Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Akash Arani April 9th 2021

Allele Frequency Spectra for 1-Dimensional Circular Populations

by

Akash Arani

Daniel B. Weissman
Adviser

Physics

Daniel. B Weissman
Adviser

Katharina V. Koelle
Committee Member

John Lindo
Committee Member

2021

Allele Frequency Spectra for 1-Dimensional Circular Populations

By

Akash Arani

Daniel B. Weissman
Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Physics

2021

Abstract

Allele Frequency Spectra for 1-Dimensional Circular Populations
By Akash Arani

The natural limit on geographic genetic dispersal given through isolation by distance encourages the use of spatial structure to analyze genetic diversity, commonly analyzed using allele frequency spectra. This paper aims to analyze the spectra of one of the simplest forms of discrete population structure: 1-dimensional circular populations. Msprime was used to model the genetic variation of a 1-dimensional circular population and create genealogical trees using coalescent theory. As mutations were assumed to have no effect on fitness, neutral mutations were applied post coalescent simulation. These genealogical trees were then converted to allele frequency spectra by counting the number of single nucleotide polymorphisms (SNPS) per branch group.

The results focused on strong structured 1-dimensional circular populations, where the configuration of the demes and the intra-deme migration rate had a significant impact on genetic variation. Simulation results showed that the expected allele frequency spectra of strong-structured 1-dimensional circular populations matched the predicted allele frequency spectra for populations of census size at low frequencies and transitioned to match the predicted allele frequency spectra for populations of effective size at higher frequencies. These results indicate that 1-dimensional circular populations undergo changes in genetic variation at different frequencies and populations need to be deeply sampled to analyze a population's genetic variation. Further research is needed to identify if this behavior translates to higher-dimensional structures and continuous space.

Allele Frequency Spectra for 1-Dimensional Circular Populations

By

Akash Arani

Daniel B. Weissman
Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Physics

2021

Table of Contents

Introduction

The theory of isolation by distance [1] suggests mating pairs generally tend to live not only near to each other but near to their eventual offspring as well. This implies that individuals that live near each other are more genetically similar than individuals that live further away. The isolation by distance model has been used to describe the evolutionary trends of several organisms, including humans. One such trend was analyzed on human inhabitants of the Sanday, Orkney Islands whereby using genealogical data, it was concluded that the degree of genetic relatedness declined as geographic distance increased [2]. This naturally limited geographical dispersal of genes has allowed scientists to discretize natural populations and create spatial structures, such as the stepping stone model [3] or the large metapopulation model [4], to make assumptions about the real world. This paper analyzes one of simplest forms of spatial structure in 1-dimensional circular populations.

Allele frequency spectra (AFS) are powerful tools used to describe genetic variation present in a population. AFS are created by taking sequences of samples within a population and comparing their genetic make up to an ancestral reference sequence. Differences between the specific sample sequences and the ancestral sequences at specific sites on the ancestral genome are noted and referred to as Single Nucleotide Polymorphisms (SNPs). The number of SNPs for each site on the ancestral genome is referred to as the allele frequency. For example, if only three samples out of 3000 differ from the ancestral sequence at a specific site, the allele frequency for that specific site is 3/3000= 0.1%. The distribution of these allele frequencies is what is plotted as the allele frequency spectrum.

One way to model the effect of spatial structure on the evolution of a population is to attempt to find an effective population size [5]. The effective population size is the size of an idealized population that experiences the same rate of loss in genetic diversity as an observed population. In the context of spatially structured populations, the effective population size is the size that a well-mixed population would have to match the genetic diversity of a spatially structured population. Effective population sizes have been used in population biology to understand the behavior of important population biology concepts such as heterozygosity through f-statistics [6] and important models such as the Moran model [7]. For a one-dimensional circular population, Maruyama [8] found simple approximate expressions for the effective population size in the limits of strong and weak spatial structure, which will be discussed in the Model section. Finite circular 1-dimensional populations, along with their linear offshoots, are popular choices of structure to study as they are easy to visualize and yet, are rich with population dynamics [8,9];[10].

Coalescent theory is a powerful tool used to model the genealogy of a population. Originally created by Kingman[11], coalescent theory looks at the entire gene pool and alleles available in a given population and attempts to trace the lineages of these alleles back in time. This paper uses the Wright-Fisher model where the population is well-mixed and stays constant at size N each generation through immediate replacement of dead individuals by their offspring[12]. If we consider a population of N diploid individuals (two alleles per individual), the probability that a random allele in the current generation has the same parent as another random allele in the previous generation, defined as the probability of coalescence one generation ago, is $= \frac{1}{2N}$ [13]. This means that the probability that two alleles do not share a parent in the previous generation, or do not coalesce in the previous

generation, is $= 1 - \frac{1}{2N}$, implying that the probability of coalescence between two inhabitants in t generations is geometrically distributed via P (coalescence t generations ago) = $(1 - \frac{1}{2N})^{t-1} \frac{1}{2N}$ [13]. For large values of N, this probability approximates to P (coalescence t generations ago) = $\frac{1}{2N} e^{(1-t)/2N}$, giving an expected time to coalescence = 2N [13]. Modelling this distribution on a genealogical tree allows us to trace back ancestry and effectively model the genealogy of a population.

Wakeley's observation of large metapopulations states that populations structures that are subdivided into discrete demes with equal density, coalesce through a scattering-collecting decomposition. The scattering phase represents the most recent phase, where inhabitants either coalesce intra-deme or their lineages moves away from the original demes they were sampled in[14]. It is important to note that in the scattering phase, individuals sampled from different demes do not coalesce [15]. As the scattering phase is the most recent phase and probability of coalescence decreases with time as P (coalescence t generations ago) = $(1 - \frac{1}{2N})^{t-1} \frac{1}{2N}$, the fastest coalescence events happen in the scattering phase. Once all the samples have either coalesced intra-deme or the each of the lineages as moved to a different deme, the collecting phase begins as seen in Fig 1. The collecting phase represents the time frame when pairs of lineages undergo large migration events to eventually collect together in a deme to coalesce. The collecting phase ends when the most recent common ancestor for the entire sample is found. Genetic diversity in this phase is modeled by a well-mixed population of effective size N_e.

