Theoretical Population Biology 153 (2023) 91–101

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

Determining the most recent common ancestor in a finite linear habitat with asymmetric dispersal

Kyle G. Teller [a,*], James M. Pringle [b]

[a] Department of Mathematical Sciences, Salisbury University, Salisbury MD 21801, United States of America
[b] Department of Earth Sciences, and the Institute of Earth, Oceans, and Space, University of New Hampshire, Durham NH 03824, United States of America

ARTICLE INFO

Article history:
Received 22 September 2022
Available online 13 July 2023

Keywords:
Asymmetric dispersal
Coalescence
Marine population genetics
Effective population size
genetic diversity

ABSTRACT

Many species that are birthed in one location and become reproductive in another location can be treated as if in a one-dimensional habitat where dispersal is biased downstream. One example of such is planktonic larvae that disperse in coastal oceans, rivers, and streams. In these habitats, the dynamics of the dispersal are dominated by the movement of offspring in one direction and the distance between parents and offspring in the other direction does not matter. We study an idealized species with non-overlapping generations in a finite linear habitat that has no larval input from outside of the habitat and is therefore isolated from other populations. The most non-realistic assumption that we make is that there are non-overlapping generations, and this is an assumption to be considered in future work. We find that a biased dispersal in the habitat reduces the average time to the most recent common ancestor and causes the average location of the most recent common ancestor to move from the center of the habitat to the upstream edge of the habitat. Due to the decrease in the time to the most recent common ancestor and the shift of the average location to the upstream edge, the effective population size (N_e) no longer depends on the census size and is dependent on the dispersal statistics. We determine the average time and location of the most recent common ancestor as a function of the larval dispersal statistics. The location of the most recent common ancestor becomes independent of the length of the habitat and is only dependent on the location of the upstream edge and the larval dispersal statistics.

1. Introduction

Dispersal has an important role in understanding the genetic connectivity of a population. Low dispersal tends to lead to individuals near each other being more genetically similar than individuals further apart (Wright, 1943). This in turn leads to gene flow restrictions that correlate to the amount of dispersal in the population (Sawyer, 1977). Therefore, understanding the effects of dispersal on the genetic structure will help to make inferences about the overall geographic structure of a population of individuals.

Some of the most well-known migration models are the island model (Maruyama, 1970; Wright, 1943), the steppingstone model (Kimura and Weiss, 1964), the spatial Λ-Fleming–Viot model (Barton et al., 2010), and the continuous, finite, linear population model (Wilkins and Wakeley, 2002). In the island and steppingstone models, the population is broken down into demes or subpopulations that are connected through migration (Fig. 1). In the island model, there is no spatial distribution of offspring, but all larvae migrate from a common pool. In the steppingstone model, demes are aligned linearly, and migration only occurs between adjacent demes. In the island model, larvae are dispersed from a common pool and therefore would not be a good representation of a linear population of individuals. In the stepping-stone model, two assumptions would not work well for a linear population of individuals. The dispersal is assumed to occur between neighboring populations and this exchange is assumed to be isotropic. These assumptions have been relaxed in the literature, but it is unclear how this model can be applied to the distribution of individuals along a linear habitat such as a coastline or river where there are no natural barriers to subdivide the population into subpopulations. The spatial Λ-Fleming–Viot model helps to overcome these limitations by considering a continuous Poisson point process of extinction-recolonization events. The continuous, finite, linear model extends the spatial structure of the steppingstone model where each individual's offspring are allowed to disperse continuously over a linear domain. The individuals are uniformly distributed across the domain and the

* Corresponding author.
E-mail addresses: kgteller@salisbury.edu (K.G. Teller), James.Pringle@unh.edu (J.M. Pringle).

https://doi.org/10.1016/j.tpb.2023.06.003
0040-5809/

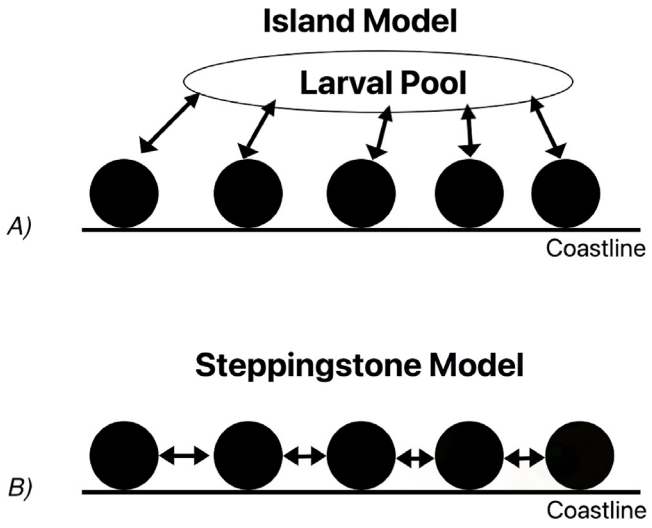

Fig. 1. Based on Selkoe and Toonen, 2011's 2011 models of dispersal. The circles represent sub-populations along a coastline. (A) Island Model: Populations can self-recruit but also contribute to a well-mixed larval pool. (B) Steppingstone Model: Populations tend to exchange migrants between nearest neighbors or a population a certain distance away in both directions (i.e., 2 populations away).

locations of offspring are distributed symmetrically around their parents. Near the edge of the boundaries, the distribution of offspring becomes skewed towards the inside of the domain and the distribution is not fully symmetric but still centered around the parents' location.

We focus on expanding the continuous, finite, linear model of Wilkins and Wakeley (2002) to one that accounts for an asymmetry in the dispersal of offspring; in the model presented here, the average offspring settles to one side, downstream, of its parents. The continuous, finite, linear model described by Wilkins and Wakeley (2002) assumes that every individual produces a very large number of gametes, which are distributed according to a normal distribution around the individual's location with variation σ_m. The boundaries of the continuous, finite, linear model, are assumed to be reflecting, meaning that gametes that land outside the domain would end up the same distance away from the edge but inside the domain. No physical justification is given for these boundaries except that they would maintain a constant population density and allow migration to be conservative. From this model, Wilkins and Wakeley (2002) showed that the coalescent location is biased towards the center of the domain when the migration length is smaller than the length of the domain.

To account for the asymmetry in dispersal the model presented here assumes that individuals produce a very large number of gametes that are moved an average distance downstream from their parents with some variation around this mean. Throughout this paper, our model assumes absorbing boundary conditions, where gametes that land outside the domain do not survive. This is more realistic in most ecosystems where the habitat beyond the extent of the population is unsuitable for survival and there would be high mortality. A constant population density is assumed to be maintained by each adult producing enough offspring so that all of the suitable habitat is occupied. Therefore, the model presented here represents the continuous, finite, linear model with asymmetric dispersal and absorbing boundaries.

One example of where this model would be applicable is marine and riverine environments, where there is physical transport of offspring due to the currents or wind (Cruzan and Hendrickson, 2020; Kling and Ackerly, 2021; Pringle et al., 2011). In these systems, the offspring are transported an average distance in a particular direction (downstream) of their parents, though there can be a substantial deviation from the average (Siegel et al., 2003; White et al., 2010). Thus in these environments, there is a bias in the direction that offspring will be dispersed. This bias in offspring movement will also create a bias in gene flow that will be observable from different measures of genetic diversity (Sundqvist et al., 2016).

One way to interpret the genetic diversity of a population is through coalescent theory and the time to the most recent common ancestor. Coalescent theory has played an important role in population genetics since its introduction in the 1980s by Kingman (1982). Kingman found that a larger population size led to a proportionally larger time to the most recent common ancestor. The Kingman coalescent has been used to develop many models and to extract parameter values but has only recently been applied to models with a spatial distribution of individuals and has been termed the "structured coalescent" (Barton et al., 2010; Wilkins and Wakeley, 2002; Wilkinson-Herbots, 1998). The coalescent is the backward-in-time process that can explain how a common gene in a population originates from a single individual in the past. The coalescent process takes the genes in the current population and models the ancestry back in time to the most recent common ancestor of each of those genes. The time it takes for all individuals that share a common gene from this common ancestor is called the time to the most recent common ancestor (T_{MRCA}), and the physical location of this common ancestor is called the location of the most recent common ancestor (X_{MRCA}). Both the average time to the most recent common ancestor and location play key roles in determining the genetic diversity of the population.

