeps can indeed affect patterns of genetic diversity [5] and shift the neutral lineages towards the sweep origin.

We did not incorporate selection to our model but even if neutral lineages are shifted by recurrent selective sweeps, the patch shapes should be largely unbiased. The methods introduced here can help to distinguish range expansions from recurrent sweeps. Nevertheless, this question requires further investigation.

12.5.2. A generalized model

The front–quenched coalescent model we presented here represents a scenario in which the initial colonization of a deme is easier than taking over a already colonized deme. The front–quenched coalescent is, of course, not the only way one can model such a situation.

The concept of high–density blocking [100] motivates the following modified rules for a colonization: All colonized demes remain active for the entire colonization process. Colonization attempts are successful with probability p_1 if the target deme is empty and succeed with probability $p_2 < p_1$ if the target deme is occupied.

As in the models presented so far, it is not necessary to simulate the full forward population dynamics. Instead, we redefine the coalescent according to the forward model: In the bulk of the population, the lineages move with probability p_{bulk} to one of the neighboring demes, at the front the lineages move with probability p_{front} to the parental deme.

For simplicity, set $p_{\text{front}} = 1$. Then $p_{\text{bulk}} = 0$ is equivalent to the fully quenched coalescent and $p_{\text{bulk}} = 1/4$ is equivalent to the front–quenched coalescent of section 10.2.2. Choosing $p_{\text{front}} = p_{\text{bulk}}$ neglects the colonization paths and if the population front is flat, we end up with the unquenched coalescent of section 10.2.3.

For this version of the coalescent, we suggest the term *p–quenched coalescent*. By tuning the two probabilities, it would be possible to define a *'quench–factor'* for the coalescent.

Outlook

In this thesis, we modeled populations under different scenarios of range expansions in order to increase the understanding of the impact of range expansions on the neutral genetic diversity.

In part I, we introduced the concept of boundary–limited range expansions and established the distinction towards phenotype–limited range expansions. We were able to demonstrate that these two types can have quite different consequences on neutral genetic diversity.

Populations that track a relatively slow moving boundary during, for instance, a scenario of climate change can maintain high population density up to the expansion front. In this case, the loss of genetic diversity typically associated with range expansions can be drastically reduced and relatively high levels of diversity are expected even close to the expansion front.

Our observations have direct implications for experimental applications: for instance, range expansions are commonly identified on the basis of clines in genetic diversity. These clines are much harder and sometimes impossible to detect in boundary–limited range expansions. Based on our results, it is possible to assess the appropriate sampling scheme for the detection of a boundary–limited range expansion or to assert that such a sampling scheme does not exist.

In part II, we generalized the concept of the boundary–limitation to two–dimensional habitats. Our linear habitat model is, to a certain extent, equipped to deal with two-dimensional habitats but it describes only the clines of genetic diversity along the expansion axis. Therefore, we addressed the two–dimensional patterns of diversity in the context of a new model.

In two spatial dimensions, new phenomena arise. The well–mixed population front of the linear habitat is replaced by a front line. This front line is not only spatially structured (topologically one–dimensional) but can also be rough: some regions of the front will expand faster than others, some patches remain temporally unpopulated albeit accessible.

The boundary limitation does not only influence the population density at the front but it can limit the roughness of the front, determine the front shape or introduce an external source of noise to the process.

The roughness of the front can have a crucial influence on the lineage movement and on the coalescent as we demonstrated with the *fully quenched* and the *front–quenched* coalescent: lineages move super–diffusive in phenotype–limited and diffusive in boundary–limited scenarios. So far, we were not able to fully understand the details of the transition between diffusion and super–diffusion and this problem requires further investigation.

Based on the comparison between the *fully quenched*, the *front–quenched* and the *unquenched* coalescent we tried to quantify the impact of the detailed colonization paths on the patterns of neutral genetic diversity. Drastic differences are apparent between the fully quenched coalescent as compared to the other versions. However, we were not yet able to establish the transition. The p–quenched coalescent proposed in section 12.5.2 is a promising concept to close that gap.

With the methods and observables developed and presented in this thesis, we aim at providing tools that are applicable in experiments. So far, no other publications made direct use of the concept of boundary–limited range expansion published in 2013 but we are confident that this will change soon. With the analysis of the spatial distribution of alleles we have provided guidelines for a new perspective on genetic data. After all, it is often the representation of data that permits deeper understanding.

A. Appendix to part I

A.1. Supplementary Figures

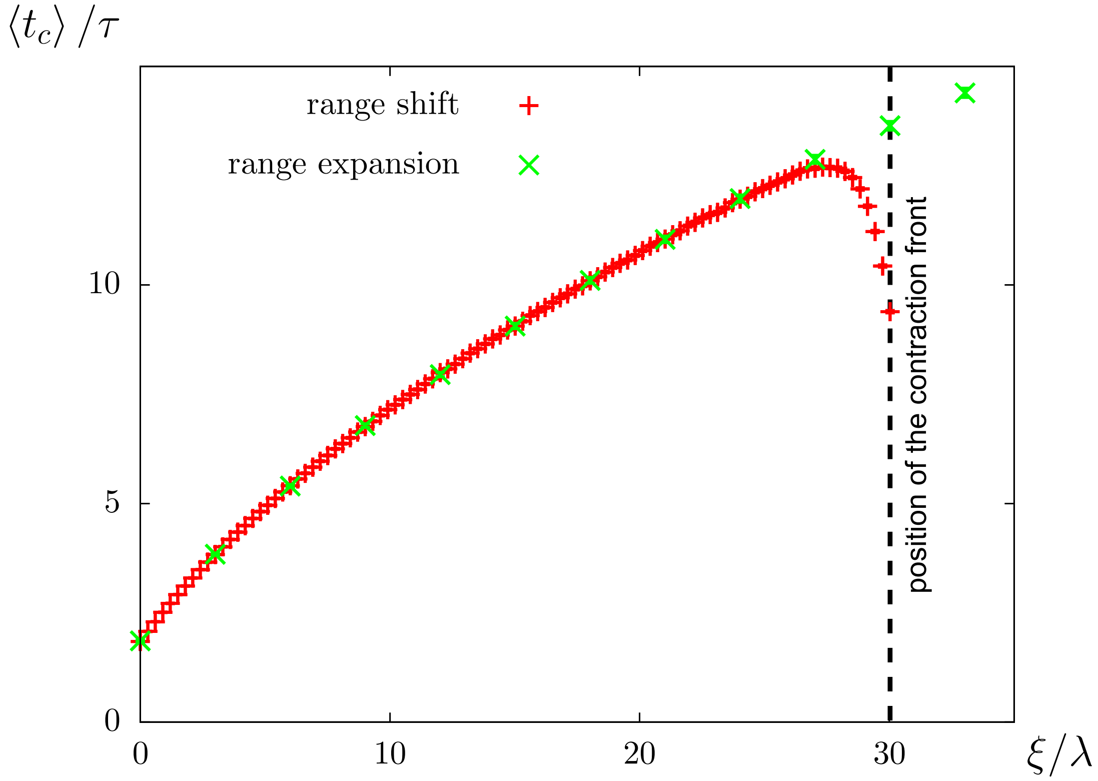


FIGURE A.1.: **Comparison of mean coalescence times (within deme sampling) between the range expansion model and the range shift model.** The habitat of the range shift model is extended from 0 to 30λ while the habitat of the range expansion model is extended from 0 to infinity. All boundaries move at the same velocity and the parameters of the expansion are as in Figure 4.1. The distance in which the contraction front has a significant influence on the coalescence process is on the order of λ. The Figure is from the SI of [78].

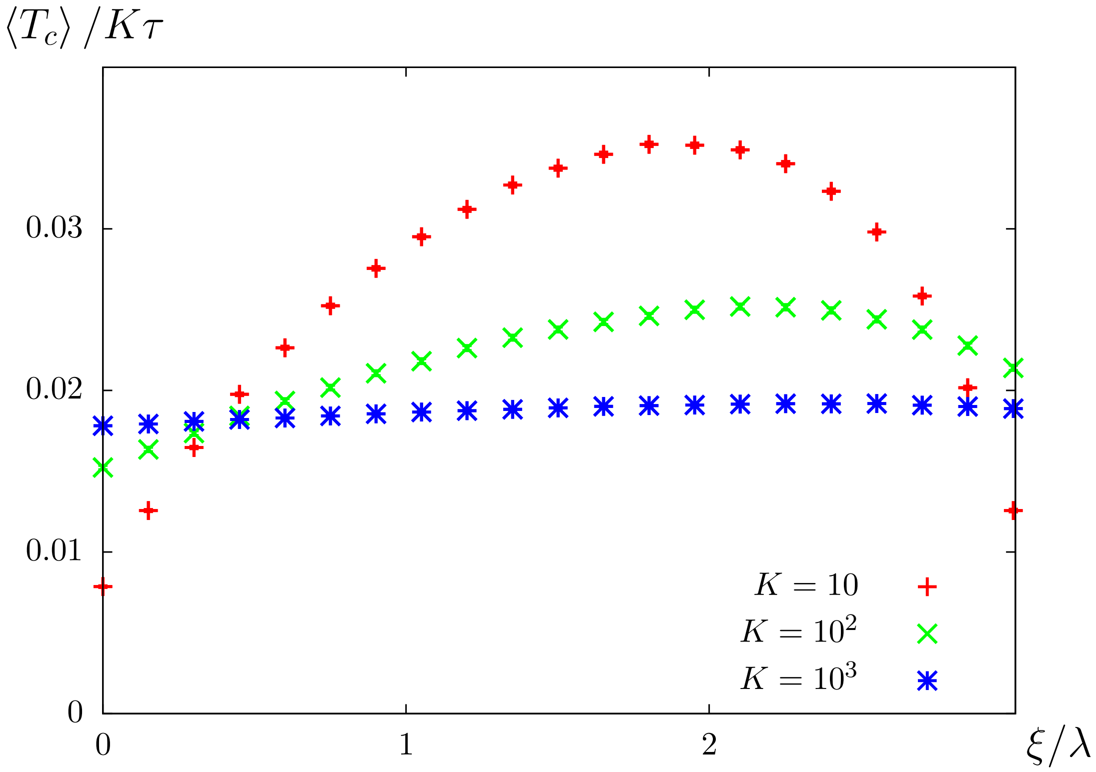


FIGURE A.2.: **Range shifts.** Mean coalescence times (scaled by the deme size K) are shown for a habitat of length $l = 100$ undergoing a range shift with velocity $v = 10^{-2}$. Note that as we increase deme sizes, the mean coalescence time becomes independent on the sampling location. Other parameters were $m = 0.33$ and $v = 10^{-2}$. The Figure is from the SI of [78].

A.2. Simulation code

The simulation code is written in Python 2.7.3 (`http://www.python.org/`), the Plots2 were created with Gnuplot [103], the Figures were created with Inkscape 0.48 [26].

The coalescence process in the linear stepping stone habitat The core of the coalescence simulation in the linear stepping stone habitat is based on two random walks with transition probability $m \leq 0.5$ to the left and right on $[0, l]_{\mathbb{Z}}$ with constant drift of one deme every ex generations. Coalescence occurs with probability K^{-1} if the lineage positions are identical at the end of a generation. These parameters are passed to the coalescence function along with the initial positions x and y. The function returns the coalescence time of a single realization.

```
def coalescencefunction(x, y, l, ex, m, K):
    t = 0
    coalescence = 0

    while not coalescence:
        qx = random.random()
        if x == 0:
            if qx < m:
                x = 1
        elif x == l:
            if qx < m:
                x = l-1
        else:
            if qx <  m:
                x += 1
            elif m < qx < 2*m:
                    x -= 1
        qy = random.random()
        if y == 0:
            if qy < m:
                y = 1
        elif y == xMax:
            if qy < m:
                y = l-1
        else:
            if qy <  p:
                y += 1
            elif m < qy < 2*m:
                y -= 1
            # end of a single generation
            t += 1
            if t % ex == 0:
                if x > 0:
                    x -= 1
                if y > 0:
                    y -= 1
            # coalescence:
            if x == y:
                qCoal = random.random()
                if qCoal < (1.0 / K):
                    coalescence = 1
    return t
```

When considering a pure range expansion we simply omitted the boundary condition at $x = l$.

The coalescent in the 2D continuous habitat model For the continuous model, we consider a habitat $[0, \infty] \times [-k/2, k/2]$, and pass sampling sites $x1 = [x_1, y_1]$ and $x2 = [x_2, y_2]$ and a contact distance δ.

```
def contcoalfunct(k, x1, x2, delta, KtwoD, v, sigmax):
    RW1 = x1[:]
    RW2 = x2[:]
    coalescence = False
    t = 0
    while Koaleszenz == 0:
        RW1 = GaussStep(RW1, sigmax, k, v)
        RW2 = GaussStep(RW2, sigmax, k, v)
        # end of a single generation
        t += 1
        dist1 = RW1[1]-RW2[1]
        dist2 = RW1[1]-RW2[1]+2*k
        dist3 = RW1[1]-RW2[1]-2*k
        disty = min(dist1, dist2, dist3)
        # periodic boundary conditions considered
        distance = math.sqrt((RW1[0]-RW2[0])**2 + disty**2)
        if distance<delta:
            q = random.random()
            if q < 1./KtwoD:
                coalescence = True
    return t
```

The function *GaussStep* moves a lineage from $[x_i, y_i]$ according to a two dimensional Gaussian distribution with mean $[x_i, y_i]$ and standard deviations $[\sigma_x, \sigma_x]$ conditioned on jumping to a place inside the habitat of the next generation (backwards in time).