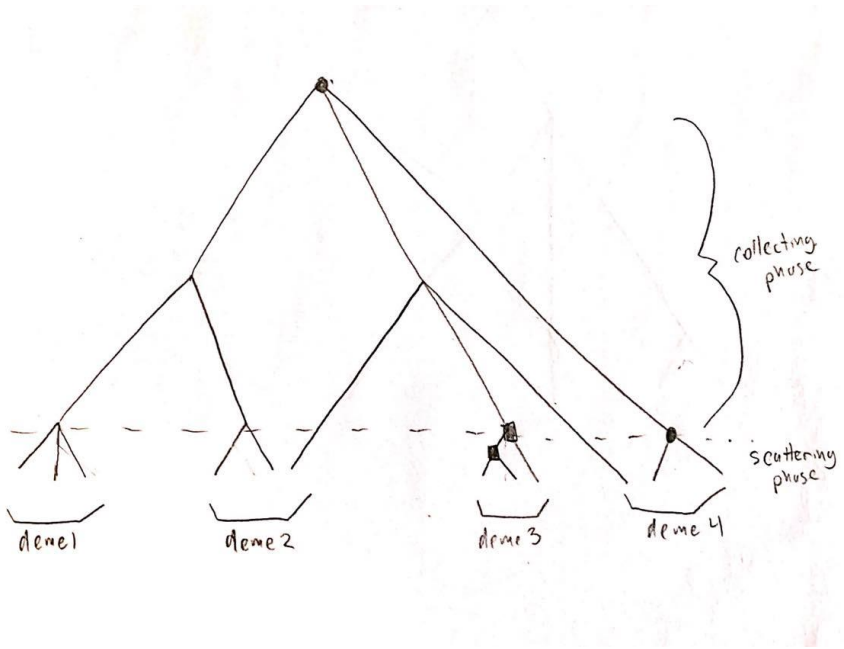

Fig 1: This figure represents an example of a coalescent tree for a structured population with four demes, each with three inhabitants. The genealogical history of the tree is separated via the scattering-collecting decomposition. The scattering phase consists of intra-deme coalescence and lineages moving to different demes. The collecting phase begins when all the lineages are in different demes and coalescence only occurs due to inter-deme migration events that bring lineages together. In deme 3, all three inhabitants coalesce intra-deme via the two coalescent events indicated by squares, signifying the end of its scattering phase. For deme 4 however, the scattering phase ends with one coalescent event and one migration event where a sample's lineage moves to a different deme. The circular nodes signify the two coalescent events that occur with respect to the inhabitants in deme 4, where the more recent coalescent happens intra-deme in the scattering phase while the more ancient coalescent happens at the end of the collecting phase with the most recent common ancestor for the entire population.

This paper aims to combine the concept of scattering-collecting decomposition with Maruyama's work on 1-dimensional circular populations to explain the behavior of said populations through an allele frequency spectrum and understand how they conform to a census population size (N) vs an effective population size (N_e).

Model

For this paper, there are three important variables that will be modified to understand 1-dimensional circular populations:

1) Migration rate: m, defined as the probability an inhabitant leaves its current deme for the neighboring demes per generation.
2) Deme density: ρ, defined as the number of diploid inhabitants per deme.
3) Length of population: L, defined as the total number of demes in the population.

which define the census population size – defined as the total number of individuals in a spatially structured population given by size N= Lρ, and the effective population size – defined as the size of an idealized well-mixed population used to describe the genetic diversity of a spatially structured population, given by size N_e.

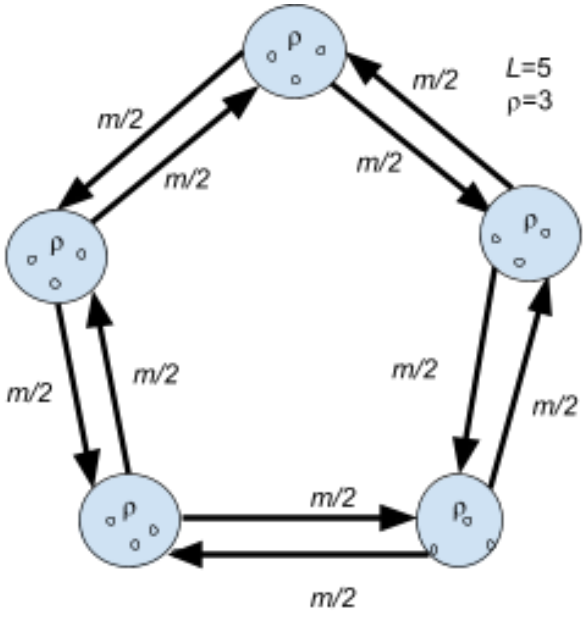

Fig 2: Example of a circular population. There are L=5 demes, each with ρ=3 diploid inhabitants. Individuals migrate to neighboring demes at rate m each generation. Total number of individuals in the entire population = deme density * number of demes -> N= ρL

All coalescent simulations were done using msprime [16] with diploid populations. Msprime was used as it allowed genealogical trees to be simulated under the spatial structure shown in Fig 2. For all populations, we simulate exhaustive sampling, in which the entire population is sampled. Mutations were assumed to have no effect on the fitness of the individuals in the population; therefore, no actual stochastic mutations were applied; however, as the genealogical trees used by msprime are measured in units of branch length, the expected mutation rate per generation was set equal to 1 for simplicity's sake. Therefore, the mutation parameter for the effective population is $\Theta_e = 4N_e\mu = 4N_e$[4] and $\theta = 4N\mu = 4N$ for the census population.

While the genealogical tree gives information on the parental history of every node it does not give us information on the genetic polymorphism of the population. This is why expected allele frequency to evaluate genetic variation. To calculate the allele frequency spectrum for a well-mixed population of given size $N = L\rho$, we use $\text{SNP count} = \frac{\theta}{i}$[13], where i is the corresponding allele count. As mutations are applied per branch length, the allele frequency spectra produced in this paper are expected allele frequency spectra.