Models of the coalescent take information that we have in the present and use it to constrain the possible ancestry of the individuals in the past. In his original definition of the coalescent, Kingman assumed that there were non-overlapping generations with a constant population size and that an individual's ancestor could have come from anywhere in the domain. Kingman's last assumption means that the location of the parent and offspring are not correlated, however, this is not realistic for spatially extended linear populations with physically realistic dispersal.

Our work expands upon the coalescent process described by Wilkins and Wakeley (2002) to include asymmetric dispersal. Estimates of the average time and location of the most recent common ancestor are derived for a linear population. A numerical simulation is used to validate these estimates. When the asymmetry becomes significant the average location of the most recent common ancestor no longer is dependent on the length of the habitat and is located near the upstream edge. Also, the average time to the most recent common ancestor is reduced to less than what would be expected for a population with isotropic dispersal. When the asymmetry is small, or the spread is large both the time and location of the most recent common ancestor do converge to the results of Wilkins and Wakeley (2002) and Kingman (1982).

2. The model

Many different species have quasi-one-dimensional habitats these could be along coastlines, in rivers, in streams, or in wind-dispersed plants. These are habitats where the length of the habitat is much larger than the width, and the dispersal statistics along the width are constant. This includes species like trees and shrubs, the green crab, barnacles, and other near-shore, coastal, or inter-tidal benthic species (Hyder et al., 1998; Kling and Ackerly, 2021; Siegel et al., 2003; Wares and Pringle, 2008; White et al., 2010). The population that is modeled here would be representative of these species in a finite linear habitat with a current that sends propagules a mean distance downstream with some

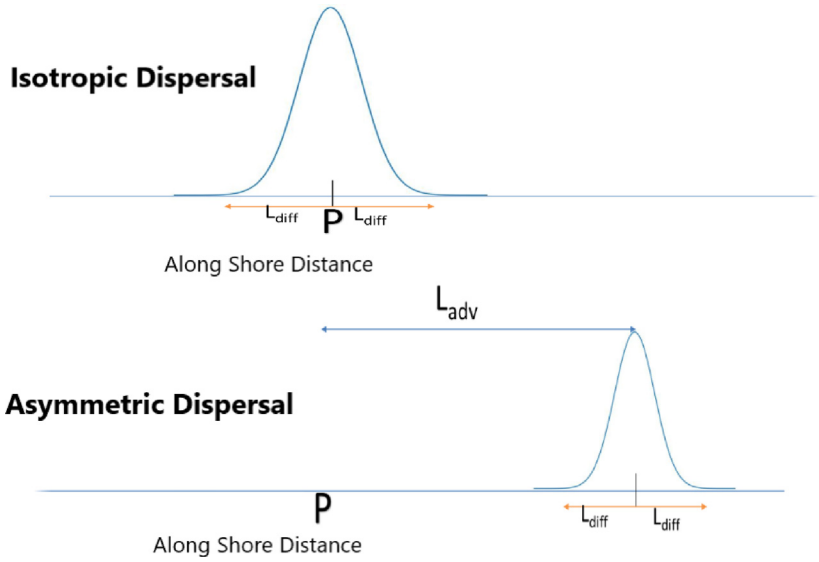

Fig. 2. Depiction of two different dispersal scenarios. In both figures, P is the parent's location and the distribution represents the possible locations of larvae. The top figure is a depiction of isotropic dispersal where larvae are equally likely to settle in either direction away from their parent but most likely near their parent. The bottom figure is a depiction of asymmetric dispersal where larvae are more likely downstream from their parent and have variation around that downstream location. Both figures have variation around a mean labeled L_{diff}, but the bottom figure has a non-zero mean that is termed L_{adv}.

variation around this mean. Competition occurs through habitat limitation with only one individual being able to occupy each habitable location. The model assumes a uniformly distributed population of individuals over a length L. All individuals are haploid with non-overlapping generations. Each adult individual has a location of birth and a location of reproduction. Each individual in the population produces enough larvae such that the population density remains constant in space and time. The boundary conditions are assumed to be absorbing; this means that any larvae that land outside the habitable length of the habitat do not survive.

To model the dynamics of parents and offspring the probability distribution of a parent's location (x_{parent}) given an existing offspring location ($x_{offspring}$) is defined by a truncated Gaussian distribution:

$$f\left(x_{parent}; x_{offspring}, L, L_{adv}, L_{diff}\right) = \frac{\frac{1}{\sqrt{2*\pi}} * e^{-\frac{\left(x_{parent}-(x_{offspring}-L_{adv})\right)^2}{2*L_{diff}^2}}}{L_{diff} * \left[\frac{\left(1+\mathrm{erf}\left(\frac{L-(x_{offspring}-L_{adv})}{L_{diff}*\sqrt{2}}\right)\right)}{2} - \frac{\left(1+\mathrm{erf}\left(\frac{0-x_{offspring}-L_{adv}}{L_{diff}*\sqrt{2}}\right)\right)}{2}\right]}. \quad (1)$$

There are three parameters: L_{adv}, L_{diff}, and L, where L is the length of the habitable domain and goes from 0 to L, L_{adv} is the mean distance downstream larvae travel, and L_{diff} is the stochastic variation of larval dispersal. L_{adv} represents the difference between the mean settlement location of larvae from the location of the parents and includes direction; here a positive L_{adv} is downstream. L_{diff} is the standard deviation of locations where the larvae settle and accounts for the variability in the dispersal of individual larvae (Fig. 2). In coastal oceans and rivers, this stochastic variation is a function of the variation of alongshore currents, the Lagrangian timescale of fluctuations of the mean currents, and the time that the larvae spend in the plankton stage (Siegel et al., 2003). The distribution above is normalized to unity because a parent must have existed in the previous generation. Therefore, probabilities for all possible locations of a parent for a given individual must integrate to 1. If $L_{adv} = 0$ we recover the continuous, finite, linear model of Wilkins and Wakeley (2002), but with absorbing boundary conditions. The results of Wilkins and Wakeley (2002) are not qualitatively different when the boundary conditions are changed from reflecting to absorbing. A Gaussian dispersal kernel shifted a mean distance downstream, as used in our model, is a good representation of the movement of passive planktonic larvae that are dispersed by the currents along coasts and rivers (Siegel et al., 2003).

To implement the model, every individual, in a domain of length L, in the starting generation shares a common gene and is labeled as a separate lineage. Time is stepped backward to determine how all the lineages coalesce into fewer and fewer lineages as time moves into the past. The problem of clumping of individuals in the domain is avoided because there is habitat competition and population density remains constant from generation to generation (Felsenstein, 1975). Each generation backward in time the parent of an individual was found by randomly generating the location of the parent from the probability density function (Eq. (1)) and using the individual's current location in the domain along with the parameters L_{diff} and L_{adv}. When two or more individuals share a parent in the previous generation, a coalescent event has occurred, and the lineages involved reduce to one lineage. The location of each of these events is recorded. Each coalescent event reduces the number of uncoalesced lineages by at least 1. This process is repeated with the remaining lineages until the final coalescent event. The location and generation of this event were recorded. This simulation was then run thousands of times to get an average time and location of the most recent common ancestor.