B. Appendix to part II

B.1. The phenotypical expansion velocity

In the Eden model, the roughness of the population front increases and saturates after a while after the onset of the colonization from a flat initial colony.

The phenotypical expansion velocity in our version of the Eden growth model depends on the proportion of successful colonization attempts. This proportion is given by the average number of free neighbors per active deme divided by the number of neighbors.

Assume that after the burn–in phase an average of k active demes exist and the average number of free neighbors is p_k in a habitat of width W. Then each colonization attempt has a probability of $p_k/4$ to colonize a new deme.

For k active demes, each generation consists of $\approx 1/k$ colonization attempts. Thus, in average, $kp_k/4$ demes are colonized per generation and the mean expansion velocity is

$$v_{\text{pheno}} = \frac{k}{W}\frac{p_k}{4}. \tag{B.1}$$

The term k/W is a posible measure for the roughness of the population front in our model.

B.2. The shape of the bubbles

B.2.1. The mean in a periodic interval

Defining the mean in a periodic interval is not straightforward and sometimes not even well–defined. For instance, consider two points at $y = 0$ and $y = W/2$ in $[0, W)_{\mathbb{R}}$ with periodic boundaries. Both, $y = W/4$ and $y = 3W/4$ are possible choices for $\overline{y}$.

We use the following algorithm to calculate the bubble mean perpendicular to the expansion axis:

- Map the interval $[0, W)_{\mathbb{R}}$ onto the unit circle in $\mathbb{R}^2$.
- Identify each y–value with the corresponding unit vector in $\mathbb{R}^2$.
- Calculate the vector mean $\overline{\mathbf{v}} = B^{-1} \sum_y \mathbf{v}_y$.
- Normalize $\overline{\mathbf{v}}$ and map it back onto $[0, W)_{\mathbb{R}}$.

Unless $\overline{\mathbf{v}} = \mathbf{0}$ this procedure yields a well–defined mean. Note that $\overline{\mathbf{v}} \approx \mathbf{0}$ occurs only if the mutant samples are approximately uniformly distributed perpendicular to the expansion axis. Such bubbles are not used in our analysis and we therefore neglect this case.

B.2.2. Possible values of SD$_x$ and SD$_y$

Estimation of the standard deviation Estimating the standard deviation of distribution based on a sample is not trivial. Depending on the distribution, different formulas must be used to obtain an unbiased estimator. Here, we do not aim at describing the the standard deviation itself but the ratio of the standard deviations along two coordinate axes. Therefore, we neglect the statistical details and use the uncorrected sample standard deviation

$$\mathrm{SD} = \sqrt{\frac{1}{B} \sum_{i=1}^{B} (x_i - \overline{x})^2}, \tag{B.2}$$

where B refers to the number of mutants in the bubble.

B.2.3. Supplementary Figures

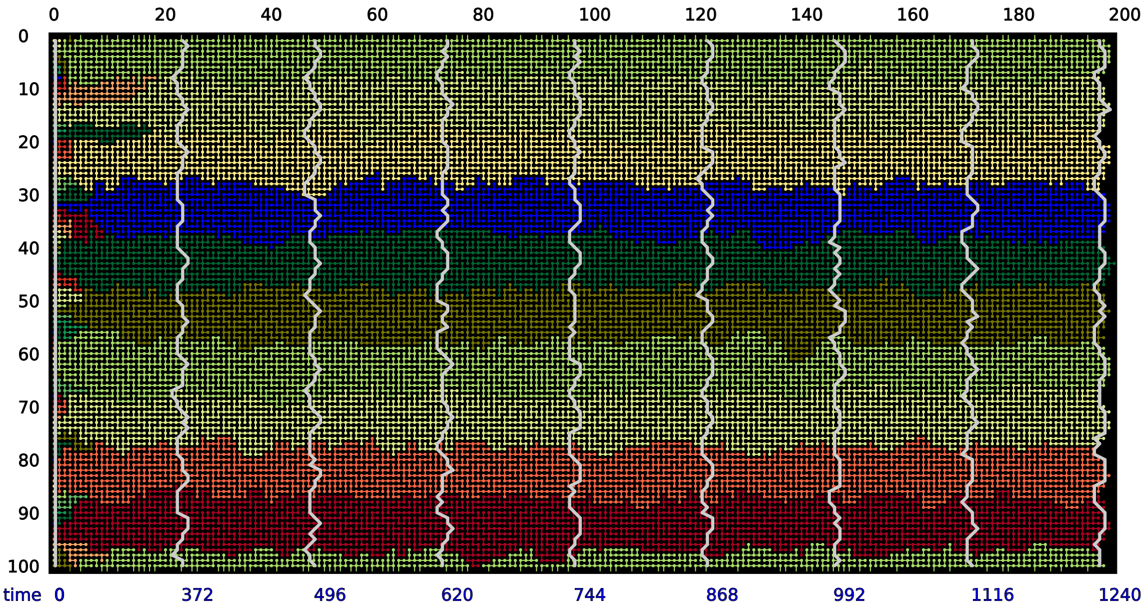


FIGURE B.1.: **Sectoring in a boundary–limited range expansion with sinusoidal boundary shape.** In this example we imposed a boundary with sinusoidal variation of the boundary position. The light grey lines indicate the position of the expansion front at the time indicated at the lower horizontal axis. The boundary stabilizes a sectors for each period of the sine function.

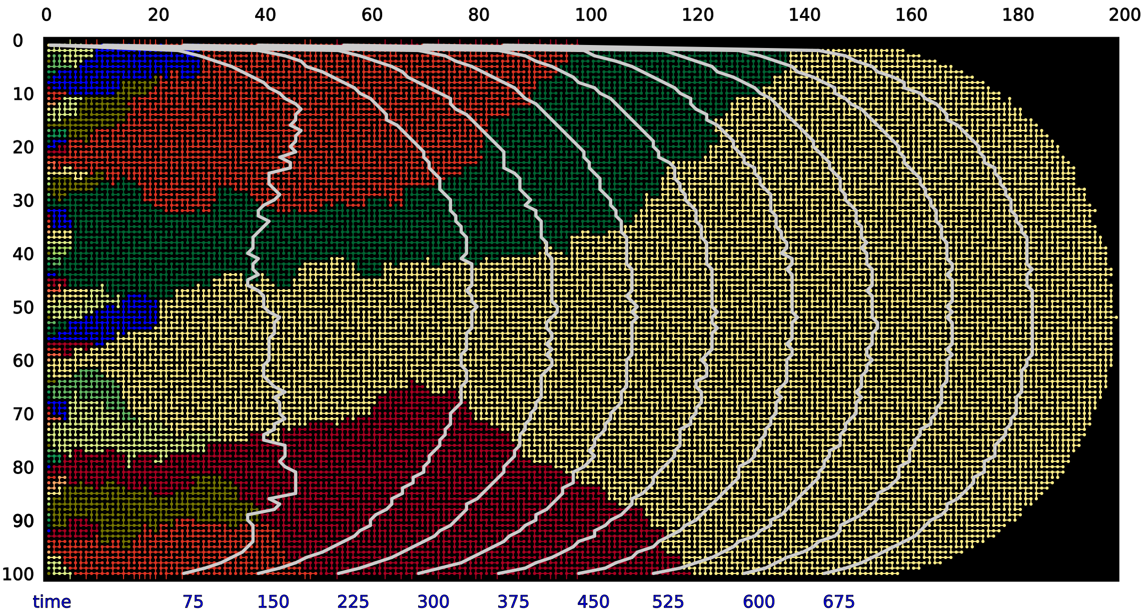
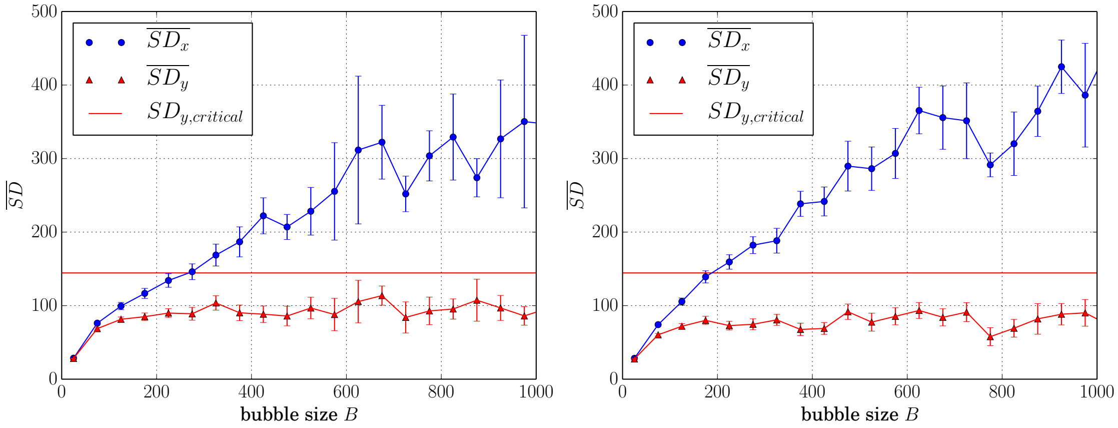
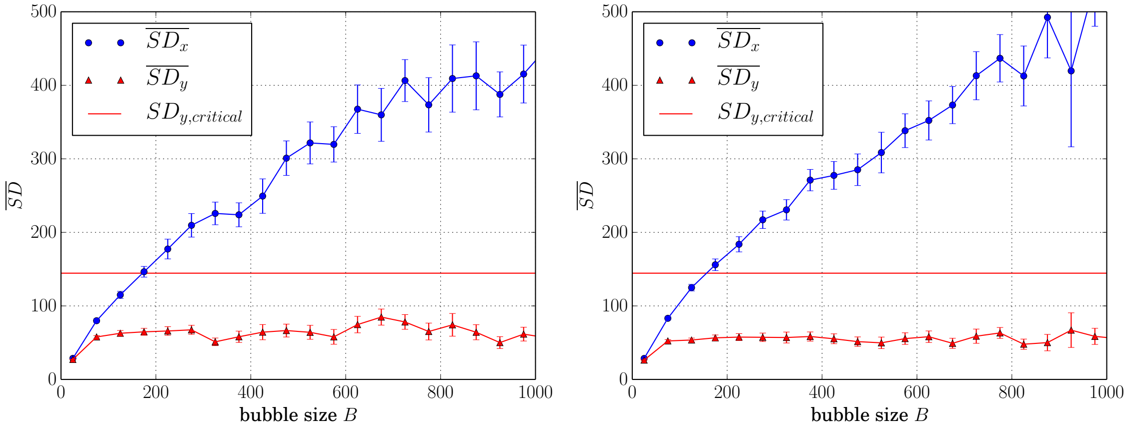


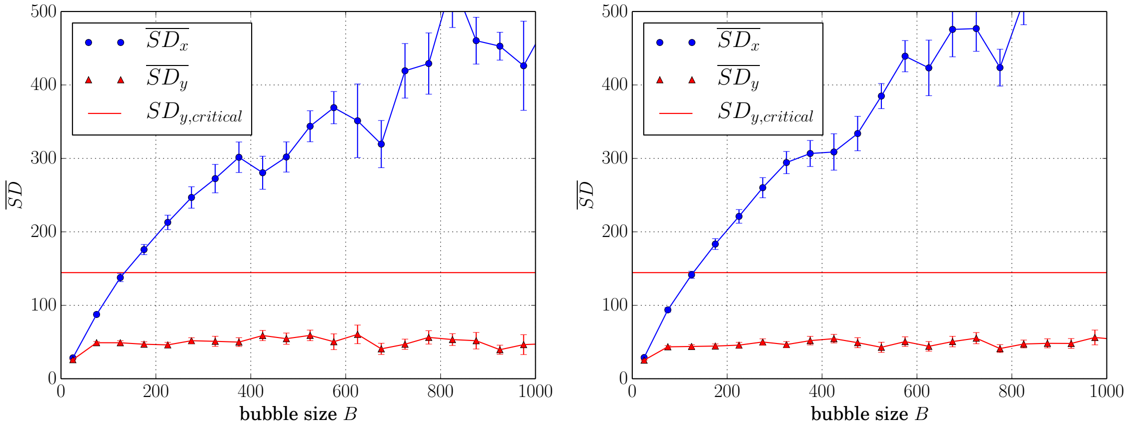
FIGURE B.2.: **Sectoring in a boundary–limited range expansion with rounded boundary shape.** In this example we imposed a boundary with rounded shape. The light grey lines indicate the position of the expansion front the time indicated at the lower horizontal axis. The boundary shape enforces the tip of the expansion front around $y = 50$ and confers an advantage to the genotype that happens to be there.



(a) $v = 0.1$. (b) $v = 0.2$.

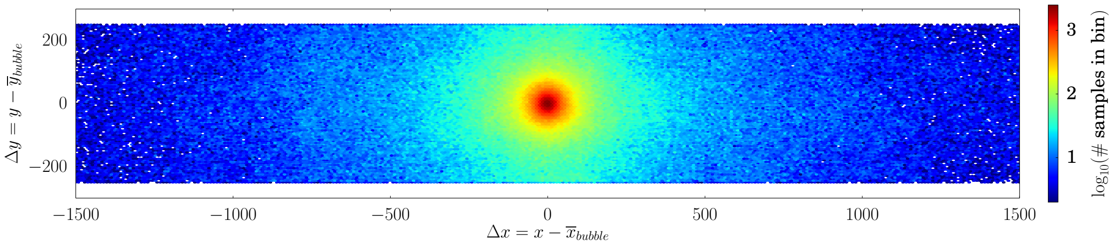


(c) $v = 0.3$. (d) $v = 0.4$.