The allele frequency spectrum for the simulation is calculated from the genealogical tree by first realizing that as mutations are applied in units of branch length, the number of mutations per branch depends on the length of the branch. Looking at Fig 3, if we want to find the number of sites with derived allele count = 1, we look at the length of the branches over nodes 0,1,2,3…,9 and as mutations are applied to every unit length of branch, the sum of the length of the branches over all the single samples is the count of SNPs for derived allele count = 1. If we want to find the number of sites with derived allele count = 2 , we look at the sum of the length of the branches over pairs of the nodes: 0,1,2,3…,9 , which

would be equivalent to looking at the sum of the length of the branches over nodes: 10,11, 12, 13, 14. Qualitatively, we can see that the sum of length of the branches over nodes 10,11, 12, 13, 14 is far larger than the sum of the length of the branches over nodes 0-9, therefore leading to a higher number of mutations and a higher SNP count for derived allele count = 2.

This process is repeated till we get to the final node, which in this case is node 18, thereby creating an array of SNP counts for different allele counts. As we want the final spectrum to be in terms of allele frequencies rather than allele counts, we divide the allele counts by the sample size to get the corresponding allele frequency. As this paper deals with exhaustive sampling, the allele frequency will be obtained by dividing the allele count will be divided by the census population size. Plotting the SNP counts over their corresponding allele frequencies is what creates an allele frequency spectrum.

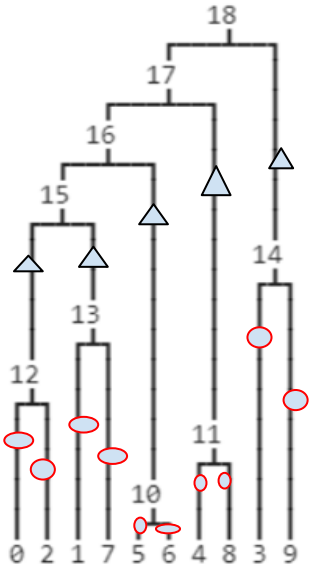

Fig 3: The sum of the lines labeled by the ovals is the count of the SNPs for derived allele count = 1. The sum of the lines labeled by the triangles is the count of the SNPs for derived allele count = 2.

At high allele frequencies, the intrinsic stochasticity in the coalescent process produces noisy allele frequency spectra. To reduce the number of simulations we needed

to run, we therefore smoothed the simulated spectra at high frequencies. Past a chosen allele frequency threshold of f = 100, the smoothing function replaces individual SNP counts with the mean count of the neighborhood. The higher the frequency, the larger the neighborhood the new count was averaged over, allowing for a drastic improvement in readability. As the smoothing function used the mean, the error was calculated to be the standard error of the mean. Please refer to [17] for the full implementation of the smoothing function

Results

According to Maruyama, there exist weakly structured and strongly structured 1-dimensional circular populations[8]. If a 1-dimensional circular population satisfies $m\rho >> L$, then structure is weak and the circular structure has little effect on the allele frequency spectrum. This weak structural effect is seen in Fig 4, where a circular population with variables $m\rho >> L$ has an expected allele frequency spectrum that exactly matches the predicted allele frequency spectrum of a well-mixed population with census population size $N=\rho L$. We therefore focus on the opposite limit, $m\rho << L$, where structure is said to be strong, and the circular population has genetic diversity equivalent to a well-mixed population of effective size

$$N_e = \frac{L^2}{2m\pi^2} \quad (2)$$

[10]

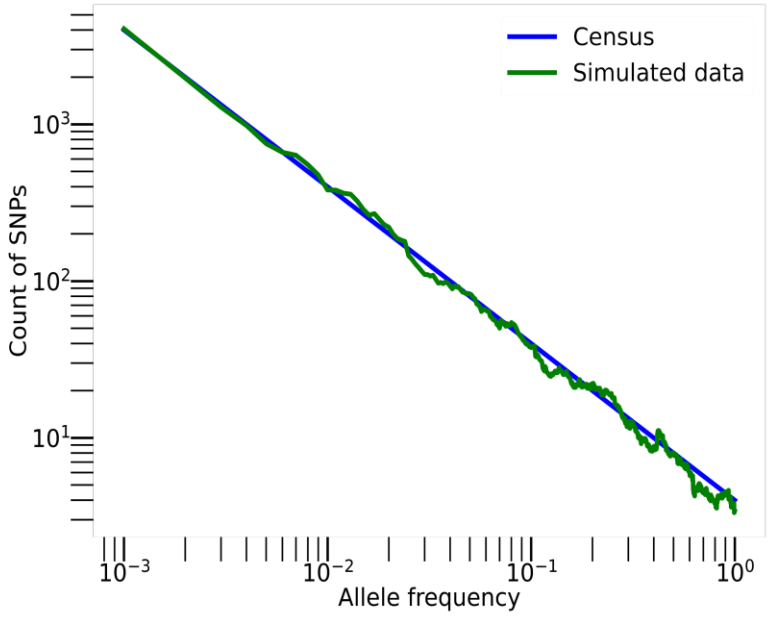

Fig 4. Weakly structured populations are approximately well-mixed**.** The expected allele frequency spectrum found in simulations of a circular population with mρ>>L (green) closely matches that predicted for a well-mixed population with the same census size N=Lρ(blue)., m=0.5, ρ=200, L=5. The simulated coalescent data was averaged over 50 trials, which will be replicated for the result of the figures.