3. Results

3.1. The source region

In a population with asymmetric dispersal, if we consider any location more than an average dispersal distance from the upstream edge of the habitat, it is both a source of larvae for other locations downstream and a sink for larvae from locations further upstream. Therefore, throughout the domain, most regions are both sources and sinks of genes. The upstream edge of the habitat, however, will act as a pure genetic source, as no larvae can enter the habitat from even further upstream. Furthermore, because

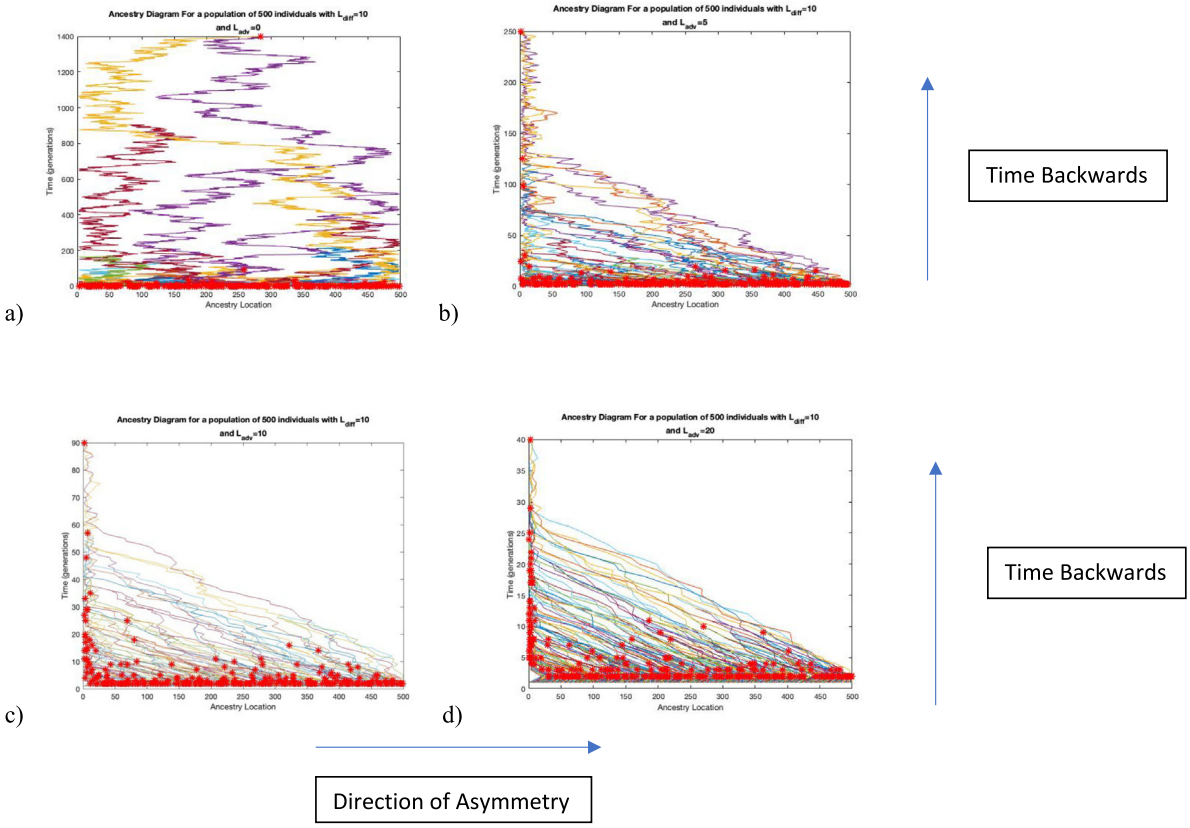

Fig. 3. Ancestry of all individuals in a population of 500 individuals. The red stars represent coalescent events. In Figure a, dispersal is symmetric ($L_{adv} = 0$), while Figures b, c, and d exhibit asymmetric dispersal ($L_{adv} = 5$, 10, and 20 respectively). When there is no asymmetry, the linages occur throughout the domain and it takes a much longer time to get to the most recent common ances, while as the asymmetry increases the lineages end up in the upstream edge. With increasing asymmetry, the number of generations it takes for the asymmetry to cause lineages to be contained in the upstream edge decreases.

the mean currents of the ocean send most propagules further downstream, to maintain the constant population density, this upstream region must be a region that is mostly maintained by the few larvae that originated at the upstream edge and return to the upstream edge. The dynamics at the upstream edge are discussed in detail by Byers and Pringle (2006), Pringle and Wares (2007), Pachepsky et al. (2008). The population range can thus be split into two regions: a region at the upstream edge in which most new recruits are from that region (“the source region”) and the rest of the domain where each location in the habitat is maintained mostly by larvae from even further upstream (the “sink region”) (Doebeli, 1995; Lebreton et al., 2000).

The existence of an upstream source region can be illustrated by tracing the spatial distribution of lineages back in time through the domain. In Fig. 3, there are four ancestry diagrams for a variety of dispersal scenarios with increasing amounts of mean downstream dispersal of larvae (e.g., increasing L_{adv}). In Fig. 3a, there is no asymmetry in the dispersal, and the ancestors of individuals move according to the random distribution defined by equation one. When lineages overlap in a generation, they coalesce, and this coalescence would be expected to occur according to a random distribution in space within the habitat. When the dispersal is made asymmetric in Fig. 2b-d, (L_{adv}>0) as one moves back in time the locations of ancestral lineages are found to move nearer to the upstream edge of the habitat, and the coalescent events are expected to occur at the upstream edge of the habitat. When L_{adv} has increased in the bottom two graphs, the lineages are seen to be constrained to the upstream edge more recently in the past and the lineages originate closer to the upstream edge. This upstream region that is the origin of all lineages in the population is the population source region discussed above.

Since L_{adv} and L_{diff} determine how far away from their parent a larva disperses, the length of the population source region will depend on L_{adv} and L_{diff}. The upstream population source region will be the portion of the domain where the likelihood that larvae are moved downstream out of the domain by the mean currents (L_{adv}) is balanced by the likelihood that the stochastic currents (L_{diff}) return them to this region. For the average currents, the distance that a propagule would travel would be n^*L_{adv} after n generations. Then, since the stochastic spread is a diffusive process it will move a propagule through a length of $n^{0.5} * L_{diff}$ after n generations (Siegel et al., 2003). Therefore, setting the length of movement by the mean and stochastic components of dispersal equal, after $n = \frac{L_{diff}^2}{L_{adv}^2}$ generations on average the movement of propagules by the mean currents and the stochastic variance of dispersal are equal. Therefore using $n = \frac{L_{diff}^2}{L_{adv}^2}$ in either length equation (n^*L_{adv} or $n^{0.5} * L_{diff}$) produces the fundamental length scale (Wares and Pringle, 2008):

$$L_{Source} = \frac{L_{diff}^2}{L_{adv}} \quad . \tag{2}$$

This is the length of the domain over which the mean transport and stochastic components of dispersal take the same amount of time to move a propagule.

3.2. The location of the most recent common ancestor

Since the mean current and diffusion balance each other in the population source region, larvae that settle in the population source region are most likely to have had a parent from

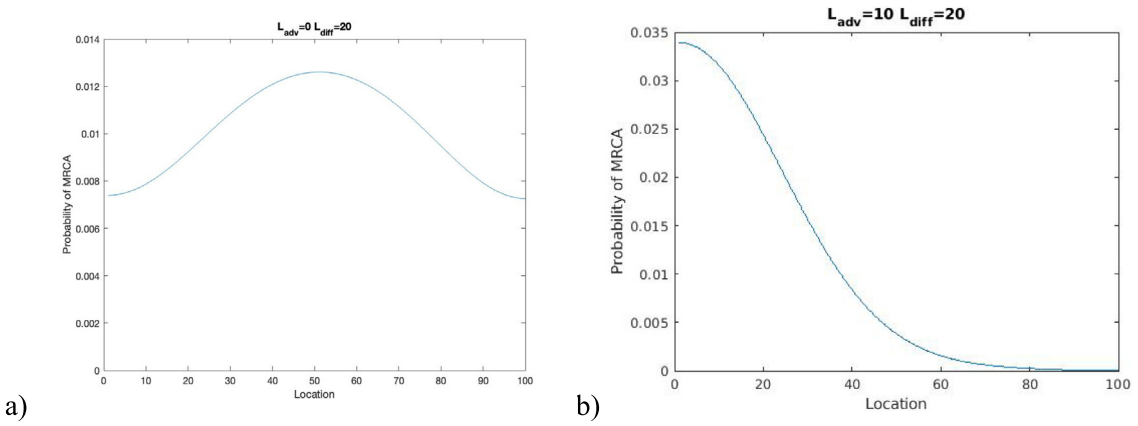

Fig. 4. Probability density function of the location of the most recent common ancestor with and without an asymmetry. The figure on the left is for no asymmetry in the dispersal while the figure on the right includes asymmetric dispersal. When there is no asymmetry, the probability is centered in the middle of the domain, while when there is an asymmetry it is skewed to the left (the upstream edge).

the population source region. In the population source region, there is no larval input from farther upstream since there is no population farther upstream. However, all other regions in the domain receive more larvae from upstream than they do from local production because the dispersal is defined (when L_{adv}>0) to be biased downstream. For the neutral genetics considered here, the populations downstream from the upstream population source region are genetic sinks, and the genes in these regions will over time be dominated by the contributions of migrants from even farther upstream.