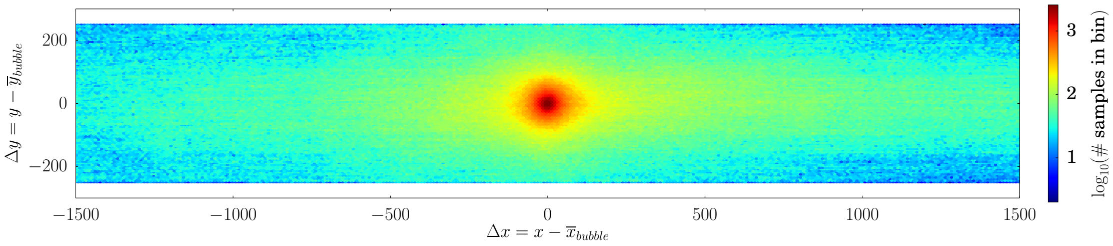


(e) $v = 0.5$. (f) $v = 0.6$.

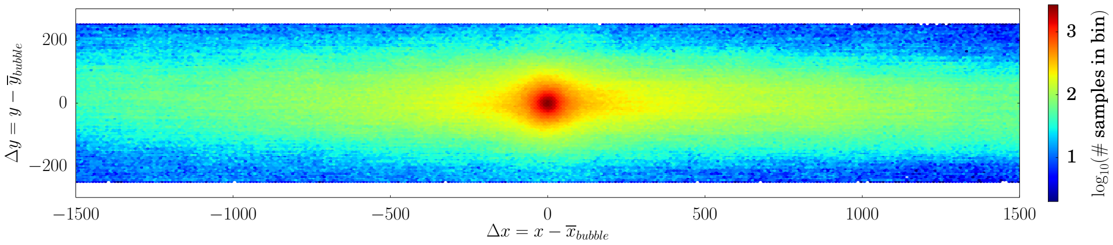
FIGURE B.3.: **Mean bubble dimensions.** For a habitat of length L (x–axis) and width W (y–axis) and the coalescent without impact of the colonization path, the average width and length of mutation bubbles is shown as a function of the bubble size B. The bubble sizes were binned, averages are taken over bubbles from the same size bin. See Figure 11.2 for a more detailed description.



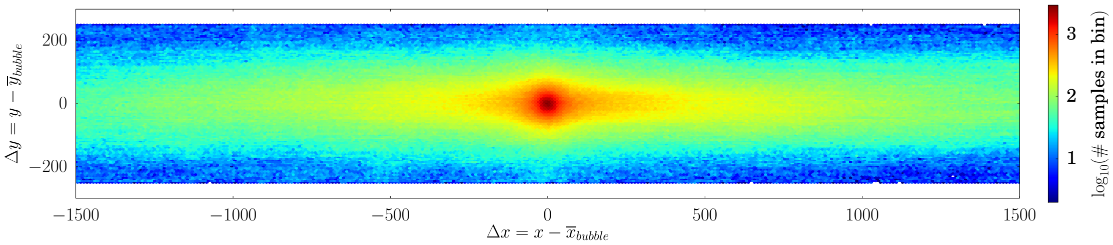
(a) $v = 0.0$.



(b) $v = 0.1$.

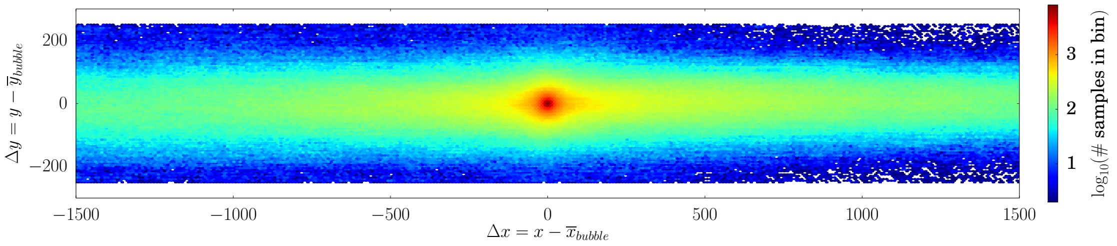


(c) $v = 0.2$.

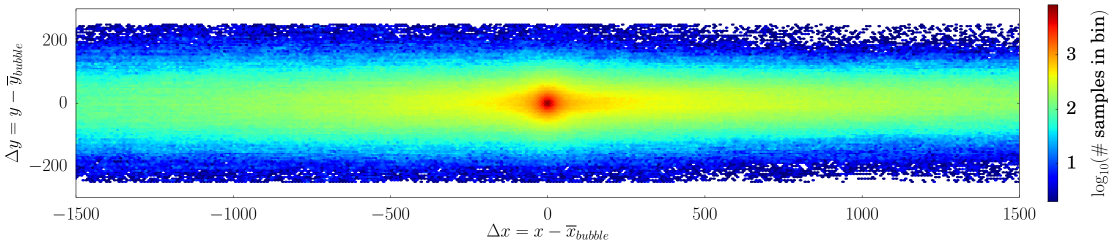


(d) $v = 0.3$.

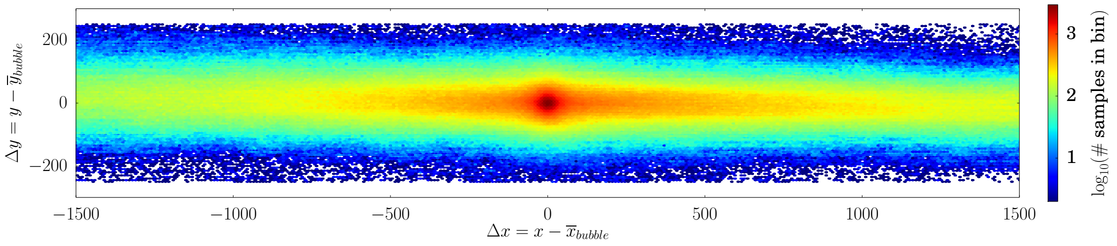
FIGURE B.4.: **Centered bubbles for different expansion velocities.**



(a) $v = 0.4$.

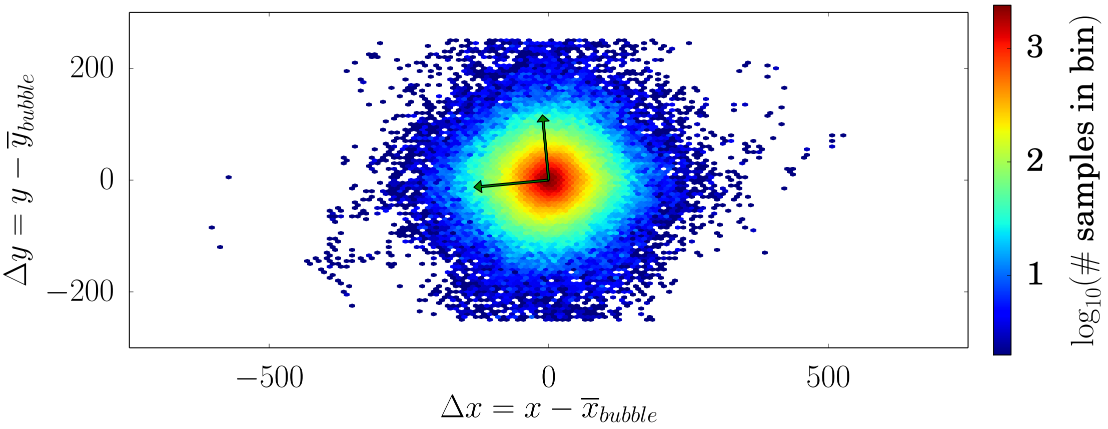


(b) $v = 0.5$.

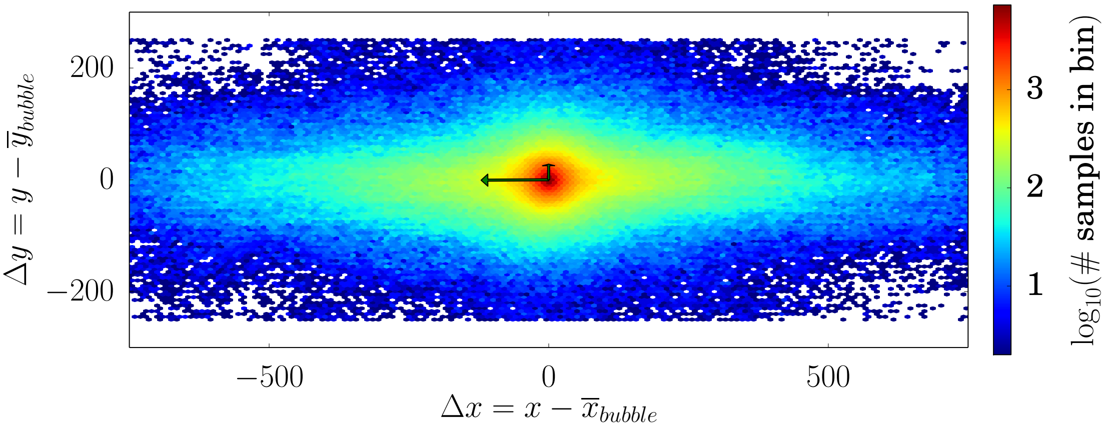


(c) $v = 0.6$. default

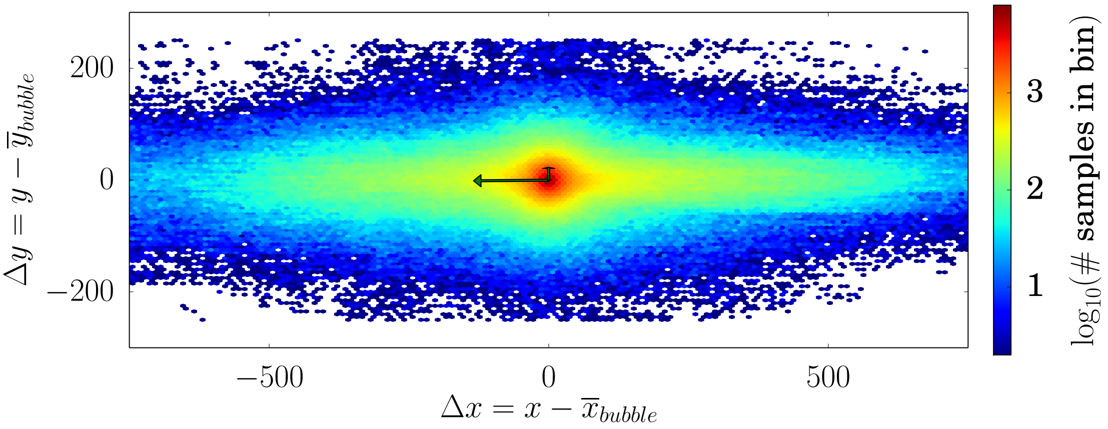
FIGURE B.5.: **Centered bubbles for different expansion velocities.**



(a) $v = 0.2,\ B_{critical} = 100$.

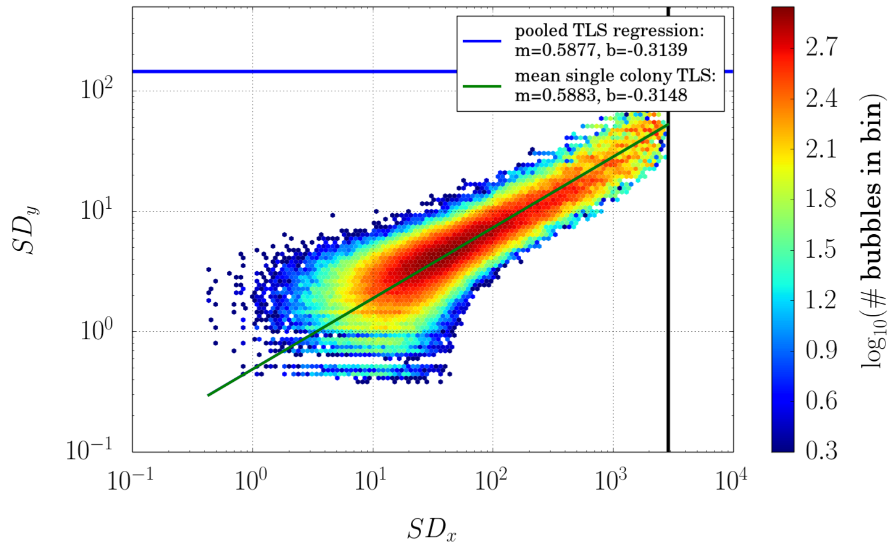


(b) $v = 0.4,\ B_{critical} = 1000$.

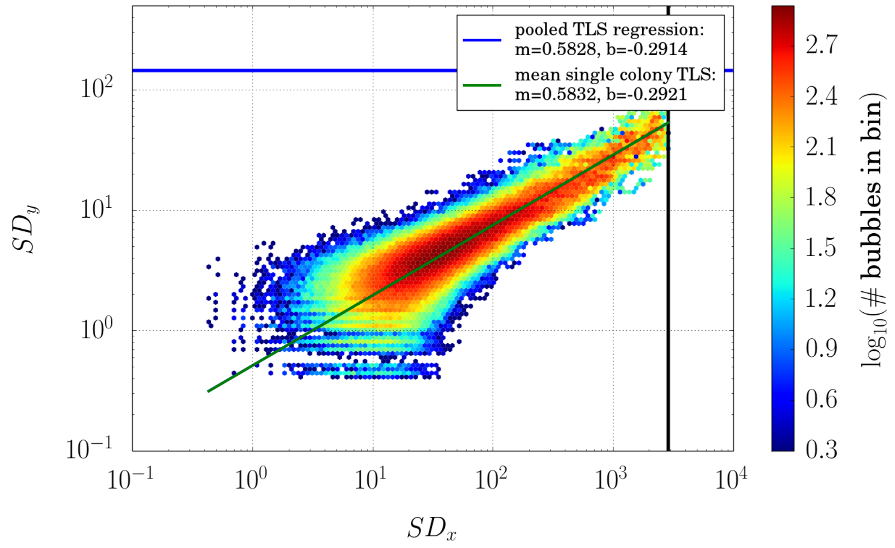


(c) $v = 0.5,\ B_{critical} = 1000$.