Strong structure implies that the number of demes and migration between said demes should play a significant role in the spread of mutation. N=ρL and $N_e = \frac{L^2}{2m\pi^2}$ will be referred to as Maruyama's limiting expressions for weak and strong structure respectively. There are two variations of strong structure where mρ << L:

1) mρ>>1: Structure is discrete, but the overall structured population behaves similarly to a continuous population due to rapid migration between demes
2) mρ<<1: Weak continuity as there are too few migrants per generation for alleles to spread inter-deme

Allele frequency spectra where $1 << m\rho << L$

Making a circular population comply to the inequality $1 << m\rho << L$ makes sure that while structure is strong due to L being larger than $m\rho$ [8], the migration rate and deme density are kept high enough for inter-deme dynamics and intra-deme dynamics respectively to be apparent as well. Qualitatively, this balance of intra-dynamics and inter-deme dynamics should lead to behavior that is not entirely well-mixed throughout the frequency range.

This qualitative conclusion is proven to be true as keeping $1 << m\rho << L$ allowed the simulated circular populations to exhibit transitional behavior between the census and effective population. As seen in Fig 5 where $m = 0.05, \rho = 200$ and L=1000, the circular population initially conforms to the census population but transitions at a specific transition frequency to the effective population from the scattering phase to the collecting phase. The same behavior was confirmed with different values of L from 1000 - 2000.

In the scattering phase, coalescent events only happen intra-deme and therefore the habitat's genetic diversity is independent on the spatial structure of the population, making the expected allele frequency spectra for simulated circular population match the predicted allele frequency spectra for the census population. In the collecting phase, coalescent events occur by bringing lineages from different demes together to coalesce, making the genetic diversity of population highly dependent on spatial structure. Therefore, in the collecting phase, the expected allele frequency spectra for simulated circular population match the predicted allele frequency spectra for the effective population.

This behavior is fascinating as setting $1 << m\rho << L$ gives the population strong structure and therefore, the population should behave like a well-mixed population with

effective size according to Maruyama's limiting expressions (Maruyama 1971) (Equation 2). However, the population instead satisfies the census population at low frequencies and transitions to satisfy the effective population at higher frequencies (Fig 5).

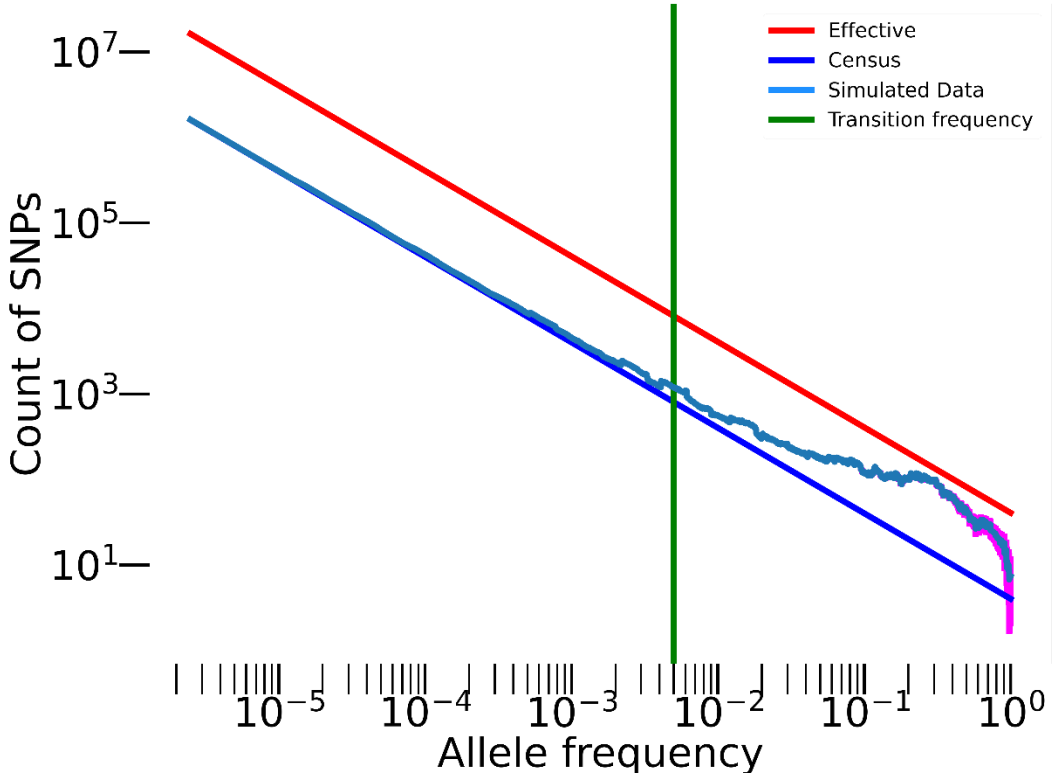

Fig 5: The expected allele frequency spectrum for a one-dimensional circular population with variables L =1000, $\rho = 200$, m=0.05 along with the predicted allele frequency spectra for well-mixed populations of census and effective size. Strongly structured allele frequency spectrums where $1 << m\rho << L$ do not solely behave like a well-mixed population with effective size N_e (Equation 2) but instead behave like the census population at lower frequencies, and transition at a specific transition frequency to behave like the effective population at higher frequencies.

The transition frequency was calculated by looking at the time scale for one-dimensional diffusion $t_{diff} = \frac{x^2}{m}$ [18]. This time scale was created by stating that during coalescence, a

lineage takes a one-dimensional random walk. On average, going backwards in time, a lineage can only take one step at a time, where $\Delta s = \pm 1$ and the sign/direction of the step is random. This implies that the average distance the lineage travels after n steps is

$$\bar{x} = \overline{\sum_{i=1}^{n} \Delta s_i}$$

but that makes $\bar{x}$=0 as the probability of a + or – step is ½ and taking the average of all these steps should make the steps cancel each other out. Therefore, we instead consider

$$\overline{x^2} = \overline{\sum_{i=1}^{n} \Delta s_i{}^2} = \overline{\sum_{i=1}^{n} \Delta s_i{}^2} + 2 \sum_{i=1}^{n} \sum_{j=1}^{n} \overline{\Delta s_i \Delta s_j}$$

But in $\Delta s_i \Delta s_j$, both Δs terms are independent therefore the average of the sum of steps that are $\pm 1 = 0$

$$\therefore \sum_{i=1}^{n} \overline{\Delta s_i \Delta s_j} = 0$$

$$\therefore \overline{x^2} = \overline{\sum_{i=1}^{n} \Delta s_i{}^2} = n$$

as $\Delta s_i{}^2 = 1$. This implies that the average distance the lineage travels is

$$\sqrt{\overline{x^2}} = \bar{x} = \sqrt{n}$$

for step size = 1.