Therefore, the population source region is most likely to contain the most recent common ancestor of the population. When looking at a geographically structured population the location of the most recent common ancestor defines the genetic structure of the population and suggests where the genetic diversity of the population originates (Austerlitz et al., 1997). In a finite linear population with symmetric dispersal and nonoverlapping generations, the ancestors of individuals will have a random distribution that is biased towards the center of the domain, as a result, the average location of these events will occur in the middle of the domain (Wilkins and Wakeley, 2002).

In Fig. 4, five thousand runs of the model are compiled to show the spatial probability distribution of the location of the most recent common ancestor of a gene found in the entire population in runs with either symmetric or asymmetric dispersal. When the dispersal of larvae is symmetric the probability peaks at the center of the domain and is distributed symmetrically around the center of the domain. However, when the dispersal is asymmetric the highest probability of the location of the most recent common ancestor occurs near the upstream edge. The probability of a location being the location of the most recent common ancestor diminishes rapidly in the downstream (to the right) direction. This means that the ancestor of an individual living today most likely came from the upstream edge. When there is an asymmetry, the genes of the current population have descended from a smaller population in the population source region.

Figs. 2 and 3 can help explain how genetic diversity is maintained from the population source region. Because the highest probability of the most recent common ancestor is in the population source region (Fig. 4) and all ancestors eventually originate from the population source region (Fig. 4), alleles in that region are more likely to become fixed than in the rest of the habitat. Therefore, most alleles that become fixed in the population come from the population source region and the most recent common ancestor of a gene will most likely be from the population source region.

The location of the most recent common ancestor is biased upstream because dispersal is biased downstream. Each generation into the past, the expected average location of the ancestor of an individual should be a distance of L_{adv} closer to the upstream edge. This phenomenon is shown in Fig. 3; as L_{adv} increases, the location of ancestors becomes closer and closer to the upstream edge. From Fig. 3 when $L_{adv} = 20$ after 20 generations ancestors only come from the upper half of the domain and no further. Further back in time the reduction of ancestors to the upstream edge will increase until all ancestors must have come from the population source region.

The most recent common ancestor is the culmination of several coalescent events. A coalescent event takes place when two or more individuals have the same ancestor in the previous generation. Early coalescent events between individuals are randomly distributed throughout the domain, but when L_{adv}>0, the coalescent location is most likely to be L_{adv} upstream of the individuals whose lineage combines (Fig. 3). Therefore, while coalescent events can take place everywhere in the habitat, they are more likely to occur upstream of the location of the individuals who came from a common parent. As time goes further back and the number of distinct lineages decreases, the influence of the asymmetry will cause the location of coalescent events to shift towards the upstream edge. Since an ancestor is on average a distance L_{adv} upstream the time it takes for all lineages to be contained in the population source region will be dependent on $L_{adv.}$ The average time for lineages to be contained in the upstream edge would be on the order of L/L_{adv} generations. After this time all lineages and coalescent events will be very likely to occur in the population source region (Fig. 3).

Since the length of the population source region is defined by $L_{Source} = \frac{L_{diff}^2}{L_{adv}}$, and it does not receive many migrants, it follows the assumptions of a population described by Wilkins and Wakeley (2002). Therefore, the ancestor of an individual in the population source region will be biased towards the middle of the population source region. The most recent common ancestor of the population must also occur within the population source region because it serves as the source for all lineages throughout the habitat. This suggests that the average location of the most recent common ancestor will come from the middle of the population source region. The resulting expression for the average distance from the upstream edge of the most recent common ancestor is:

$$X_{MRCA} = \frac{L_{Source}}{2} = \frac{L_{diff}^2}{2 * L_{adv}} \tag{3}$$

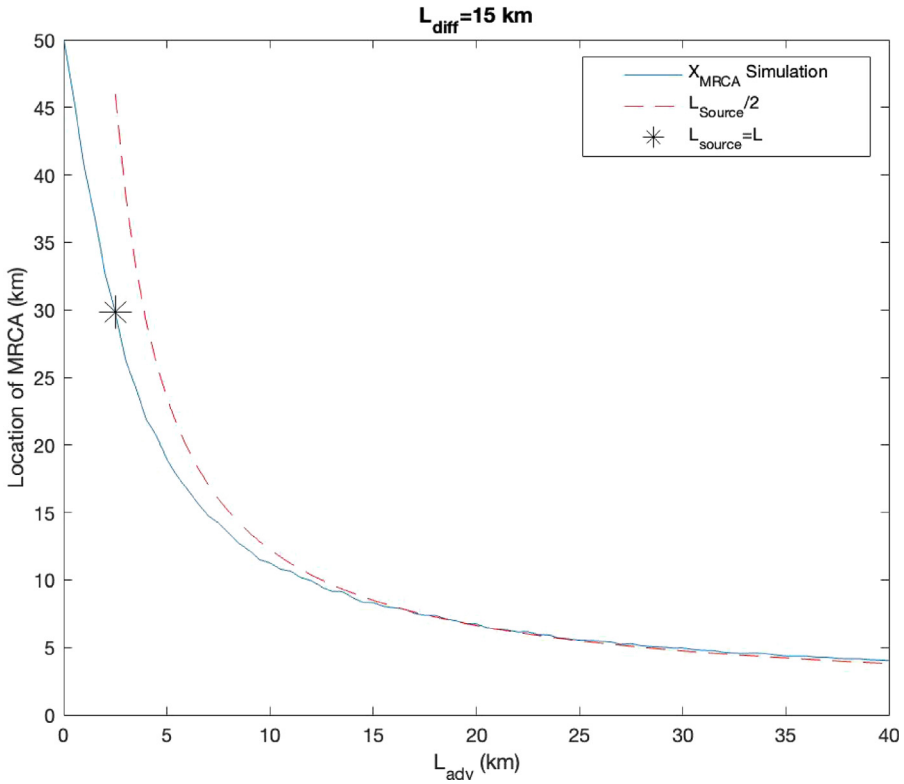

Fig. 5. Simulation of the average coalescent location with changing mean downstream dispersal (L_{adv}) and constant stochastic spread (L_{diff}) in a population of 100 individuals and a domain length of 100. The solid line is data from Monte Carlo simulation runs of the model while the dashed line represents the expression from Eq. (3). For large $L_{adv,}$ this does well to approximate the location but as L_{adv} gets smaller the expression begins to become more and more inaccurate because the length of the source region becomes larger than the length of the habitat. The black star represents when the source region becomes as large as the length of the habitat.

To test the expression for the average location of the most recent common ancestor, simulations of the model were run with 100 individuals each separated by 1 km. The first simulation had a constant L_{diff} of 15 km and an increasing L_{adv} that ranged from 0 km to 40 km (Fig. 5). The second simulation had a constant L_{adv} of 10 km and an increasing L_{diff} that ranged from 5 km to 40 km (Fig. 6). Both sets of simulations were run 5000 times. For these simulations, the average location of the most recent common ancestor was plotted and compared to the estimate for the location of the most recent common ancestor (Eq. (3)). When L_{Source} is less than the length of the habitat L, the asymmetry matters and Eq. (3) is a good first-order predictor of the location of the most recent common ancestor. When L_{Source} is greater than L, it is not. This suggests that when the asymmetry is large enough that the scale for the size of the source region (Eq. (2)) is smaller than the domain size, the genetic diversity of the population is maintained by the individuals in the population source region.

In the limit of small L_{adv} and/or large L_{diff}, the expression for L_{source} (Eq. (2)) gives a length that is larger than the length of the habitat. In these limits (L_{Source}>L) the whole habitat will be the population source region and the effects of dispersal are the same as if it is isotropic, as the stochastic and mean components are balanced. This leads to ancestors being biased towards the center of the domain as in the isotropic dispersal scenario (Fig. 3a and Fig. 4). Thus, the average location of the most recent common ancestor will converge back to the middle of the entire domain when L_{adv} is small, and the results become similar to those in Wilkins and Wakeley (2002).