FIGURE B.6.: **Histograms of centered bubbles for different expansion velocities.** The critical bubble size is chosen according to the observation in Figure 11.4.

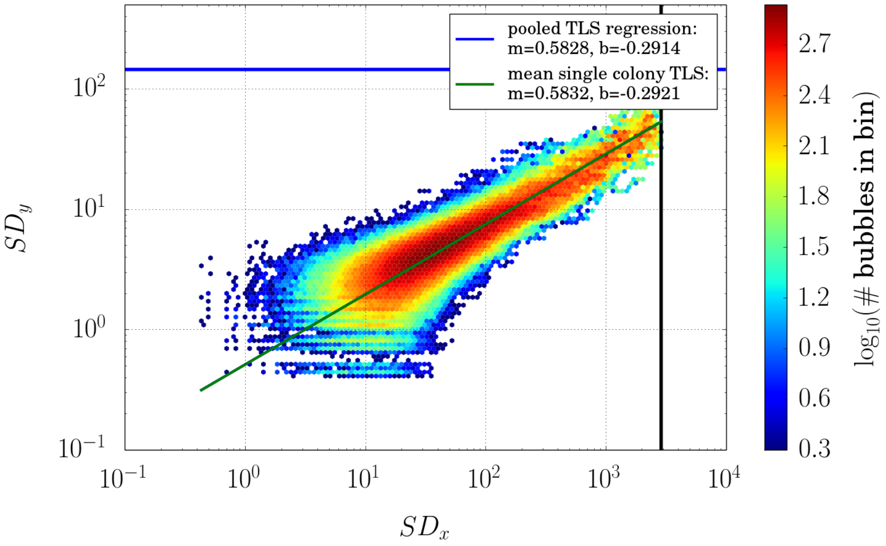


(a) $v = 0.2$.

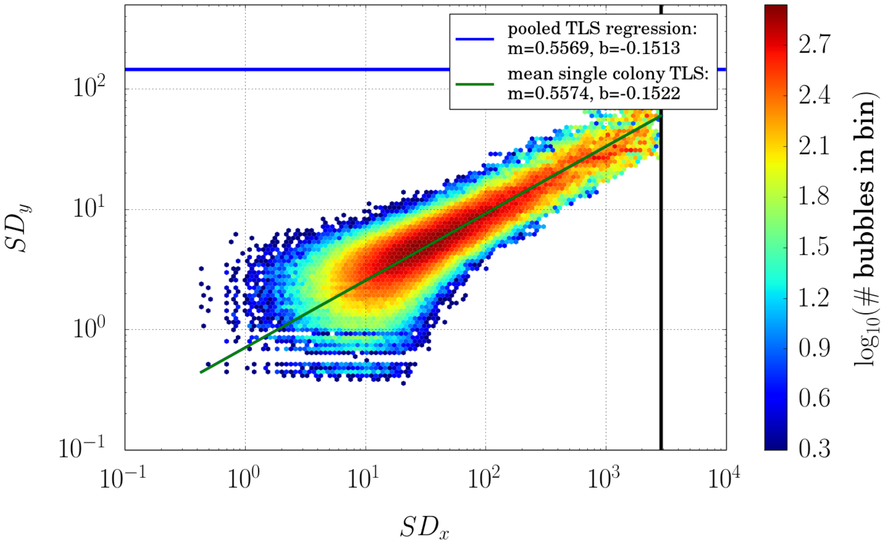


(b) $v = 0.3$.

FIGURE B.7.: **Bubble shapes for the fully quenched coalescent.** See Figure B.10 for the description.

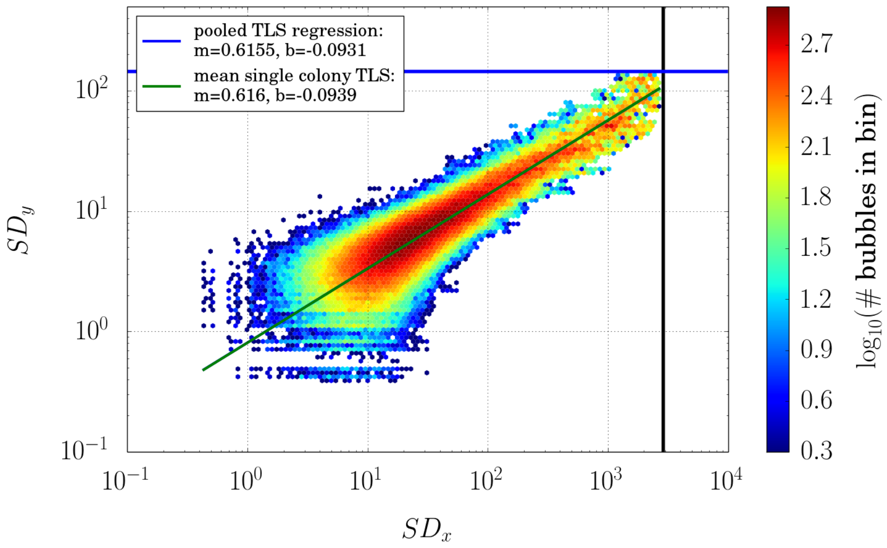


(a) $v = 0.4$.

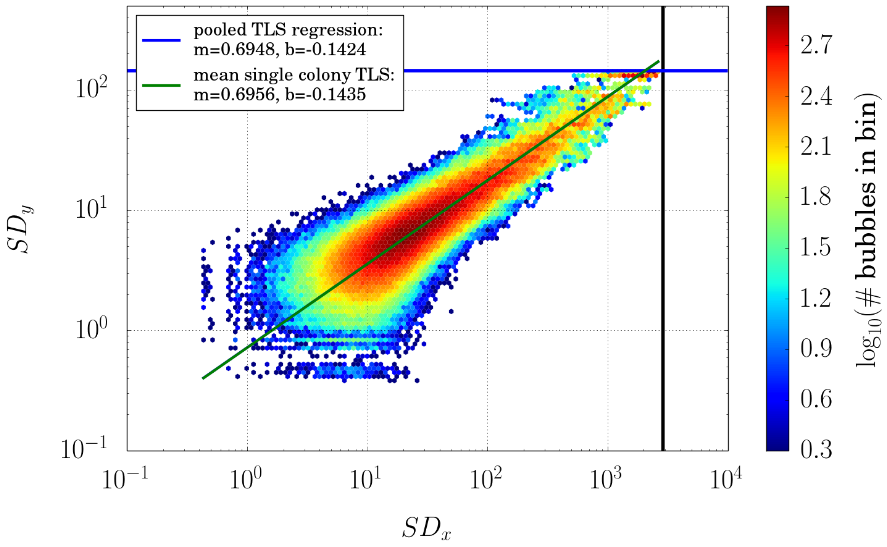


(b) $v = 0.5$.

FIGURE B.8.: **Bubble shapes for the fully quenched coalescent.** See Figure B.10 for the description.

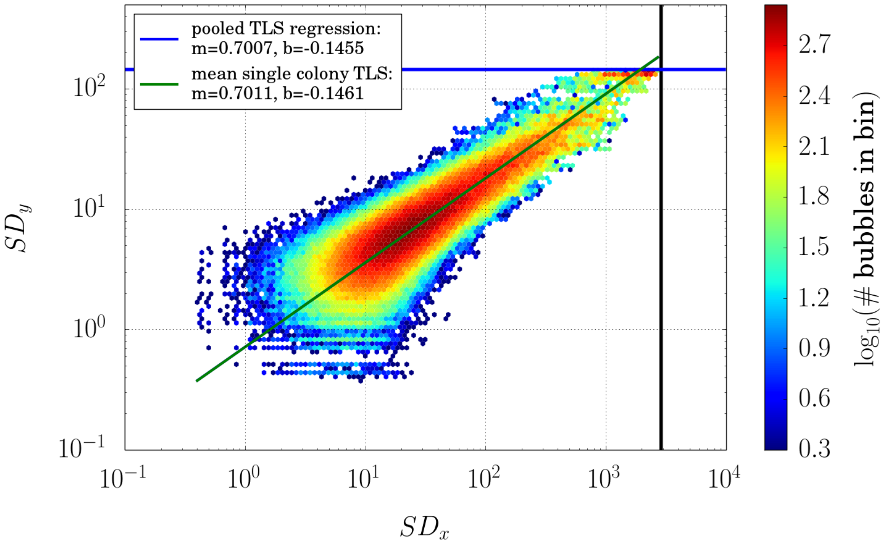


(a) $v = 0.6$.



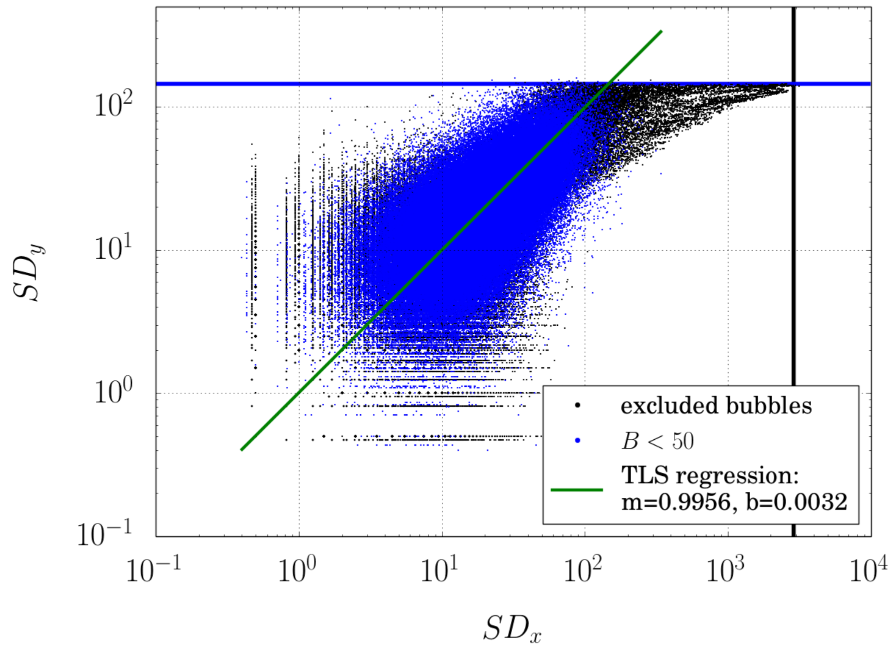
(b) $v = 0.7$.

FIGURE B.9.: **Bubble shapes for the fully quenched coalescent.** See Figure B.10 for the description.

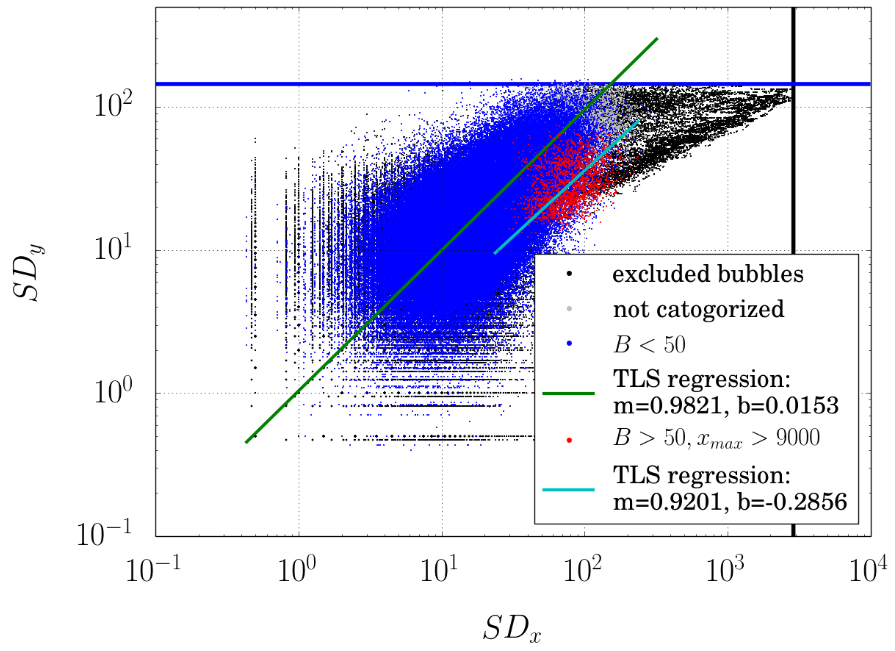


(a) $v = 1.0$.

FIGURE B.10.: **Bubble shapes for the fully quenched coalescent.** All bubble shapes form a single cluster, as expected. The results for the different velocities differ mainly in the slope of the total least squares regression (see Figure 11.29).