The diffusion time scale corresponds to the entire travel time of the lineage backwards in time:

$$t_{diff} \sim \frac{\text{number of steps}}{\text{speed at which each step is taken}}$$

where on average, the lineage takes a ± 1 step every $\frac{1}{m}$ generations

$$\therefore t_{diff} \sim \frac{\text{number of steps}}{\text{average speed at which each step is taken}} = \frac{n}{\frac{\text{magnitude of each step}}{\text{number of generations per step}}} = \frac{n}{\frac{1}{\frac{1}{m}}} = \frac{n}{m} = \frac{\overline{x^2}}{m}$$

and we can simplify $\overline{x^2}$ to x^2 because we are using the expected allele frequency spectra for our simulations.

According to Weissman [19], the number of mutants directly correlates to the generational age of the mutant via $n(t) = t$; therefore, by equating n to t, we get the following math:

$$t = \frac{x^2}{m} \rightarrow x = (mt)^{\frac{1}{2}}, n = t \ \therefore \ \frac{n}{x} = (\frac{t}{m})^{\frac{1}{2}} = \rho \ \ \therefore \ t_{wm} = m\rho^2 \quad (3)$$

Where t_{wm} is the time at which the well-mixed nature of the circular population breaks. As the allele frequency is simply the fraction of the total population that are mutants, the frequency where the well-mixed time scale breaks is defined by

$$f_{transition} = \frac{t}{\rho L} = \frac{mp}{L} \quad (4)$$

The transition frequency represents the point when the purely well-mixed nature of the system breaks and the rare mutant allele transitions from the scattering phase to the collecting phase. Once this transition phase finishes, the rare allele now becomes the wild-type allele, and the simulation behaves like a well-mixed population with an effective size $N_e = \frac{L^2}{2m\pi^2}$ instead of N=ρL.

Allele frequency spectra where $m\rho << L$ **and** $m\rho < 1$

Keeping $m\rho << L$ but allowing $m\rho < 1$ keeps structure strong and allows the simulated circular populations to similarly exhibit transitional behavior; however, in a different format to circular populations that follow $1 << m\rho << L$ with notedly different transition frequencies. There are two ways to break the inequality $1 << m\rho << L$ via $m\rho << 1$. The first case is by keeping $\rho = 200$ and $L = 1000$ from Fig 5 but lowering m to 0.00025 to make$m\rho = 0.05$<<1 (Fig 6). Lowering the migration rate makes the chance

of a mutant migrating intra-deme less likely, making the circular structure less important than it is in the original inequality $1 << m\rho << L$.

While the circular population does transition in Fig 6, the equivalently calculated transition frequency from Equation 4, $f = \frac{mp}{L} = 5 * 10^{-5}$, undershoots the actual transition phase as it places itself in the scattering phase. The lesser intra-deme movement increases the initial frequency range where well-mixed dynamics dominate; invalidating the transition frequency calculated via Equation (4). This large well-mixed frequency range implies a proportionally higher scattering phase, as the end of the scattering phase is defined as when all samples have either coalesced or migrated to a non-birth deme. By keeping ρ and L constant lowering m, the rate at which all samples either coalesce or migrate to different demes is reduced.

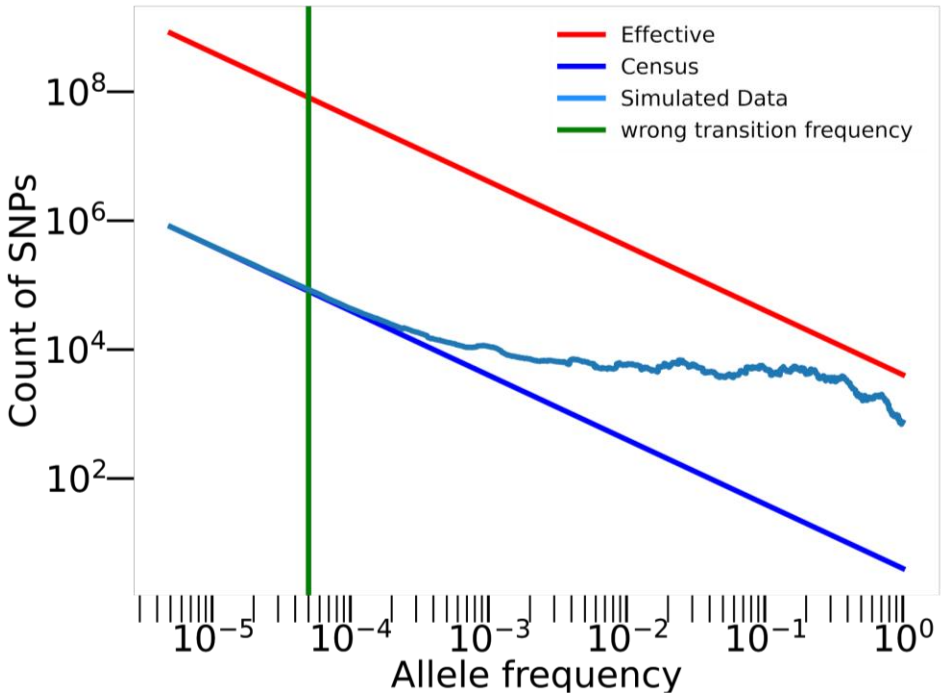

Fig 6: The expected allele frequency spectrum for a circular population with variables L=1000, $\rho = 200$ m=0.00025 along with the predicted allele frequency spectra for well-mixed populations of census and effective size. All variables from Fig

5 are kept constant but m=0.00025 to make $\frac{m\rho}{L} = 5 * 10^{-5}$<1. The calculated transition frequency undershoots the transition phase as the lower migration rate extends the amount of time that well-mixed dynamics dominate, leading to a higher scattering phase that is not described by $\frac{m\rho}{L}$. This invalidates $\frac{m\rho}{L}$ as the transition frequency for this population and highlights the need for a new transition frequency.