3.3. Effective population size

In the simplest model of common ancestry with random mating and no population structure, the time to the most recent common ancestor is proportional to the population size (Kingman, 1982). However, many real populations do not align with these assumptions and have some population structure. It is often useful to define the size of the population so that the dynamics are like those found in the simpler model of Kingman (1982). Therefore, in many populations, the effective population size (N_e) can be defined as the size of the population that would give the same time to the most recent common ancestor as a population that obeys the assumptions of Kingman. N_e often does not represent actual individuals in the population but the number of idealized individuals that produce the time to the most recent common ancestor as expected by Kingman. For a population with asymmetric dispersal, this ideal population size N_e reflects the actual size and location of the population within the domain whose descendants will persist in the domain. Since the population source region is the most likely region for the location of the most recent common ancestor the genetics and time to the most recent common ancestor of the entire population will depend on the population source region. Therefore, N_e would be the number of individuals contained in the population source region.

The population of the source region is L_{source}, the length of the region, multiplied by the population density H_{dens} in the number of individuals per unit length, so

$$N_e = \frac{L_{diff}^2}{L_{adv}} * H_{dens} \tag{4}$$

In the limit that L_{adv} goes to zero or as L_{diff} goes to infinity the effective population size will get bigger than the census size (N). Because the N_e derived above represents the actual number of individuals in the population it cannot be larger than the population size. The expression for N_e (Eq. (4)) is only valid when the predicted N_e is less than the total population, or equivalently, that

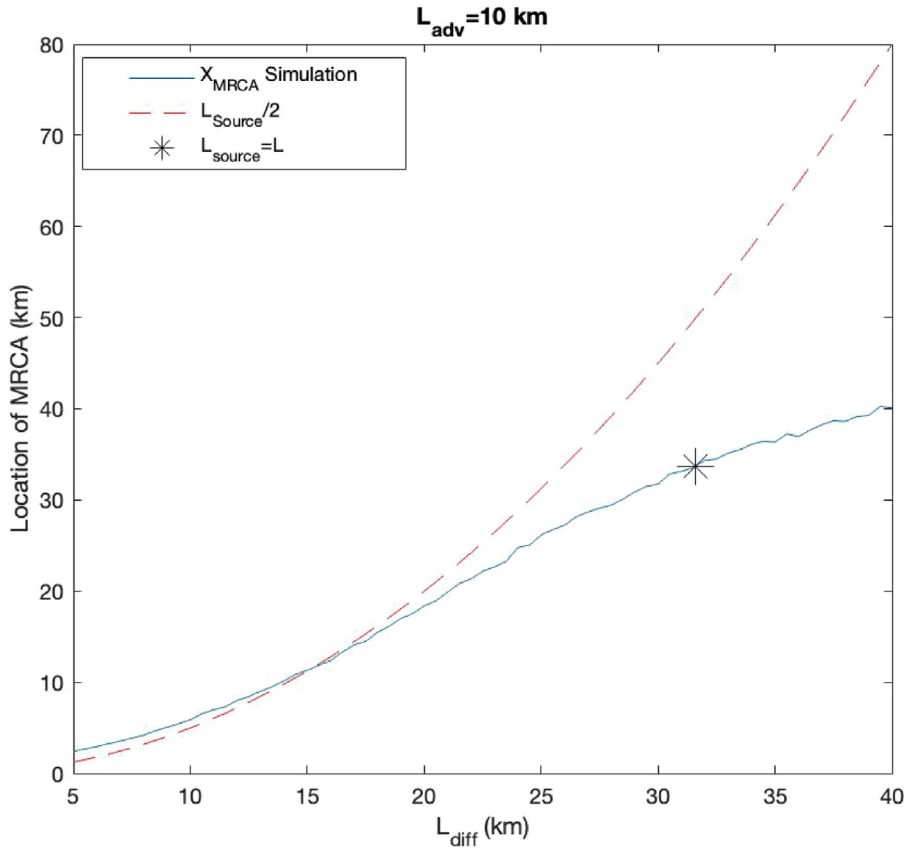

Fig. 6. Simulation of the average coalescent location with changing spread (L_{diff}) and constant mean currents (L_{adv}) in a population of 100 individuals and a domain length of 100. The solid line is data from Monte Carlo simulation runs of the model while the dashed line represents the expression from Eq. (3). For small L_{diff} this does well to approximate the location, but as Ldiff gets larger, the equation becomes more inaccurate due to the source region becoming bigger than the length of the habitat. The black star represents when the source region becomes as large as the length of the habitat.

the L_{source}<L. When the L_{source}>L, the entire population is retained, and N_e is approximately N. This estimate of N_e can be tested by comparing the time to the most recent common ancestor in the model to the time to the most recent common ancestor expected for a population of size $N = N_e$.

3.4. The time to the most recent common ancestor

For many populations, the time to the most recent common ancestor controls the diversity in the population and allows an estimate of the age of evolutionary divergence (Eizirik et al., 2010). The time to the most recent common ancestor can also give insight into the probability of relatedness between individuals given their geographic separation (Neigel et al., 1991).

Kingman assumed that any individual in a population could mate with any other individual, unlike the assumption in this work where mating is with nearby individuals. Given Kingman's assumptions for a haploid population with *N* individuals with random mating, the average time to the most recent common ancestor for a sample of n individuals is 2N*(1-1/n) generations (Kingman, 1982). Although the assumptions of Kingman's model do not fit this model, within the source region the assumptions are nearly met, except that everywhere in the model mating is with adjacent individuals and there are absorbing boundaries. However, in one dispersal event, every individual within the source region could be adjacent to another individual whose parents were somewhere else within the source region. This suggests that the population of the source region, or N_e, could be used to estimate the time to the most recent common ancestor.

When considering the relationship between coalescent times and N_e, a smaller coalescent time will define a N_e that could be smaller than the census population size. Therefore, for a haploid population with non-overlapping generations, the average time to the most recent common ancestor for n individuals would be $2^*N_e(1\text{-}1/n)$ generations. The (1-1/n) accounts for sampling only n individuals in the population. Using the effective population size defined above (Eq. (4)) with the average time to the most recent common ancestor derived by Kingman, the average time to the most recent common ancestor for individuals in the source region for a sample of n individuals is:

$$T_{MRCA} = 2 * \frac{L_{diff}^2}{L_{adv}} * H_{dens}(1 - \frac{1}{n}) \tag{5}$$

when L_{Source}<L. When L_{Source}>L the coalescent time converges to $T_{MRCA} = 2 * N * (1 - \frac{1}{n})$.

The expression above (Eq. (5)) fails to account for the time it takes lineages outside of the source region to migrate backward in time into the source region. For lineages that start furthest from the upstream edge of the domain, this timescale should be on the order of L/L_{adv}. To correct for this time, the time it would take descendants from the source region to move throughout the domain, or equivalently the time in the past it would take lineages from the downstream most part of the population to have come from the source region, Eq. (5) must be increased by L/L_{adv}. The average time to the most recent common ancestor for individuals anywhere in the domain, for a sample of n individuals, should then be:

$$T_{MRCA} = 2 * \frac{L_{diff}^2}{L_{adv}} * H_{dens}\left(1 - \frac{1}{n}\right) + \frac{L}{L_{adv}} \tag{6}$$

To test the expressions for the average time to the most recent common ancestor, simulations of the model were run with 100