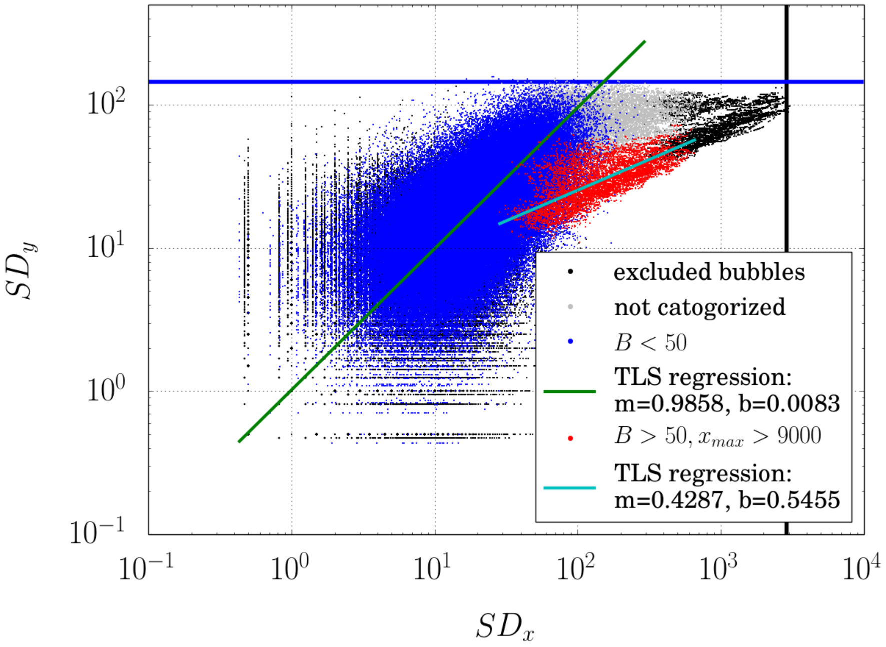


(a) $v = 0.1$.

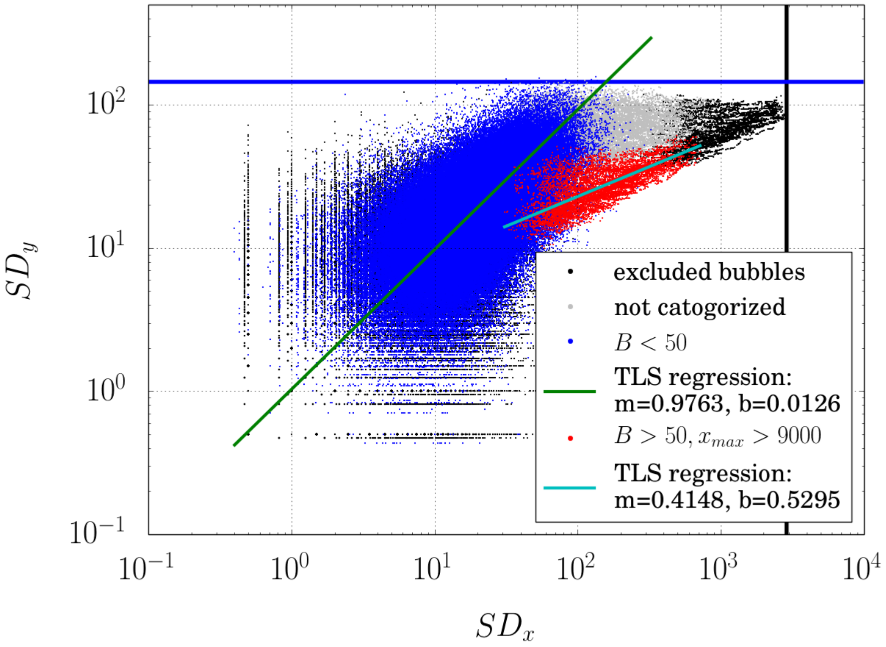


(b) $v = 0.2$.

FIGURE B.11.: **Bubble shapes for the quenched–front coalescent.** See Figure B.13 for the description. Note that, the bubble shapes for $v = 0.2$ show an 'power–law bar' but these bubbles where excluded based on our choice of the critical bubble size. The (red) group of bubbles that was chosen for the regression did not give a correct prediction in this case. The bubble size filter is not perfect and in real data sets an optimal filter might require a more detailed analysis.

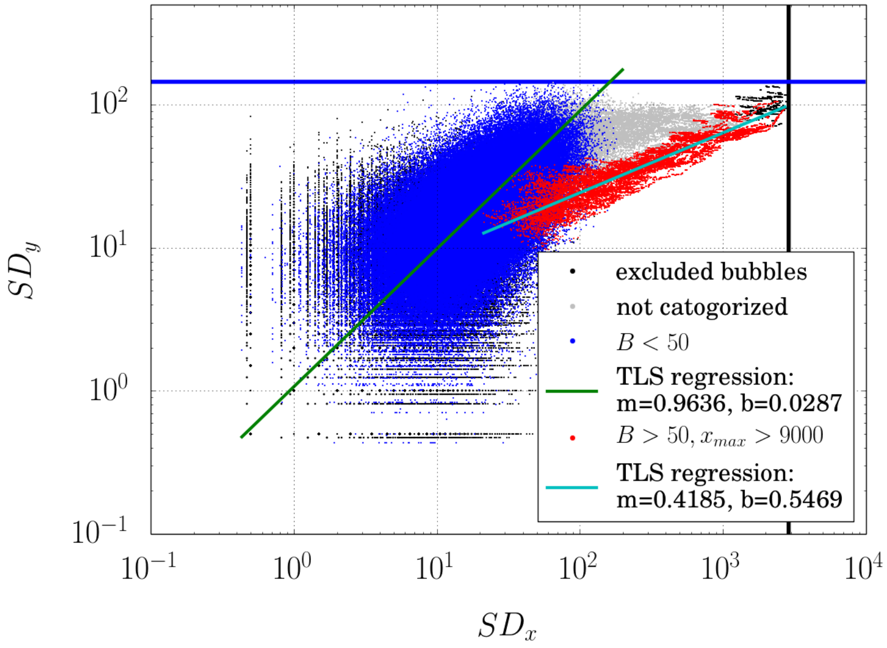


(a) $v = 0.3$.



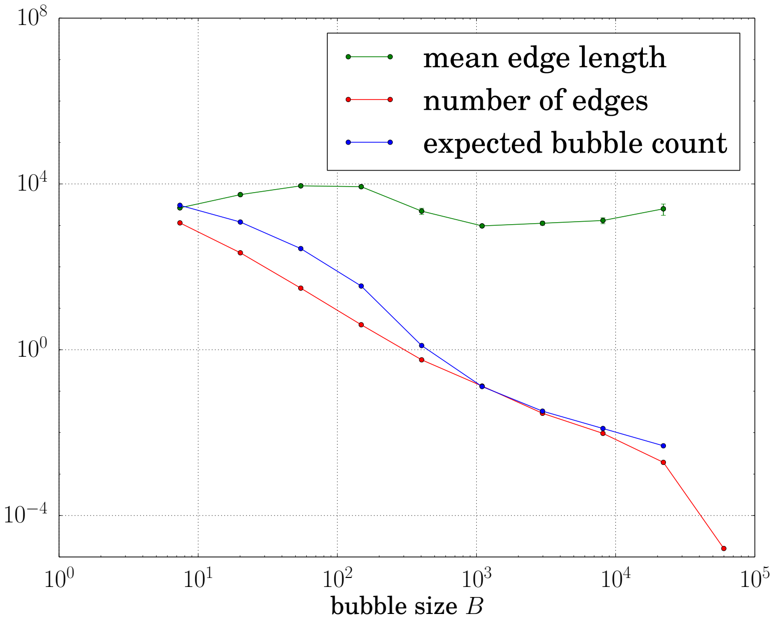
(b) $v = 0.4$.

FIGURE B.12.: **Bubble shapes for the quenched–front coalescent.** See Figure B.13 for the description.

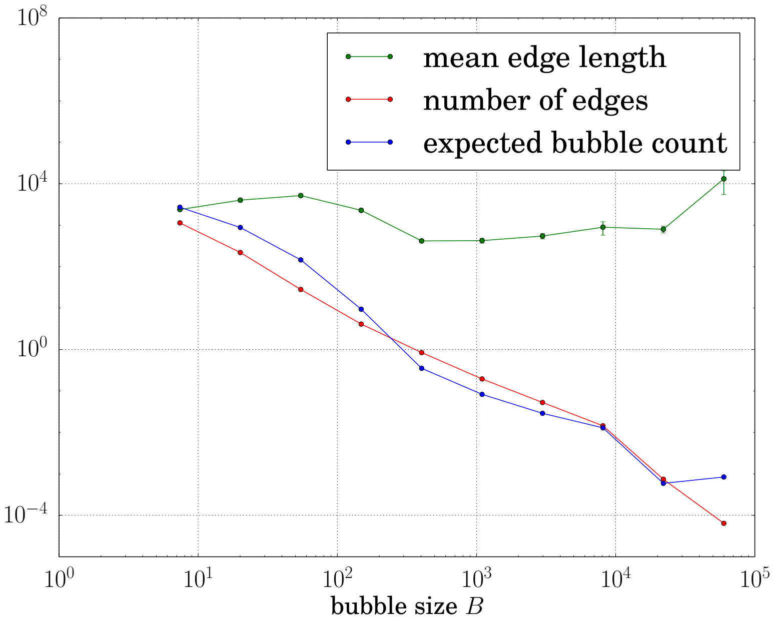


(a) $v = 0.6$.

FIGURE B.13.: **Bubble shapes for the quenched–front coalescent.** The results for the velocities below the phenotypical expansion velocity differ only in details from the unquenched coalescent. The detailed analysis is equivalent to the corresponding analysis in Figures 11.11 to 11.18.

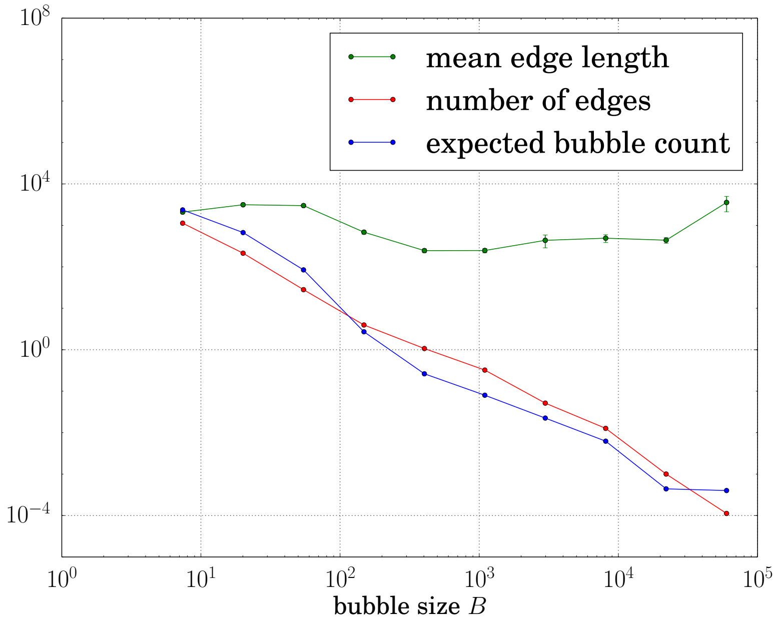


(a) **Bubble occurrence for** $v = 0.1$**.**

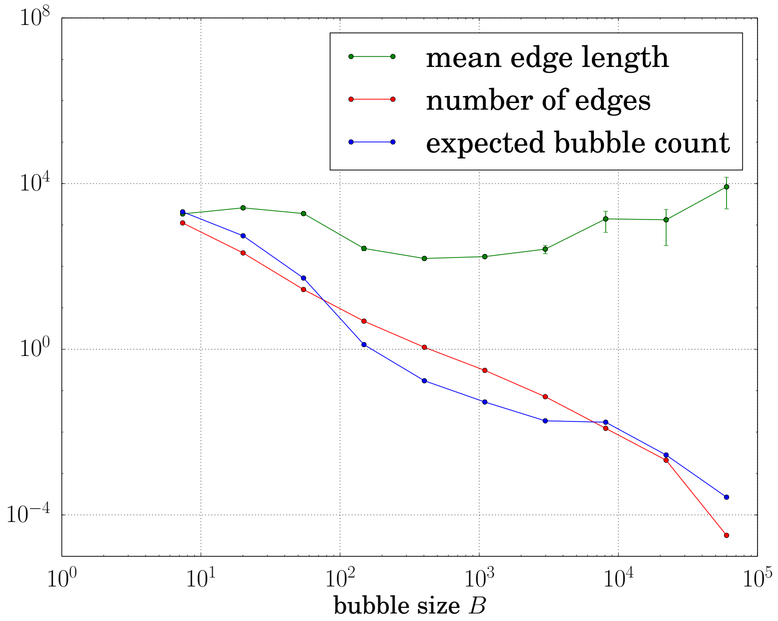


(b) **Bubble occurrence for** $v = 0.2$**.**

FIGURE B.14.: The mean number edges on the coalescence tree and the mean edge length in generations are displayed as a function of the bubble size. The bubble sizes are binned logarithmically. The expected bubble count refers to the number of bubbles expected in a sample. Of course, it depends on the mutation rate. Here we display the product of the number of edges and the mean edge length (rescaled by 10^{-3} for convenience).