Invalidation occurs once again when the inequality is broken by keeping m constant and making ρ small. This was done by reducing Fig 5 's $\rho = 200$ to ρ=1 while keeping m=0.05 and L=1000. As the allele frequency spectrums work within the range of N where N=ρL, reducing ρ reduces the domain of frequencies available within the population. In the case of Fig 7, reducing $\rho = 200$ to ρ=1 reduces N =200000 to N=1000. The issue lies in that f=mρ/L does not adjust to this change in N, giving us the same transition frequency as in Fig 6; however, N being 1000 instead of 200000 limits the allele frequency domain to 10^{-3}. This leads to the transition frequency of $5 * 10^{-5}$ undershooting the entire spectrum as seen in Fig 7.

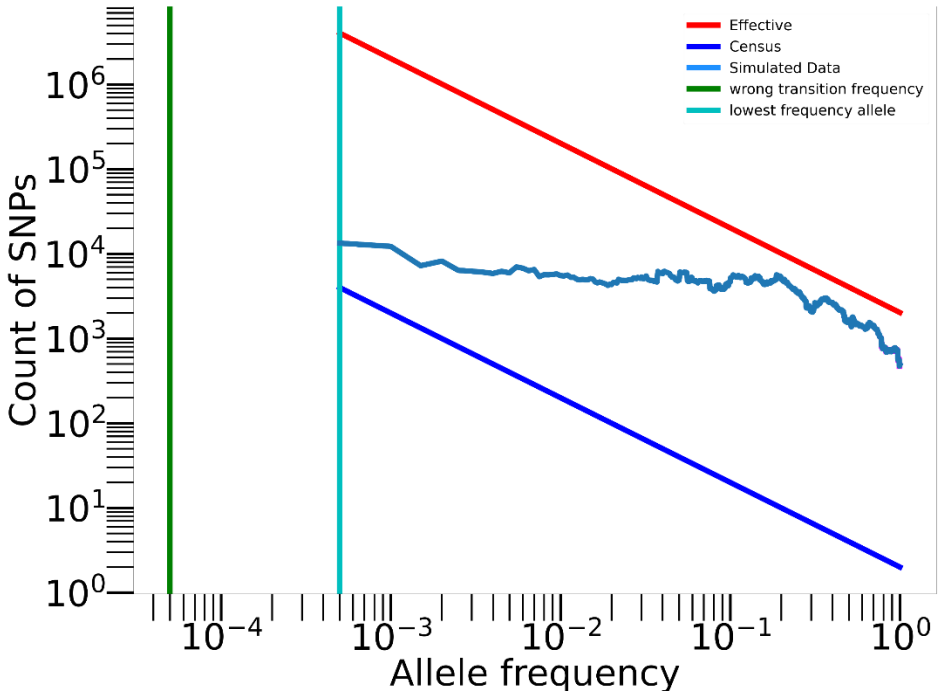

Fig 7: The expected allele frequency spectrum for a circular population with variables L=1000, $\rho = 2$, m=0.05 along with the predicted allele frequency spectra for well-mixed populations of census and effective size. All variables from Fig 5 are kept constant but ρ=2 to make mρ<1. While the transition frequency is the same as it is in Fig 6 and also undershoots the transition phase, it does so in a different manner. Reducing ρ reduces N which in turn reduces the frequency domain of the allele frequency spectrum. This makes the transition frequency of 0.0001 not only undershoots the transition phase but even the lowest frequency allele possible. This invalidates $\frac{m\rho}{L}$ as the transition frequency for this population as well and further highlights the need for a new transition frequency.

Calculating transition frequencies

It is evident through Fig 6 and Fig 7 that the transition frequency calculated via Equation (3) does not work when $m\rho < 1$. It is important to note that along with t_{wm} , the time scale at which a deme fills up with the mutation,

$$t_d \sim \frac{\text{number of individuals in a deme}}{\text{mutation rate}} = \frac{\rho}{1} = \rho \qquad (5)$$

needs to be considered as well as once a deme fills up with the mutation, inter-deme dynamics become irrelevant, and the population dynamics starts to become dependent on spatial structure. Time scales $t_{wm} = m\rho^2$ and $t_d = \rho$ represent the balance between the inter-deme and intra-deme respectively in the spread of the mutation, where once the dominant of the two time scales is broken, the genetic diversity of the 1-dimensional circular population becomes spatially dependent and the transition occurs. Therefore, to find the correct transition frequency, the following adjusted method needs to be applied:

$$x = (mt)^{\frac{1}{2}}, n = t \therefore \frac{n}{x} = (\frac{t}{m})^{\frac{1}{2}} = \rightarrow t = \max(\rho, m\rho^2) \quad (6)$$

if $\rho > m\rho^2$

$$f_d = \frac{1}{L} = \text{transition frequency} \quad (7)$$

if $m\rho^2 > \rho$

$$f_{wm} = \frac{mp}{L} = \text{transition frequency}$$

In both Fig 6 and Fig 7, $\rho > m\rho^2$; therefore, the correct time scale to associate with the population's transition to spatial dependence should be tt_d. leading to $f_d = \frac{1}{L}$ via Equation (7). Fig 8 and Fig 9 depict the exact same simulation data to Fig 6 and 11 respectively; however, with f_d instead of f_{wm}.