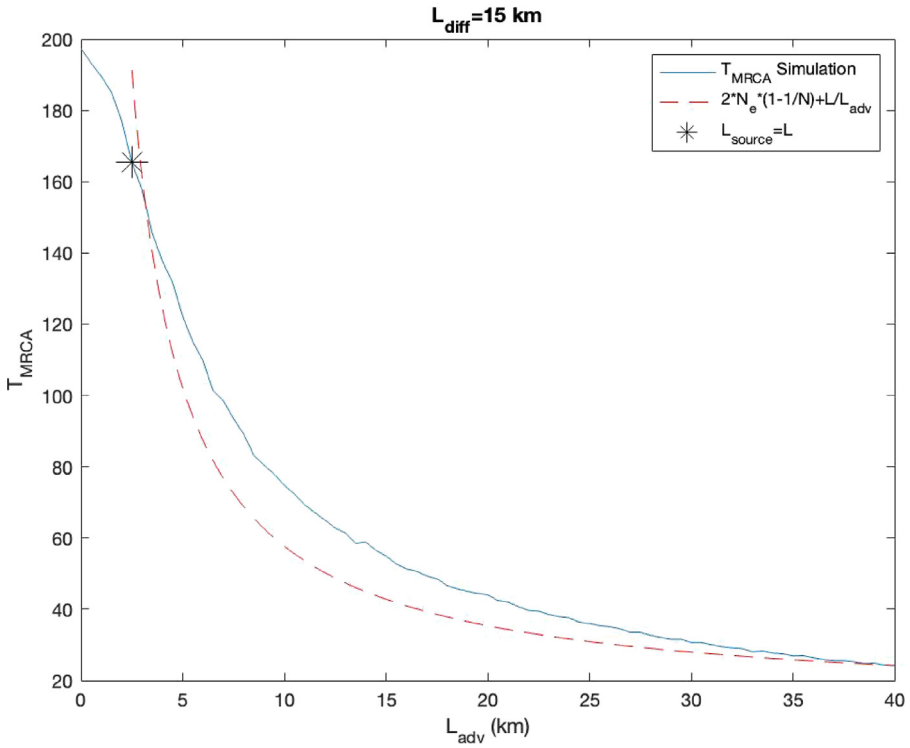

Fig. 7. T_{MRCA} from the model for increasing average dispersal distance (L_{adv}) and constant spread (L_{diff}) in a population of 100 individuals and a domain length of 100. The solid line is data from Monte Carlo simulation runs of the model while the dashed line represents the expression from Eq. (6). The scaling from Eq. (6) matches the first-order behavior of the time to the most recent common ancestor. As L_{adv} becomes smaller the expression for N_e becomes larger than the census population and therefore our scaling begins to break down and the actual time to the most recent common ancestor returns to the classic Kingman result. The black star represents when the effective population becomes larger than the census population.

individuals each separated by 1 km. The first simulation had a constant L_{diff} of 15 km and an increasing L_{adv} that ranged from 0 km to 40 km (Fig. 7). While the second simulation had a constant L_{adv} of 10 km and an increasing L_{diff} that ranged from 5 km to 40 km (Fig. 8). Both simulations were run 5000 times over their respective ranges. For these simulations, the average time to the most recent common ancestor was plotted against the changing parameter and compared to the expression above (Eq. (6)). The expression matches the first-order behavior of the coalescent time. The addition of the L/L_{adv} term makes the expression more accurate when compared to the simulated data by accounting for the average time it takes for lineages to reach the source region. As with the location of the most recent common ancestor, the expression for the time to the most recent common ancestor (Eq. (6)) breaks down when the length of the source region is greater than the length of the habitat. When the length of the source region does become larger than the length of the habitat, every individual in the habitat nearly conforms to the original assumptions of Kingman. Therefore, the time to the most recent common ancestor returns to $2N^*(1\text{-}1/n)$. Overall, when the asymmetry causes L_{source} to be less than the length of the habitat the time to the most recent common ancestor decreases and is dependent on the dispersal statistics and the population density.

3.5. Effect of limited population size in model on results

In many natural populations, there are thousands to millions of individuals in a habitat. This, however, is not feasible to replicate in our numerical models due to the time it would take to track the genetics of the population. Therefore, the numerical model used to validate our estimates of the time to the most recent common ancestor and the location of the most recent common ancestor is run with a much smaller population size than would be expected for most species to which this model is applicable. The population must be kept small because the time it takes to make a model run scales as the number of individuals (N). The time for the model to reach equilibrium scales as *N* which leads to a computer run time that scales as N^2. This time becomes impracticable for large N.

However, the estimates discussed above are still valid. The expression for the location of the most recent common ancestor (Eq. (3)) does not depend on population size while the expression for the time to the most recent common ancestor (Eq. (4)) accounts for the population size in the population density (H_{dens}) term. In Fig. 9, an ancestry diagram for a population of 500 individuals and a population of 1000 individuals is shown with the same L_{adv} and L_{diff} and population densities of one. This figure illustrates the effect of population size on the time to the most recent common ancestor (T_{MRCA}) and the lack of dependence on the location of the most recent common ancestor. The location of the final coalescent event and the length of the population source region do not change between the two populations. The expressions for the average location of the most recent common ancestor do not depend on population size and only depend on the sizes of L_{adv} and L_{diff}. T_{MRCA} has been shown above to scale proportionally with the effective population size. In Fig. 10, the time to the most recent common ancestor for two populations is shown one with an $H_{dens} = 1$ and the other with an $H_{dens} = 2$. Just as with the previous figure the population with a $H_{dens} = 2$ has a time to most recent common ancestor roughly twice the time for the population with $H_{dens} = 1$. Therefore, the addition of H_{dens} in the expression for the average time to the most recent common ancestor (Eq. (4)) allows this expression to work for any population size.

Population density also affects the rate at which single coalescent events occur. The average rate of coalescent events will

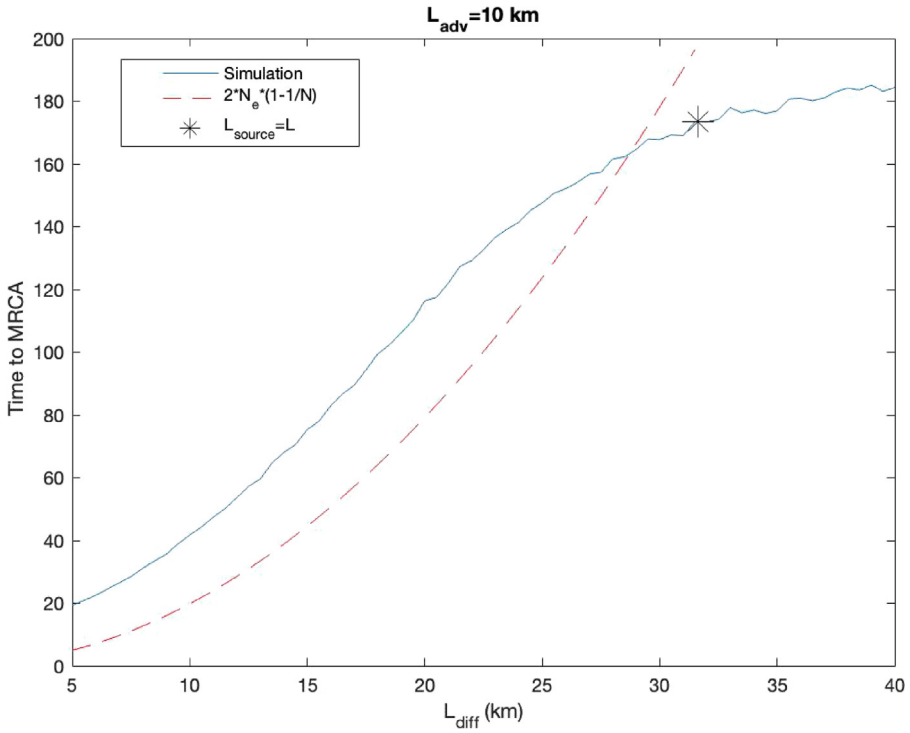

Fig. 8. Simulation of T_{MRCA} with increasing spread (L_{diff}) and constant average current distance (L_{adv})) in a population of 100 individuals and a domain length of 100 km. The solid line is data from simulation runs of the model while the dashed line represents the expression from Eq. (6). The scaling from Eq. (6) matches the first-order behavior of the time to the most recent common ancestor. As L_{diff} gets large the expression for N_e becomes larger than the census population and therefore our scaling begins to break down and the actual time to the most recent common ancestor returns to the classic Kingman result. The black star represents when the effective population becomes larger than the census population.