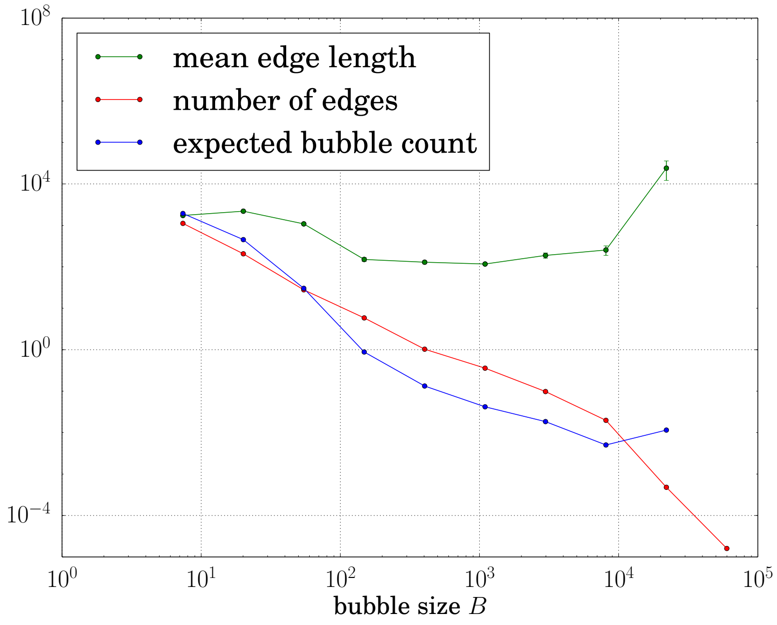


(a) **Bubble occurrence for $v = 0.3$.**



(b) **Bubble occurrence for $v = 0.4$.**

FIGURE B.15.: The mean number edges on the coalescence tree and the mean edge length in generations are displayed as a function of the bubble size. The bubble sizes are binned logarithmically. The expected bubble count refers to the number of bubbles expected in a sample. Of course, it depends on the mutation rate. Here we display the product of the number of edges and the mean edge length (rescaled by 10^{-3} for convenience).



(a) **Bubble occurrence for** $v = 0.5$**.**

FIGURE B.16.: The mean number edges on the coalescence tree and the mean edge length in generations are displayed as a function of the bubble size. The bubble sizes are binned logarithmically. The expected bubble count refers to the number of bubbles expected in a sample. Of course, it depends on the mutation rate. Here we display the product of the number of edges and the mean edge length (rescaled by 10^{-3} for convenience).

B.3. Simulation code

The simulation code is written in Python 2.7.3 (`http://www.python.org/`). The Figures in part II were all created using the matplotlib library developed by John D. Hunter [53].

B.3.1. The spatial distribution of alleles in expanding populations

The analysis of the spatial distribution of mutations and of the impact of the colonization on the coalescent encompasses a forward simulation based on the Eden model and three versions of a spatial coalescent.

All simulations are written in Python 2.7.3 (`http://www.python.org/`).

B.3.1.1. Colonization based on the Eden model

The function `periodicEDENFUNCTION` creates a colony according to the Eden model on a two–dimensional Karthesian lattice. The parameters are the velocity of the boundary v, the width of the habitat w, the habitat length L, and the starting position of the boundary.

```
def periodicEDENFUNCTION(w, v, L, BoundaryStart):
    # create dictionary of active demes.
    # The keys are the active demes, the value is the list of directions to open
        demes.
    activeDemes      = {}
    # define possible growth directions
    allDirections    = [(1,0), (-1,0), (0, 1), (0,-1)]

    # set up initial configuration of active demes:
    for yKoord in range(w):
        activeDemes[(0, yKoord)] = [(1, 0)]
        # the demes in the initial colony can only colonize their neighbors to the
             right.

    colonyList       = [(i, (0,0), 0) for i in activeDemes.keys()]
    # Entries of the colonyList: ( offspringDeme, parentDeme, ColonizationTime )
    InitialColony    = [i for (i, j, k) in colonyList]
    print 'Colonization starts from\n', InitialColony

    # create dictionary for the identification of the genotypes
    # This information is used to determine each deme's color for the plot of the
        colony.
    SourceDict       = {}
    for deme in InitialColony:
        SourceDict[deme] = deme

    # set additional parameters
    realtime         = 0    # time in generations
    steps            = 0    # number of simulation steps
    colonizations    = 0    # number of successful colonizations

    # The colony growth is stopped, when the first deme at x=L is colonized.
    # xMax holds the current rightmost position inside the colony.
    xMax = 0
    while xMax < L:
        survivors = [SourceDict[deme] for deme in activeDemes.keys()]
        # Choose parentDeme and offspringDeme:
        parentDeme    = random.choice(activeDemes.keys())
        direction     = random.choice(allDirections)
        offspringDeme = ( parentDeme[0]+direction[0], parentDeme[1]+direction[1] )
        realtime     += 1./len(activeDemes)
        steps        += 1

        # Apply periodic boundary:
        if offspringDeme[1]   == w:
            offspringDeme = (offspringDeme[0], 0)
        elif offspringDeme[1] == -1:
            offspringDeme = (offspringDeme[0], w-1)

        # Determine if the colonization attempt is successful:
        if direction not in activeDemes[parentDeme]:
            # not successful, target deme is passive.
            pass
        elif offspringDeme[0] > realtime*v + BoundaryStart:
            # not successful, target deme is blocked by the boundary.
           pass
        else:
            # colonization sucessful.
            if offspringDeme[0] > xMax:
                xMax = offspringDeme[0]
                print 'xMax =', xMax
            colonizations += 1
            SourceDict[offspringDeme] = SourceDict[parentDeme]
            # Add the new deme to the set of active demes.
            activeDemes[offspringDeme]=[]
            # Remark: The colonized neighbors of the new demes must be all active.
            # Therefore: Check in the activeDemes-dict, which of the 4 neighbors
                are free.
            # For up and down, respect periodic boundary:
            upperNeighbor = (offspringDeme[0],offspringDeme[1]+1)
            if upperNeighbor[1]==w:
                upperNeighbor = (offspringDeme[0],0)
            if upperNeighbor not in activeDemes.keys():
                activeDemes[offspringDeme].append((0,1))
            lowerNeighbor = (offspringDeme[0],offspringDeme[1]-1)
            if lowerNeighbor[1]==-1:
                lowerNeighbor = (offspringDeme[0],w-1)
            if lowerNeighbor not in activeDemes.keys():
                # The deme below is free (and inside the habitat).
                activeDemes[offspringDeme].append((0,-1))
            if not ((offspringDeme[0]+1,offspringDeme[1]) in activeDemes.keys() ):
                # The deme to the right is free.
                activeDemes[offspringDeme].append((1,0))
            if not ((offspringDeme[0]-1,offspringDeme[1]) in activeDemes.keys() ):
                # The deme to the left is free (and inside the habitat).
                activeDemes[offspringDeme].append((-1,0))
            if activeDemes[offspringDeme]==[]:
                # The new deme is not active.
                del activeDemes[offspringDeme]
            # Now, update the d-values of the new deme's neighbors.
            # Check for all four directions, if an active Deme ist there.
            # If so, refresh its list of free neighbors.
            for direction2neighbor in allDirections:
                # determine the neighbor:
                neighbor = (offspringDeme[0]+direction2neighbor[0],
                            offspringDeme[1]+direction2neighbor[1])
                # apply periodic boundary to neighbor coordinates:
                if neighbor[1]==w:
                    # the y-coordinate y=w is outside the habitat, it corresponds
                        to y=0.
                    neighbor = (neighbor[0],0)
                elif neighbor[1]==-1:
                    # y=-1 corresponds to y=w-1.
                    neighbor = (neighbor[0],w-1)
                # remove new deme from its neighbors freedoms:
                if neighbor in activeDemes.keys():
                    inverseDirection = (-direction2neighbor[0], -
                        direction2neighbor[1])
                    # that's the direction to delete from the neighbor.
                    activeDemes[neighbor].remove(inverseDirection)
                    if activeDemes[neighbor] == []:
                        # In that case, this neighbor is not longer active.
                        del activeDemes[neighbor]
            colonyList.append( (offspringDeme, parentDeme, realtime) )
    print 'Colonization is finished.'
    # colonyList holds the complete information on the colonization process.
```

B.3.1.2. The fully quenched coalescent

The coalescence functions are called with the Eden colony as `ForwardColonyDict`. The output contains the coalescence tree with all intermediate coalescence events and the samples involved. A dictionary with the detailed coalescence paths with every single step of the lineages is optional.

```
def EdenCoalescenceA(   ForwardColonyDict,
                        positionVector,
                        T,
                        fT,
                        detailedPaths = False
                        ):
    ColonyDict = ForwardColonyDict
    finaltime  = fT

    print 'Starting the coalescence process with:'
    print 'End of colonization =', T

    # Initialize dictionary for the coalescence paths
    pathDict = {}
    if detailedPaths:
        # enter sampling position to path:
        print '\tRecording detailed paths.'
        for site in positionVector:
            pathDict[site[1]] = [site[0]]

    # Initialize dictionary for the coalescence tree.
    # The items are edges with extra info.
    # keys: all sampling sites merged into that lineage
    # values: time of coalescence that created the edge  (creation time)
    #         place where this happened                  (creation place)
    #         time of next merging into that lineage     (melting time)
    #         place where this happened                  (melting place)
    # Rq: Values 1 and 2 are written when the edge is created,
    #     values 3 and 4 when the edge is closed.
    coalDict = {}
    for site in positionVector:
        # site[0] = location, site[1] = ID, site[2] = IDcollection
        coalDict[ site[1] ] = [ (site[0], site[2][:], finaltime) ]


    # Start coalescence process:
    while T > 0 and len(positionVector) > 1:
        T += -1
        # move the lineages:
        for index in range(len(positionVector)):
            while ColonyDict[positionVector[index][0]][1] > T:
                # move to ancestor (forced move)
                positionVector[index][0] = ColonyDict[positionVector[index][0]][0]
            # enter position into path dictionary
            if detailedPaths: pathDict[positionVector[index][1]].append(
                positionVector[index][0])

        # check for coalescence
        # sort positionVector => lineages at the same place get next to each other
            .
        positionVector.sort()
        killVector       = []
        coalescenceState = False
        # iterate over positionVector:
        for index in range(len(positionVector)-1):
            if positionVector[index][0] == positionVector[index+1][0]:
                # Coalescence!

                # Note the indices of the others in the killVector:
                killVector.append(index+1)

                # Append the coalescence event to the coalDict-entry (This will be
                     the last entry).
                finalIDs = positionVector[index+1][2][:]
                coalDict[positionVector[index+1][1]].append( (positionVector[index
                    +1][0], finalIDs, T) )

                # The first lineage at each coordinate remains in the list.
                # It collects the lineage IDs of the dying lineages.
                if coalescenceState == False:
                    # This happens, only for the first one of the lineages at the
                        same place.
                    # This lineage will survive in the positionVector.
                    survivorIDandTime = (positionVector[index][:], T)
                    survivorIndex     = index
                    # The entry in the coalPath for the survivorIndex will be
                        added below.

                # add the coalescing IDs to the list 'currentIDs'.
                # (The object 'survivorIndex' is created in the first occurrence
                    of a coalescence series.
                #  The vanishing lineage IDs are added here.)
                positionVector[survivorIndex][2] += positionVector[index+1][2]
                # Set the coalescenceState = True, to indicate that we are now
                    inside a coalescence group.
                coalescenceState = True

                # Append the survivor ID and the current time to the dead lineage.
                if detailedPaths: pathDict[positionVector[index+1][1]].append(
                    survivorIDandTime)

            elif coalescenceState == True:
                # This happens only for the last one of the lineages at the same
                    place.
                # The series of local coalescences is over.
                # Set the coalState to False:
                coalescenceState = False
                # Append the coalescence event to the coalDict-entry
                currentIDs = positionVector[survivorIndex][2][:]

                # Append the coalescence event to the survivor coalPath:
                coalDict[positionVector[survivorIndex][1]].append( (positionVector
                    [survivorIndex][0], currentIDs, T) )

        if coalescenceState == True:
            # When the last two entries of the positionVector are at the same
                place,
            # we must add the path vertex of the survivorIndex here.
            currentIDs = positionVector[survivorIndex][2][:]
            coalDict[positionVector[survivorIndex][1]].append( (positionVector[
                survivorIndex][0], currentIDs, T) )

        # update the positionVector
        if killVector != []:
            # only now, we have to update the position vector
            newPositionVector = [positionVector[i] for i in range(len(
                positionVector)) if i not in killVector]
            positionVector    = newPositionVector[:]

    # If the coalescence is not complete, add the edges to the source demes in the
         initial colony.
    if len(positionVector) > 1:
        for entry in positionVector:
            coalDict[entry[1]].append( (entry[0], entry[2], T) )

    if detailedPaths:
        # write final entry to survivor lineage entries
        for entry in positionVector:
            pathDict[entry[1]].append( (entry, T) )

    print 'Super ancestor(s) at:'
    for entry in positionVector:
        print entry[0]
    print 'time of final coalescence / end of process:', T
    return (pathDict, T, coalDict)
```