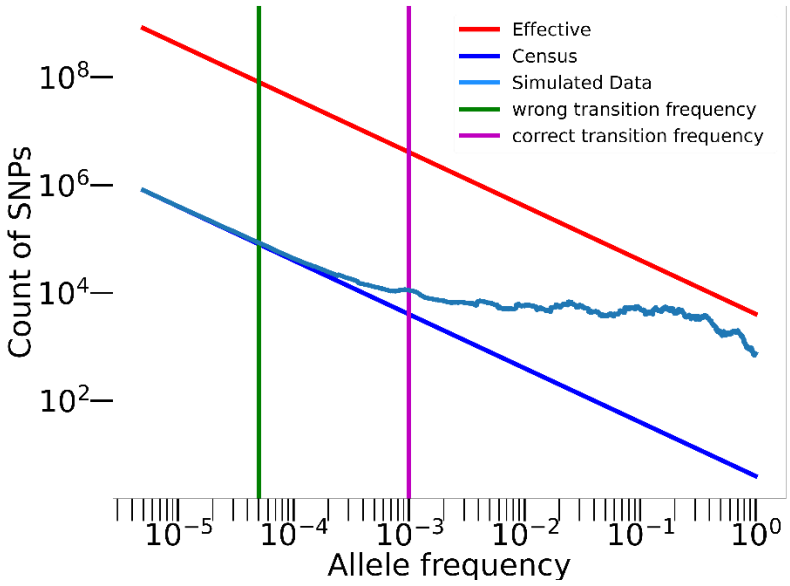

Fig 8: The expected allele frequency spectrum for a circular population with variables L=1000, $\rho = 200$ m=0.00025 along with the predicted allele frequency spectra for well-mixed populations of census and effective size. Same data as Fig 6 but with the corrected $f_d = 10^{-3}$ instead of $f_{wm} = 5 * 10^{-5}$. The transition frequency indicates the transition phase with far greater accuracy than it did in Fig 6, implying that using the deme-filling time scale is highly effective when $m\rho < 1$.

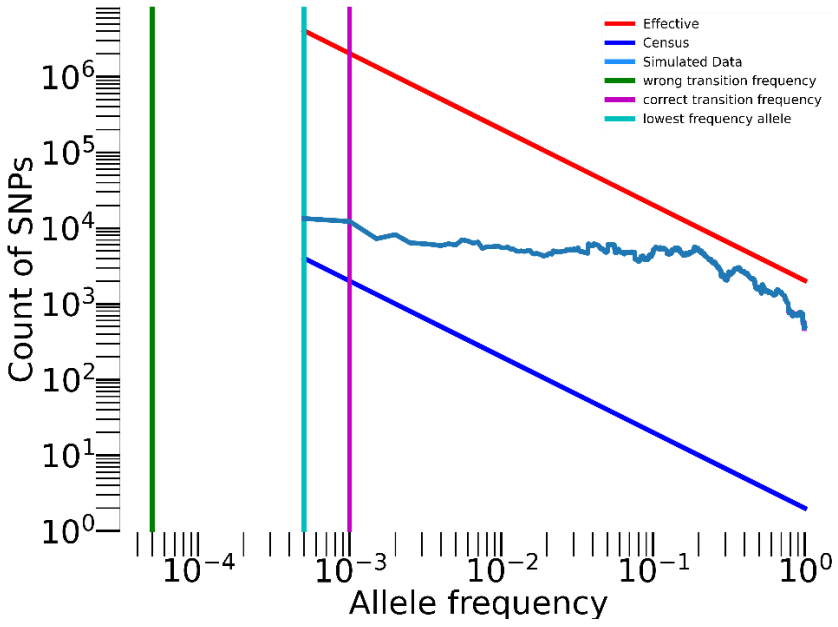

Fig 9 Same data as Fig 7 but with the corrected $f_d = 10^{-3}$ instead of $f_{wm} = 5 * 10^{-5}$. L=1000, $\rho = 1$, m=0.05. The transition frequency indicates the transition phase with far greater accuracy than it did in Fig 6. Just like with Fig 8, the accurate transition frequency implies that using the deme-filling time scale is highly effective when $m\rho < 1$

Comparatively, the threshold frequencies in Fig 8 and Fig 9 perform better than their counterparts in Fig 6 and Fig 7. This suggests that it is imperative to consider which of the well-mixed or deme-filling time scale dominates the other to define the threshold frequency at which the population breaks its spatial independence, leading to f=1/L via Equation (4).

This logic follows through to the earlier examples as well. In Fig 10, $1 << m\rho << L$ itself implies that the well-mixed time scale $m\rho^2$ dominates the deme-filling time scale as$1 << m\rho$ is equivalent to $\rho << m\rho^2$; therefore the well-mixed time scale is what defines the limit at which spatial independence breaks and f_{mp} becomes the correct transition frequency via Equation (4), which is correctly plotted in Fig 5. Erroneously applying f_d causes the threshold frequency to significantly undershoot the actual transition region as the mutant rare type has not begun transitioning from the scattering to the collecting phase yet.

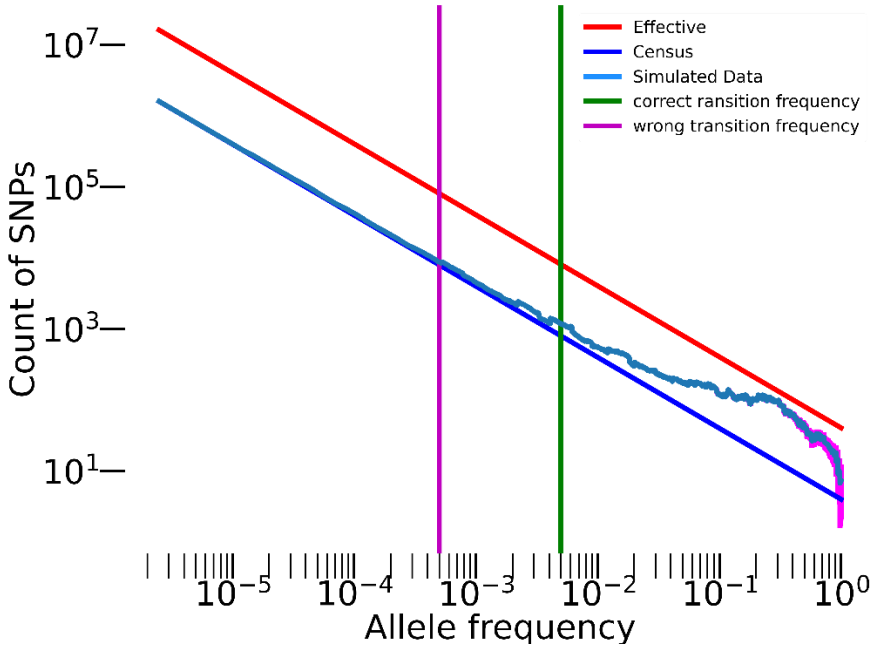

Fig 10: Erroneously applying f_d instead of f_{wm} to Fig 5's allele frequency spectrum. L =1000, $\rho = 200$, m=0.05. The scattering phase is longer than what the deme filling time scale predicts. This shows that it is important to use correct time scale when deriving the transition frequency.