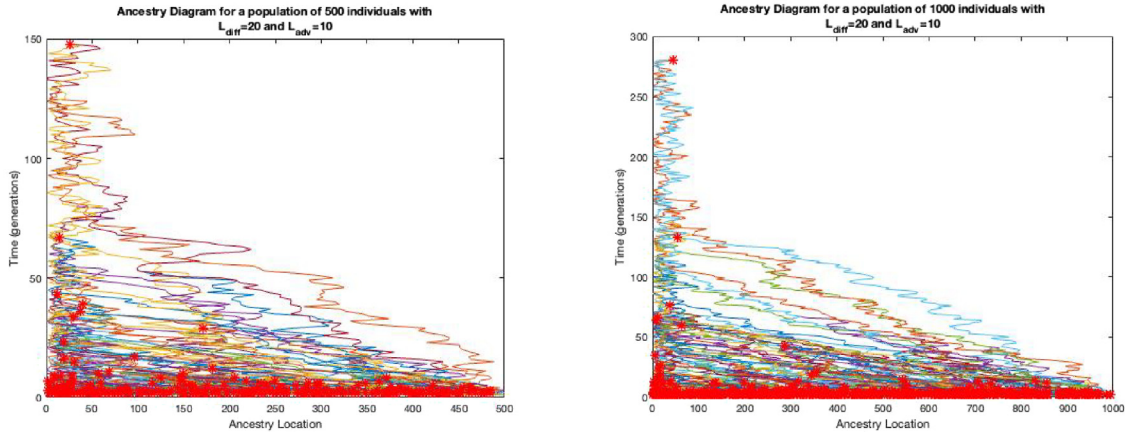

Fig. 9. Ancestry diagram of two populations with different sizes but the same average currents distance (L_{adv}) spread (L_{diff}) and population density ($H_{dens} = 1$). Red stars represent coalescent events. The location of the most recent common ancestor occurs in the same region regardless of population size. With a larger population size, the time to the most recent common ancestor is larger.

depend on two things, the average time to the most recent common ancestor and how many pairs of individuals can coalesce. Dividing the number of pairs of individuals that can coalesce by the average time to the most recent common ancestor gives the average rate of coalescent events. From Eq. (6) the time to the most recent common ancestor scales proportionally to the number of individuals in the population. However, the number of pairs of individuals that coalesce is proportional to the number of individuals squared. This means that the average rate at which single coalescent events occur is approximately proportional to the number of individuals and therefore the population density. This can be seen in Fig. 10, as the population density increases the total time to the most recent common ancestor increases but so does the rate at which coalescent events occur. When $H_{dens} = 2$ all but three coalescent events occur within the first 50 generations. When $H_{dense} = 1$ it takes longer than 50 generations for there to be only three coalescent events left. Both the average time to the most recent common ancestor and the rate of coalescent events will scale with the population density.

4. Discussion

When larval dispersal is asymmetric, due to currents in the ocean or the winds in the atmosphere, we show that there exists an upstream source region for larvae that impacts the time to the most recent common ancestor, the effective population size,

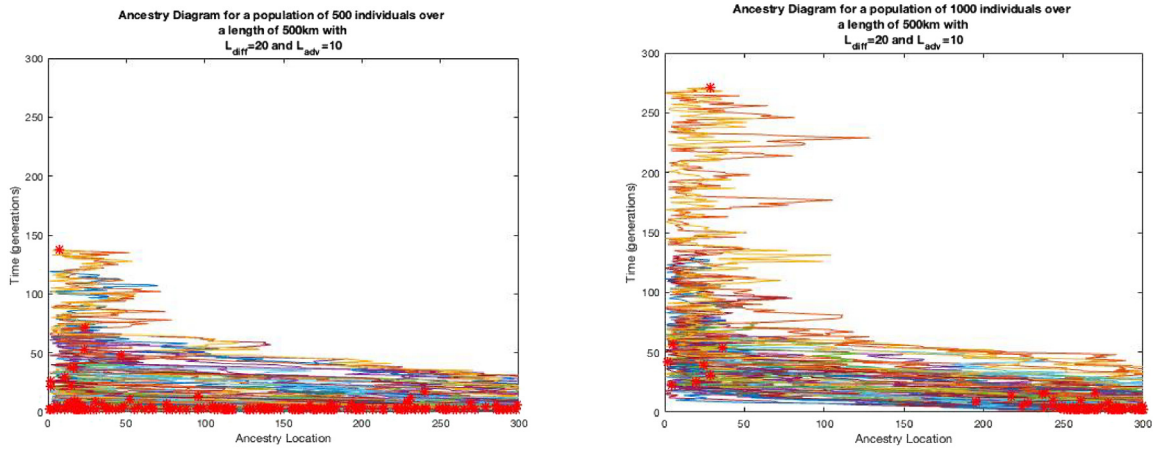

Fig. 10. Ancestry diagram of two populations with different population densities but the same average currents distance (L_{adv}), spread (L_{diff}), and length of the domain. The population on the left has an $H_{dens} = 1$ and the population on the right has an $H_{dens} = 2$. The red stars represent coalescent events. The location of the most recent common ancestor occurs in the same region regardless of population size, while the time to the most recent common ancestor grows with increased population size.

and location of the most recent common ancestor. The size of the population source region (L_{Source}) is a function of the average distance larvae flow downstream (L_{adv}) and the standard deviation of dispersal (L_{diff}). L_{Source} is the distance from the upstream edge over which the location of the most recent common ancestor can be found, and the average location is in the middle of this region. Based on L_{Source} and the population density (H_{dens}) the effective population size (N_e) can be determined. N_e can then be used to determine a scaling for the average time to the most recent common ancestor (T_{MRCA}) for the population.

4.1. Importance of the population source region

In the case of asymmetric larval dispersal, alleles that are present in the population are most likely to have come from the population source region. Therefore, a novel allele that is favored in the population source region and not through the entire habitat will tend to be retained. On the other hand, a novel allele that is favored outside the population source region will most likely be lost since this region is a sink for the population source region. Thus, alleles that are favored in the upstream edge will be favored throughout the entire domain (Kawecki and Holt, 2002). For example, if a population of individuals with larvae that are dispersed downstream is adapted to temperatures along a certain section of the coast, individuals that are moved away from that section of the coast may die out if they are moved into a region with water that is too warm or cold for the larvae to function.

One example of the importance of the population source region containing favored alleles is when climate change shifts the habitat range. Fuchs et al. (2020) have shown that benthic organisms along the Northwest Atlantic continental shelf are being transported to warmer regions where there is a high mortality. The reason for this is that transport is sending larvae farther southward than expected. This has led to a shrinking of the occupied range of the species in this region. Since larvae are being sent southward by the currents, the northern end of the domain is the upstream edge. For the population to persist alleles must be present in the upstream edge that favors these newer warmer waters. When these alleles are not present in the upstream edge the length of the habitat will shrink until the number of successfully recruiting larvae meets the threshold described by Byers and Pringle (2006) that depends on the population source region.

Although the population source region can hinder local adaptation in the sink region, selection can still play a role in allowing local adaptation. When selection is strongly favored for an allele outside the population source region it can persist throughout the population (Pringle and Wares, 2007). This selection can be due to changes in environmental conditions, the invasion of another species, or any other factor that may make a heritable trait more favored. One example of this that has been shown to be true is in the European green crab, *Carcinus maenas* (Tepolt and Palumbi, 2020). Tepolt and Palumbi have shown that the European green crab introduced along the east coast of North America had the standing genetic variation for different temperature tolerances. Although the upstream edge of the habitat would have a selection that is strongly favored for cold tolerance the green crab along the coast has been able to inherit genes for a range of temperature tolerances. This could imply that selection for different water temperature tolerance has been large enough that alleles that are not favored in the upstream edge have been able to persist throughout the population. Therefore, there is a spatial balancing selection occurring where selection for different temperature tolerances is being maintained where the green crab has invaded. The ability of these multiple alleles to be present in the European green crab has made it such a successful invasive species.

4.2. Implications on effective population size

The expression for N_e (Eq. (4)) derived above illustrates that the statistics of an asymmetric dispersal kernel can influence N_e and therefore genetic diversity (Wright, 1931). N_e and the population's genetic diversity are reduced when the dispersal is asymmetric. In many coastal oceanic species, the ratio of effective population size to census size is small (Hedrick, 2005). There are many reasons why this might be true such as sex ratios, selection, and variance in the reproductive success (Wang and Caballero, 1999). The expression for N_e derived in the results gives a new reason why N_e might be reduced in marine environments with asymmetric dispersal. When L_{source} is less than the length of the habitat then the effective population size is reduced to the individuals contained in the population source region. Therefore, the small N_e to N ratios seen in many oceanic species like damselfish and angelfish could be due, in part, to the mean downstream transport of larvae by currents in the ocean (Crane et al., 2018).

For many benthic species, the size of L_{adv} and L_{diff} will roughly be comparable (Pringle et al., 2011; Siegel et al., 2003). Therefore, the length of the population source region will be one dispersal distance away from the upstream edge. The length of the habitat

for a given population is almost always larger than one dispersal distance (Trewhella et al., 1988). Thus, many benthic marine species that have, essentially, a one-dimensional habitat structure will have reduced times to the most recent common ancestor and the location of the most recent common ancestor will occur in a region one dispersal distance away from the upstream edge. This leads to many of these species having small N_e to N ratios and lower genetic diversity than expected for their census population size.