B.3.1.3. The front–quenched coalescent

This version simulates diffusive lineage movement in addition to lineage movement strictly along the colonization paths. The block after `if not forcedMove:` adds the diffusive movement.

```python
def EdenCoalescenceB(   ForwardColonyDict,
                        positionVector,
                        T,
                        fT,
                        detailedPaths = False
                        ):
    ColonyDict = ForwardColonyDict
    finaltime  = fT

    print 'Starting the coalescence process with:'
    print 'End of colonization =', T

    # Initialize dictionary for the coalescence paths
    pathDict = {}
    if detailedPaths:
        # enter sampling position to path:
        print '\tRecording detailed paths.'
        for site in positionVector:
            pathDict[site[1]] = [site[0]]

    # Initialize dictionary for the coalescence tree. The items are edges with
        extra info.
    # keys: all sampling sites merged into that lineage
    # values: time of coalescence that created the edge  (creation time)
    #         place where this happened                  (creation place)
    #         time of next merging into that lineage     (melting time)
    #         place where this happened                  (melting place)
    # Rq: Values 1 and 2 are written when the edge is created,
    #     values 3 and 4 when the edge is closed.
    coalDict = {}
    for site in positionVector:
        # site[0] = location, site[1] = ID, site[2] = IDcollection
        coalDict[ site[1] ] = [ (site[0], site[2][:], finaltime) ]


    # Start coalescence process:
    while T > 0 and len(positionVector) > 1:
        T += -1
        # move the lineages:
        for index in range(len(positionVector)):
            forcedMove = False
            while ColonyDict[positionVector[index][0]][1] > T:
                # move to ancestor (forced move)
                positionVector[index][0] = ColonyDict[positionVector[index][0]][0]
                forcedMove = True
            if not forcedMove:
                # make 1 random step
                direction = random.choice(((1, 0), (-1, 0), (0, 1), (0, -1)))
                target    = (positionVector[index][0][0]+direction[0],
                    positionVector[index][0][1]+direction[1])
                if target in ColonyDict:
                    positionVector[index][0] = target
            # enter position into path dictionary
            if detailedPaths: pathDict[positionVector[index][1]].append(
                positionVector[index][0])

        # check for coalescence
        # sort positionVector => lineages at the same place get next to each other
            .
        positionVector.sort()
        killVector       = []
        coalescenceState = False
        # iterate over positionVector:
        for index in range(len(positionVector)-1):
            if positionVector[index][0] == positionVector[index+1][0]:
                # Coalescence!
                print '\t', len(positionVector), 'lineages left.'

                # Note the indices of the others in the killVector:
                killVector.append(index+1)

                # Append the coalescence event to the coalDict-entry (This will be
                     the last entry).
                finalIDs = positionVector[index+1][2][:]
                coalDict[positionVector[index+1][1]].append( (positionVector[index
                    +1][0], finalIDs, T) )

                # The first lineage at each coordinate remains in the list.
                # It collects the lineage IDs of the dying lineages.
                if coalescenceState == False:
                    # This happens, only for the first one of the lineages at the
                        same place.
                    # This lineage will survive in the positionVector.
                    survivorIDandTime = (positionVector[index][:], T)
                    survivorIndex     = index
                    # The entry in the coalPath for the survivorIndex will be
                        added below.

                # add the coalescing IDs to the list 'currentIDs'.
                # (The object 'survivorIndex' is created in the first occurrence
                    of a coalescence series.
                #  The vanishing lineage IDs are added here.)
                positionVector[survivorIndex][2] += positionVector[index+1][2]

                # Set the coalescenceState = True, to indicate that we are now
                    inside a coalescence group.
                coalescenceState = True

                # Append the survivor ID and the current time to the dead lineage.
                if detailedPaths: pathDict[positionVector[index+1][1]].append(
                    survivorIDandTime)

            elif coalescenceState == True:
                # This happens only for the last one of the lineages at the same
                    place.
                # The series of local coalescences is over.
                # Set the coalState to False:
                coalescenceState = False
                # Append the coalescence event to the coalDict-entry
                currentIDs = positionVector[survivorIndex][2][:]
                # Append the coalescence event to the survivor coalPath:
                coalDict[positionVector[survivorIndex][1]].append( (positionVector
                    [survivorIndex][0], currentIDs, T) )

        if coalescenceState == True:
            # When the last two entries of the positionVector are at the same
                place,
            # we must add the path vertex of the survivorIndex here.
            currentIDs = positionVector[survivorIndex][2][:]
            coalDict[positionVector[survivorIndex][1]].append( (positionVector[
                survivorIndex][0], currentIDs, T) )


        if killVector != []:
            # only now, we have to update the position vector
            newPositionVector = [positionVector[i] for i in range(len(
                positionVector)) if i not in killVector]
            positionVector    = newPositionVector[:]

    # If the coalescence is not complete, add the edges to the source demes in the
         initial colony.
    if len(positionVector) > 1:
        for entry in positionVector:
            coalDict[entry[1]].append( (entry[0], entry[2], T) )

    if detailedPaths:
        # write final entry to survivor lineage entries
        for entry in positionVector:
            pathDict[entry[1]].append( (entry, T) )

    print ’Super ancestor(s) at:’
    for entry in positionVector:
        print entry[0]
    print ’time of final coalescence / end of process:’, T
    return (pathDict, T, coalDict)
```

B.3.1.4. Diffusive lineage movement within the whole habitat

The coalescence function with only diffusive lineage movement is almost identical to the function `EdenCoalescenceB`: The block with the movement along the colonization paths is removed, of course.

Acknowledgements First of all, I would like to express my gratitude towards my supervisor Oskar Hallatschek for inviting me to join his at that time newly founded work group. My scientific background was almost exclusively pure mathematics at that point and it was thanks to his open–mindedness that I had the opportunity to spend four interesting and stimulating years in Göttingen. During my time at the Max–Planck–institute for Dynamics and Self–Organization, Oskar's scientific expertise and experience proved to be invaluable in many hours of fruitful discussions.

I thank Marc Timme and Anja Sturm for their inspiring questions and suggestions as members of my thesis committee.

Many thanks go to every single member of the Biological Physics and Evolutionary Dynamics group: Erik A. Martens, Sven Boekhoff, Robert Filter, Christopher Seedig, Marc Szabo, Niels Podewitz, Paulo Pinto, Hedvika Toncrova, Martin Ehrle, Marijn Korsten, Jan Lebert, Matthias Meschede, Clemens Buss, Fabian Stiewe, Lukas Geyrhofer, Jörn Hartung and Jean–François Flot.

Erik, Sven, Robert, Matthias, Clemens, Fabian, Lukas and Jean–François deserve special thanks for their scientific support. Lukas deserves the most special thanks for whatever. Outside our workgroup, I would like to thank Hinrich Arnoldt and Mirko Lukovic for valuable insight on evolution and random walks, respectively. Harold Gutch and Sven Jahnke proved to be indispensable in my struggle with all those little and not so little programming issues.

When I came to Göttingen, I felt welcome right away. Fabio Schittler–Neves, my first officemate, was one of the main reasons. Many other wonderful colleagues contributed to the wonderful atmosphere at the institute. Nevertheless, the whole institute would probably have collapsed without Barbara Kutz.

Particular thanks go to Erik and his wife Nadya for sharing an apartment with me and still deciding to leave me physically intact. The same is true for Kati Rimbach.

I thank my parents, my brother and my sister for their unconditional love and support.

Index

allele, 10

coalescent, 14
  p–quenched, 120
  front–quenched, 69
  fully quenched, 69
  unquenched, 70
colonization time, 66
compression factor, 74
continuous model, 26
critical patch size, 72

deme, 10, 65

enforced phase of coalescence, 23, 24
expanding stepping stone model, 25
expansion velocity
  phenotype–limited, 65

fixation, 10
free phase of coalescence, 23, 24

gene surfing, 13
genetic diversity, 11
genetic drift, 12

heterozygosity, 11
hitchhiking, 64

identity by descent, 14

limited migration, 12
linkage, 63
linkage disequilibrium, 64
locus, 10

most recent common ancestor, 14
MRCA, 14
mutant, 10
mutation, 11
  beneficial, 11
  deleterious, 11
  neutral, 11

natal dispersal distance, 12

patch
  mutant, 60, 61
  non–surfer, 60
  surfer, 60
patch shape, 71
patch size, 71
phenotype-limited range expansions, 21
phenotypical expansion velocity, 79
polymorphic site, 10

quenched randomness, 64

range shift, 45

spatial structure, 10, 12

t_{MRCA}, 14

wild–type, 10

Bibliography

[1] M. Arenas, N. Ray, M. Currat, and L. Excoffier. Consequences of range contractions and range shifts on molecular diversity. *Molecular Biology and Evolution*, 29(1):207–218, 2012.

[2] Frederic Austerlitz, Bernard Jung-Muller, Bernard Godelle, and Pierre-Henri Gouyon. Evolution of coalescence times, genetic diversity and structure during colonization. *Theoretical Population Biology*, 51(2):148–164, 1997.

[3] HG Baker. Stages in invasion and replacement demonstrated by species of melandrium. *The Journal of Ecology*, pages 96–119, 1948.

[4] J. Bart, S. Droege, P. Geissler, B. Peterjohn, and C.J. Ralph. Density estimation in wildlife surveys. *Wildlife Society Bulletin*, 32(4):1242–1247, 2004.

[5] N. H. Barton, A. M. Etheridge, J. Kelleher, and A. Véber. Genetic hitchhiking in spatially extended populations. 2012. preprint.

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

[7] Nicholas H Barton. The effect of hitch-hiking on neutral genealogies. *Genetical Research*, 72(2):123–133, 1998.

[8] A. Battisti, M. Stastny, S. Netherer, C. Robinet, A. Schopf, A. Roques, and S. Larsson. Expansion of geographic range in the pine processionary moth caused by increased winter temperatures. *Ecological Applications*, 15(6):2084–2096, 2005.

[9] David R Bentley, Shankar Balasubramanian, Harold P Swerdlow, Geoffrey P Smith, John Milton, Clive G Brown, Kevin P Hall, Dirk J Evers, Colin L Barnes, Helen R Bignell, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, 2008.

[10] J.E. Bronnenhuber, B.A. Dufour, D.M. Higgs, and D.D. Heath. Dispersal strategies, secondary range expansion and invasion genetics of the non-indigenous round goby, neogobius melanostomus, in great lakes tributaries. *Molecular Ecology*, 20(9):1845–1859, 2011.

[11] É. Brunet and B. Derrida. How genealogies are affected by the speed of evolution. *Philosophical Magazine*, 92(1-3):255–271, 2012.

[12] E. Brunet, B. Derrida, AH Mueller, and S. Munier. Noisy traveling waves: effect of selection on genealogies. *EPL (Europhysics Letters)*, 76:1, 2006.

[13] FL Bunnell and AS Harestad. Dispersal and dispersion of black-tailed deer: models and observations. *Journal of Mammalogy*, 64(2):201–209, 1983.

[14] R.B. Chandler, J.A. Royle, and D.I. King. Inference about density and temporary emigration in unmarked populations. *Ecology*, 92(7):1429–1435, 2011.

[15] B Charlesworth, D Charlesworth, and NH Barton. The effects of genetic and geographic structure on neutral variation. *Annual Review of Ecology, Evolution, and Systematics*, 34(1):99–125, 2003.

[16] I.-C Chen, J. K Hill, R Ohlemuller, D. B Roy, and C. D Thomas. Rapid range shifts of species associated with high levels of climate warming. *Science*, 333(6045):1024–1026, 2011.

[17] Robert Cornman, Michael Schatz, J Johnston, Yan-Ping Chen, Jeff Pettis, Greg Hunt, Lanie Bourgeois, Chris Elsik, Denis Anderson, Christina Grozinger, et al. Genomic survey of the ectoparasitic mite varroa destructor, a major pest of the honey bee apis mellifera. *BMC genomics*, 11(1):602, 2010.

[18] L Crozier. Warmer winters drive butterfly range expansion by increasing survivorship. *Ecology*, 85(1):231–241, 2004.

[19] M Currat and L Excoffier. The effect of the neolithic expansion on european molecular diversity. *Proceedings of the Royal Society B: Biological Sciences*, 272(1564):679–688, 2005.

[20] Andrew Curry. The milk revolution, 2013.

[21] M DeGiorgio, J.H Degnan, and N.A Rosenberg. Coalescence-time distributions in a serial founder model of human evolutionary history. *Genetics*, 189(2):579–593, 2011.

[22] Eric Dumonteil, Satya N Majumdar, Alberto Rosso, and Andrea Zoia. Spatial extent of an outbreak in animal epidemics. *Proceedings of the National Academy of Sciences*, 110(11):4239–4244, 2013.

[23] Ian Dunham, Ewan Birney, Bryan R Lajoie, Amartya Sanyal, Xianjun Dong, Melissa Greven, Xinying Lin, Jie Wang, Troy W Whitfield, Jiali Zhuang, et al. An integrated encyclopedia of dna elements in the human genome. 2012.

[24] Murray Eden. A two-dimensional growth process. *Dynamics of fractal surfaces*, pages 265–283, 1961.

[25] C.A. Edmonds, A.S. Lillie, and L.L. Cavalli-Sforza. Mutations arising in the wave front of an expanding population. *Proceedings of the National Academy of Sciences of the United States of America*, 101(4):975, 2004.

[26] B. Harrington et al. Inkscape. `http://www.inkscape.org/`, 2004-2005.

[27] WJ Ewens. A note on the sampling theory for infinite alleles and infinite sites models. *Theoretical population biology*, 6(2):143–148, 1974.

[28] L. Excoffier, M. Foll, and R.J. Petit. Genetic consequences of range expansions. *Annual Review of Ecology, Evolution, and Systematics*, 40:481–501, 2009.

[29] L Excoffier and N Ray. Surfing during population expansions promotes genetic revolutions and structuration. *Trends in Ecology & Evolution*, 23(7):347–351, 2008.

[30] L. Excoffier and N. Ray. Surfing during population expansions promotes genetic revolutions and structuration. *Trends in Ecology & Evolution*, 23(7):347–351, 2008.

[31] RA Fisher. The genetical theory of natural selection, 1930.