Data availability

All the Matlab functions as well as a Read_me.txt file are available as Supplementary file.

Acknowledgments

This work was supported by National Science Foundation, United States of America, grants 1947954 and 1459609 to James Pringle. This work was greatly improved by the helpful feedback of two reviewers and the editor.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.tpb.2023.06.003.

References

Austerlitz, F., Jung-Muller, B., Godelle, B., Gouyon, P.-H., 1997. Evolution of coalescence times, genetic diversity and structure during colonization. Theor. Popul. Biol. 51 (2), 148–164. http://dx.doi.org/10.1006/tpbi.1997.1302.

Barton, N., Etheridge, A., Véber, A., 2010. A new model for evolution in a spatial continuum. Electron. J. Probab. 15, 162–216. http://dx.doi.org/10.1214/EJP.v15-741.

Byers, J., Pringle, J., 2006. Going against the flow: Retention range limits and invasions in advective environments. Mar. Ecol. Progress Ser. 313, 27–41. http://dx.doi.org/10.3354/meps313027.

Crane, N.L., Tariel, J., Caselle, J.E., Friedlander, A.M., Robertson, D.R., Bernardi, G., 2018. Clipperton atoll as a model to study small marine populations: Endemism and the genomic consequences of small population size. PLoS One 13 (6), e0198901. http://dx.doi.org/10.1371/journal.pone.0198901.

Cruzan, M.B., Hendrickson, E.C., 2020. Landscape genetics of plants: Challenges and opportunities. Plant Commun. 1 (6), 100100. http://dx.doi.org/10.1016/j.xplc.2020.100100.

Doebeli, M., 1995. Dispersal and dynamics. Theor. Popul. Biol. 47 (1), 82–106. http://dx.doi.org/10.1006/tpbi.1995.1004.

Eizirik, E., Murphy, W.J., Koepfli, K.-P., Johnson, W.E., Dragoo, J.W., Wayne, R.K., O'Brien, S.J., 2010. Pattern and timing of diversification of the mammalian order Carnivora inferred from multiple nuclear gene sequences. Mol. Phylogen. Evol. 56 (1), 49–63. http://dx.doi.org/10.1016/j.ympev.2010.01.033.

Felsenstein, J., 1975. A pain in the torus: Some difficulties with models of isolation by distance. Amer. Nat. 109 (967), 359–368.

Fuchs, H.L., Chant, R.J., Hunter, E.J., Curchitser, E.N., Gerbi, G.P., Chen, E.Y., 2020. Wrong-way migrations of benthic species driven by ocean warming and larval transport. Nature Climate Change 10 (11), http://dx.doi.org/10.1038/s41558-020-0894-x, Article 11.

Hedrick, P., 2005. Large variance in reproductive success and the Ne/N ratio. Evolution 59 (7), 1596–1599. http://dx.doi.org/10.1111/j.0014-3820.2005.tb01809.

Hyder, K., Johnson, M., Hawkins, S., Gurney, W., 1998. Barnacle demography:evidence for an existing model and spatial scales of variation. Mar. Ecol. Prog. Ser. 174, 89–99. http://dx.doi.org/10.3354/meps174089.x.

Kawecki, T.J., Holt, R.D., 2002. Evolutionary consequences of asymmetric dispersal rates. Am. Nat. 160 (3), 333. http://dx.doi.org/10.1086/341519.

Kimura, M., Weiss, G.H., 1964. The Stepping Stone Model of Population Structure and the Decrease of Genetic Correlation with Distance. Genetics 49 (4), 561–576.

Kingman, J.F.C., 1982. The coalescent. Stochastic Process. Appl. 13 (3), 235–248. http://dx.doi.org/10.1016/0304-4149(82)90011-4.

Kling, M.M., Ackerly, D.D., 2021. Global wind patterns shape genetic differentiation, asymmetric gene flow, and genetic diversity in trees. Proc. Natl. Acad. Sci. 118 (17), http://dx.doi.org/10.1073/pnas.2017317118.

Lebreton, J.-D., Khaladi, M., Grosbois, V., 2000. An explicit approach to evolutionarily stable dispersal strategies: No cost of dispersal. Math. Biosci. 165 (2), 163–176. http://dx.doi.org/10.1016/S0025-5564(00)00016-X.

Maruyama, T., 1970. On the fixation probability of mutant genes in a subdivided population. Genet. Res. 15 (2), 221–225. http://dx.doi.org/10.1017/S0016672300001543.

Neigel, J.E., Ball, R.M., Avise, J.C., 1991. Estimation of single generation migration distances from geographic variation in animal mitochondrial DNA. Evol.; Int. J. Organic Evol. 45 (2), 423–432. http://dx.doi.org/10.1111/j.1558-5646.1991.tb04415.x.

Pachepsky, E., Nisbet, R.M., Murdoch, W.W., 2008. Between discrete and continuous: Consumer-resource dynamics with synchronized reproduction. Ecology 89 (1), 280–288. http://dx.doi.org/10.1890/07-0641.1.

Pringle, J.M., Blakeslee, A.M.H., Byers, J.E., Roman, J., 2011. Asymmetric dispersal allows an upstream region to control population structure throughout a species' range. Proc. Natl. Acad. Sci. 108 (37), 15288–15293. http://dx.doi.org/10.1073/pnas.1100473108.

Pringle, J.M., Wares, J.P., 2007. Going against the flow: Maintenance of alongshore variation in allele frequency in a coastal ocean. Mar. Ecol. Prog. Ser. 335, 69–84. http://dx.doi.org/10.3354/meps335069.

Sawyer, S., 1977. Asymptotic properties of the equilibrium probability of identity in a geographically structured population. Adv. Appl. Probab. 9 (1977), 268–282.

Selkoe, K., Toonen, R., 2011. Marine connectivity: A new look at pelagic larval duration and genetic metrics of dispersal. Mar. Ecol. Prog. Ser. 436, 291–305. http://dx.doi.org/10.3354/meps09238.

Siegel, D.A., Kinlan, B.P., Gaylord, B., Gaines, S.D., 2003. Lagrangian descriptions of marine larval dispersion. Mar. Ecol. Prog. Ser. 260, 83–96. http://dx.doi.org/10.3354/meps260083.

Sundqvist, L., Keenan, K., Zackrisson, M., Prodöhl, P., Kleinhans, D., 2016. Directional genetic differentiation and relative migration. Ecol. Evol. 6 (11), 3461–3475. http://dx.doi.org/10.1002/ece3.2096.

Tepolt, C.K., Palumbi, S.R., 2020. Rapid adaptation to temperature via a potential Genomic Island of divergence in the Invasive Green Crab, Carcinus maenas. Front. Ecol. Evol. 8, http://dx.doi.org/10.3389/fevo.2020.580701.

Trewhella, W.J., Harris, S., McAllister, F.E., 1988. Dispersal distance, home-range size and population density in the red fox (Vulpes vulpes): A quantitative analysis. J. Appl. Ecol. 25 (2), 423–434. http://dx.doi.org/10.2307/2403834.

Wang, J., Caballero, A., 1999. Developments in predicting the effective size of subdivided populations. Heredity 82 (2), 212–226. http://dx.doi.org/10.1038/sj.hdy.6884670.

Wares, J.P., Pringle, J.M., 2008. Drift by drift: Effective population size is limited by advection. BMC Evol. Biol. 8 (1), 235. http://dx.doi.org/10.1186/1471-2148-8-235.

White, C., Selkoe, K.A., Watson, J., Siegel, D.A., Zacherl, D.C., Toonen, R.J., 2010. Ocean currents help explain population genetic structure. Proc. R. Soc. B Biol. Sci. 277 (1688), 1685–1694. http://dx.doi.org/10.1098/rspb.2009.2214.

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

Wilkinson-Herbots, H.M., 1998. Genealogy and subpopulation differentiation under various models of population structure. J. Math. Biol. 37 (6), 535–585. http://dx.doi.org/10.1007/s002850050140.

Wright, S., 1931. Evolution in mendelian populations. Genetics 16 (2), 97–159.

Wright, S., 1943. Isolation by distance. Genetics 28 (2), 114–138.