[32] R.A. Fisher. The wave of advance of advantageous genes. *Annals of Human Genetics*, 7(4):355–369, 1937.

[33] Michael C Fontaine, Fréderic Austerlitz, Tatiana Giraud, Frédéric Labbé, Daciana Papura, Sylvie Richard-Cervera, and François Delmotte. Genetic signature of a range expansion and leap-frog event after the recent invasion of europe by the grapevine downy mildew pathogen plasmopara viticola. *Molecular ecology*, 2013.

[34] William E Fry and Stephen B Goodwin. Resurgence of the irish potato famine fungus. *Bioscience*, 47(6):363–371, 1997.

[35] LL Getz, FR Cole, and DL Gates. Interstate roadsides as dispersal routes for microtus pennsylvanicus. *Journal of Mammalogy*, 59(1):208–212, 1978.

[36] John H Gillespie. *Population genetics: a concise guide*. JHU Press, 2010.

[37] Rodolphe Elie Gozlan, JR Britton, I Cowx, and GH Copp. Current knowledge on non-native freshwater fish introductions. *Journal of Fish Biology*, 76(4):751–786, 2010.

[38] Dan Graur, Yichen Zheng, Nicholas Price, Ricardo BR Azevedo, Rebecca A Zufall, and Eran Elhaik. On the immortality of television sets: "function" in the human genome according to the evolution–free gospel of encode. *Genome biology and evolution*, 5(3):578–590, 2013.

[39] S.M. Haig and L.W. Oring. Distribution and dispersal in the piping plover. *The Auk*, pages 630–638, 1988.

[40] O. Hallatschek, P. Hersen, S. Ramanathan, and D.R. Nelson. Genetic drift at expanding frontiers promotes gene segregation. *Proceedings of the National Academy of Sciences*, 104(50):19926–19930, 2007.

[41] O. Hallatschek and D.R. Nelson. Gene surfing in expanding populations. *Theoretical population biology*, 73(1):158–170, 2008.

[42] O. Hallatschek and D.R. Nelson. Population genetics and range expansions. *Physics Today*, 62:42, 2009.

[43] O. Hallatschek and D.R. Nelson. Life at the front of an expanding population. *Evolution*, 64(1):193–206, 2010.

[44] D.L. Hartl, A.G. Clark, et al. *Principles of population genetics*, volume 7. Sinauer associates Sunderland, Massachusetts, 1997.

[45] ME Hellberg, DP Balch, and K Roy. Climate-driven range expansion and morphological evolution in a marine gastropod. *Science*, 292(5522):1707, 2001.

[46] G Hewitt. The genetic legacy of the quaternary ice ages. *Nature*, 405(6789):907–913, 2000.

[47] GM Hewitt. Genetic consequences of climatic oscillations in the quaternary. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 359(1442):183–195, 2004.

[48] A.T. Hitch and P.L. Leberg. Breeding distributions of north american bird species moving north as a result of climate change. *Conservation Biology*, 21(2):534–539, 2007.

[49] J.T. Houghton et al. *Climate change 2001: the scientific basis*, volume 881. Cambridge University Press Cambridge, 2001.

[50] Richard R Hudson, M Slatkin, and WP Maddison. Estimation of levels of gene flow from dna sequence data. *Genetics*, 132(2):583–589, 1992.

[51] RR Hudson. Gene genealogies and the coalescent process. *Oxford surveys in evolutionary biology*, 7(1):44, 1990.

[52] Lesley Hughes. Biological consequences of global warming: is the signal already apparent? *Trends in Ecology & Evolution*, 15(2):56–61, 2000.

[53] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95, 2007.

[54] Julian Huxley et al. Evolution. the modern synthesis. *Evolution. The Modern Synthesis.*, 1942.

[55] W. Jetz, C. Carbone, J. Fulford, and J.H. Brown. The scaling of animal space use. *Science*, 306(5694):266–268, 2004.

[56] M Kimura. A mathematical analysis of the stepping stone model of genetic correlation. *Journal of Applied Probability*, Jan 1965.

[57] M Kimura and G.H Weiss. The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, 49(4):561, 1964.

[58] J Kingman. On the genealogy of large populations. *Journal of Applied Probability*, Jan 1982.

[59] J.F.C. Kingman. The coalescent. *Stochastic processes and their applications*, 13(3):235–248, 1982.

[60] S. Klopfstein, M. Currat, and L. Excoffier. The fate of mutations surfing on the wave of a range expansion. *Molecular Biology and Evolution*, 23(3):482–490, 2006.

[61] A. N. Kolmogorov, I. G. Petrovskii, and N. S. Piskunov. A study of the equation of diffusion with increase in the quantity of matter, and its application to a biological problem. *Bull. Moscow Univ. Math. Ser. A 1*, 1:713–719, 1937.

[62] K Korolev, M Avlund, O Hallatschek, and D R Nelson. Genetic demixing and evolution in linear stepping stone models. *Reviews of modern physics*, Jan 2010.

[63] R Leblois, F Rousset, D Tikel, C Moritz, and A Estoup. Absence of evidence for isolation by distance in an expanding cane toad (bufo marinus) population: an individual-based analysis of microsatellite genotypes. *Molecular ecology*, 9(11):1905–9, 2000.

[64] R. Lehe, O. Hallatschek, and L. Peliti. The rate of beneficial mutations surfing on the wave of a range expansion. *PLoS Computational Biology*, 8(3):e1002447, 2012.

[65] Wen-Hsiung Li. Distribution of nucleotide differences between two randomly chosen cistrons in a finite population. *Genetics*, 85(2):331–337, 1977.

[66] Gordon Luikart and Jean-Marie Cornuet. Empirical evaluation of a test for identifying recently bottlenecked populations from allele frequency data. *Conservation Biology*, 12(1):228–237, 1998.

[67] Satya N Majumdar, Alain Comtet, and Julien Randon-Furling. Random convex hulls and extreme value statistics. *Journal of Statistical Physics*, 138(6):955–1009, 2010.

[68] G. Malecot. Heterozygosity and relationship in regularly subdivided populations. *Theoretical population biology*, 8(2):212–241, 1975.

[69] Marcel Margulies, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, Lisa A Bemben, Jan Berka, Michael S Braverman, Yi-Ju Chen, Zhoutao Chen, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, 2005.

[70] E. Matthysen. Density-dependent dispersal in birds and mammals. *Ecography*, 28(3):403–416, 2005.

[71] GJ McInerny, JRG Turner, HY Wong, JMJ Travis, and TG Benton. How range shifts induced by climate change affect neutral evolution. *Proceedings of the Royal Society B: Biological Sciences*, 276(1661):1527–1534, 2009.

[72] P A P Moran. Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society*, 54:60–71, Dec 1958.

[73] Lloyd W Morrison, Sanford D Porter, Eric Daniels, and Michael D Korzukhin. Potential global range expansion of the invasive fire ant, solenopsis invicta. *Biological Invasions*, 6(2):183–191, 2004.

[74] T. Nagylaki. The strong-migration limit in geographically structured populations. *Journal of mathematical biology*, 9(2):101–114, 1980.

[75] T. Nagylaki. Gustave malécot and the transition from classical to modern population genetics. *Genetics*, 122(2):253, 1989.

[76] R.A. Neher and O. Hallatschek. Genealogies of rapidly adapting populations. *arXiv preprint arXiv:1208.3185*, 2012.

[77] J Novembre, T Johnson, K Bryc, Z Kutalik, AR Boyko, A Auton, A Indap, KS King, S Bergmann, and MR Nelson. Genes mirror geography within europe. *Nature*, 456(7218):98, 2008.

[78] Jens Nullmeier and Oskar Hallatschek. The coalescent in boundary-limited range expansions. *Evolution*, 2013.

[79] Richard Ogutu-Ohwayo and RE Hecky. Fish introductions in africa and some of their implications. *Canadian Journal of Fisheries and Aquatic Sciences*, 48(S1):8–12, 1991.

[80] C. Parmesan. Ecological and evolutionary responses to recent climate change. *Annu. Rev. Ecol. Evol. Syst.*, 37:637–669, 2006.

[81] C Parmesan, N Ryrholm, C Stefanescu, J.K Hill, C.D Thomas, H Descimon, B Huntley, L Kaila, J Kullberg, and T Tammaru. Poleward shifts in geographical ranges of butterfly species associated with regional warming. *Nature*, 399(6736):579–583, 1999.

[82] A.T. Peterson. Subtle recent distributional shifts in great plains bird species. *The Southwestern Naturalist*, 48(2):289–292, 2003.

[83] A.R. Pluess. Pursuing glacier retreat: genetic structure of a rapidly expanding *Larix decidua* population. *Molecular Ecology*, 20(3):473–485, 2011.

[84] F Prugnolle, A Manica, and F Balloux. Geography predicts neutral genetic diversity of human populations. *Current Biology*, 15(5):R159–R160, 2005.

[85] S. Ramachandran, O. Deshpande, C.C. Roseman, N.A. Rosenberg, M.W. Feldman, and L.L. Cavalli-Sforza. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in africa. *Proceedings of the National Academy of Sciences of the United States of America*, 102(44):15942, 2005.

[86] S Ramachandran, O Deshpande, C.C Roseman, N.A Rosenberg, M.W Feldman, and L.L Cavalli-Sforza. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in africa. *Proc Natl Acad Sci USA*, 102(44):15942, 2005.

[87] Jonathan M Rothberg, Wolfgang Hinz, Todd M Rearick, Jonathan Schultz, William Mileski, Mel Davey, John H Leamon, Kim Johnson, Mark J Milgrew, Matthew Edwards, et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356):348–352, 2011.

[88] Joseph Rudnick and George Gaspari. The shapes of random walks. *Science*, 237(4813):384–389, 1987.

[89] Peter Savolainen, Ya-ping Zhang, Jing Luo, Joakim Lundeberg, and Thomas Leitner. Genetic evidence for an east asian origin of domestic dogs. *Science*, 298(5598):1610–1613, 2002.

[90] M. Slatkin et al. Inbreeding coefficients and coalescence times. *Genet. Res*, 58(2):167–175, 1991.

[91] M. Slatkin and L. Excoffier. Serial founder effects during range expansion: a spatial analog of genetic drift. *Genetics*, 191(1):171–181, 2012.

[92] Montgomery Slatkin and Richard R Hudson. Pairwise comparisons of mitochondrial dna sequences in stable and exponentially growing populations. *Genetics*, 129(2):555–562, 1991.

[93] Karel Šolc. Shape of a random-flight chain. *The Journal of Chemical Physics*, 55:335, 1971.

[94] C Strobeck. Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. *Genetics*, 117(1):149–53, Sep 1987.

[95] C. Strobeck. Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. *Genetics*, 117(1):149–153, 1987.

[96] Michael PH Stumpf and David B Goldstein. Demography, recombination hotspot intensity, and the block structure of linkage disequilibrium. *Current biology*, 13(1):1–8, 2003.

[97] G.D. Sutherland, A.S. Harestad, K. Price, and K.P. Lertzman. Scaling of natal dispersal distances in terrestrial birds and mammals. *Conservation Ecology*, 4(1):16, 2000.

[98] J.M.J. Travis, T. Münkemüller, O.J. Burton, A. Best, C. Dytham, and K. Johst. Deleterious mutations can surf to high densities on the wave front of an expanding population. *Molecular Biology and Evolution*, 24(10):2334–2343, 2007.

[99] J Wakeley and S Lessard. Theory of the effects of population structure and sampling on patterns of linkage disequilibrium applied to genomic data from humans. *Genetics*, 164(3):1043, 2003.

[100] Jonathan M Waters, Ceridwen I Fraser, and Godfrey M Hewitt. Founder takes all: density-dependent processes structure biodiversity. *Trends in ecology & evolution*, 2012.

[101] GA Watterson. On the number of segregating sites in genetical models without recombination. *Theoretical population biology*, 7(2):256–276, 1975.

[102] JF Wilkins and J Wakeley. The coalescent in a continuous, finite, linear population. *Genetics*, 161(2):873, 2002.

[103] Thomas Williams, Colin Kelley, and many others. Gnuplot 4.4: an interactive plotting program. `http://gnuplot.sourceforge.net/`, March 2010.

[104] Sewall Wright. Evolution in mendelian populations. *Genetics*, 16(2):97, 1931.

Zeppelinstraße 3
37083 Göttingen
☎ *+49 (0)176 97 85 42 94*
✉ *jens@nld.ds.mpg.de*

Jens Nullmeier

Education

2009–2013	**PhD candidate**, *Max-Planck-institute for Dynamics and Self-Organization, Göttingen*.
2005–2008	**Diploma student**, *Freie Universität*, Berlin, *Diplom–Mathematiker*.
2003–2004	**Diploma student**, *Université Paul Sabatier*, Toulouse.
2001–2005	**Teacher trainee**, *Freie Universität*, Berlin.
1992–2000	**Student**, *Rückert Oberschule*, Berlin, *Abitur and baccalauréat*.

Research interest

Evolution, graph theory

Languages

German, mothertongue

English and French, both fluent

Teaching experience

2005–2008	Math courses for students of economics
2009–2013	Tutorials in ”Mathematik für Physiker 2“, ”Vorkurs: Math. Methoden der Physik“, project tutorial ”PILZ“

Publications and presentations

2010	SMBE, Lyon (poster)
2011	ESEB, Tübingen (poster)
2012	PSE12, Berlin
2012	PopGroup46, Glasgow (talk)
2013	**The Coalescent in Boundary-Limited Range Expansions**, *Evolution*, (publication)