Discussion

This paper showed, using allele frequency spectrums, that 1-dimensional circular populations that comply to the inequality mp<<L defy the limiting approximations given by Maruyama [8] (Equation 2) but instead behave like well-mixed populations of size $N = \rho L$ at lower frequencies and transition to behave like well-mixed populations of size $N_e = \frac{L^2}{2m\pi^2}$ at higher frequencies. These transitions occur at specific transition frequencies that indicate when the population starts to become spatially dependent and move from the space independent scattering phase to the space dependent collecting phase. In the scattering phase, coalescence events occur very frequently but only happen intra-deme [14], hence the space independence. In the collecting phase, coalescence events are much rarer and are the product of large inter-deme migration events[14], hence the space dependence.

The transition frequencies were calculated using Equations (4), (6), and (7) by using diffusion time scale t_{diff}, along with the dominant of the two time scales: the deme-filling time scale t_d and the well-mixed time scale t_{wm}. If the deme-filling time scale dominates the well-mixed time scale, the scattering phase ends when the demes of the population fill up with mutants and the transition frequency is calculated as $f_d = \frac{1}{L}$. If the well-mixed time scale dominates, almost mutant has migrated has taken place for the scattering phase end and the transition frequency is calculated as $f_{wm} = \frac{mp}{L}$.

The existence of these transition frequencies suggests that when looking at structured populations where $m\rho << L$, it is important to deeply and accurately sequence data as behavior can evidently differ at extremely small allele frequencies. If a 1-dimensional circular population had variables m=0.01,ρ=100, and $L = 10^4$, this would give a transition frequency $f_d = \frac{1}{L} = 10^{-4}$, and would require deep enough and accurate enough sequencing to correctly notice a rare allele of 0.1% frequency and would imply transitional dynamics only present at low frequencies. Rare alleles are becoming prevalent in the study of disease control due to the advent of low-error deep sequencing; with recent studies showing newly discovered low frequency alleles of 1% or lower having a significant impact on the study of common disease [20]. This makes the study of population dynamics at low frequencies more important today than ever before.

While behavior conforming to the consensus population is defined as the scattering phase and behavior conforming to the effective population size is defined as the collecting phase, the transition phase is not defined by either phase. Coalescence events are defined to only happen intra-deme in the scattering phase, while due to large inter-deme migrations in the collecting phase[14]. The description of coalescence events in the transition phase however are not clear as during the phase, the population does not conform to the census population, implying some degree of space dependence; however, it does not conform to a calculated effectively sized population as well, making its dynamics hard to comparatively analyze. More research needs to be done specifically on the features of the transition phase, as the transition from the scattering to collecting phase is clearly not instantaneous.

An interesting observation for allele frequency spectrums where L>1000 was that at very high frequencies, the simulated allele frequency spectrum seems to transition back

from Maruyama's N_e to N (Fig 14). This can be attributed to the fixation of the non-mutant allele, leading to the loss of other alleles. This should make the circular population lose genetic diversity at the same rate as the census population.; however further research is needed on much larger L and N values to draw decisive conclusions.

A continuation to this paper would be to test whether this transitional behavior occurs in continuous space. Concepts used such as $n(t) = t$ for any well-mixed population and diffusion $x \sim (mt)^{1/2}$ are true regardless of dimensionality, while concepts such as Maruyama's effective population $N_e = \frac{L^2}{2m\pi^2}$ are only true for 1-dimensional circular populations. In reality however, natural phenomena around us takes place in continuous space and while the aforementioned isolation by distance does lead to the limited geographical dispersal of genetic data, not all studies respond well to discrete generalization. Structured population are susceptible in making assumptions about the distribution of offspring and migration that are inconsistent with continuous space [21].

While the continuation of this paper to continuous space might prove challenging, a more obvious next step would be to see whether this transitionary behavior of expected allele frequency spectra for 1-dimensional circular populations carries over to 2-dimensions. Concepts such as diffusion and deme-filling should crossover while the effective population would need to be heavily adjusted to manage the second dimension. Maruyama has already done work on the heterozygosity of 2-dimensional circular populations [22], combining his results with the concepts put forth in this paper would prove interesting as the limits of strong and weak structure should change with demes being arranged in higher dimensions.

Works Cited:

1. Wright S. Isolation by Distance. Genetics. 1943;28: 114–138.

2. Relethford JH, Brennan ER. Temporal trends in isolation by distance on Sanday, Orkney Islands. Hum Biol. 1982;54: 315–327.

3. Kimura M, Weiss GH. The Stepping Stone Model of Population Structure and the Decrease of Genetic Correlation with Distance. Genetics. 1964;49: 561–576.

4. Wakeley J, Aliacar N. Gene genealogies in a metapopulation. Genetics. 2001;159: 893–905.

5. Ewens WJ. Mathematical Population Genetics 1: Theoretical Introduction. Springer Science & Business Media; 2012.

6. Weinberg, W. ber den Nachweis der Vererbung beim Menschen. Jahres Wiertt Ver Vaterl Natkd. 1908;64: 369–382.

7. Moran. The statistical processes of evolutionary theory. The statistical processes of evolutionary theory. 1962. Available: https://www.cabdirect.org/cabdirect/abstract/19631602449

8. Maruyama T. The rate of decrease of heterozygosity in a population occupying a circular or a linear habitat. National Institute of Genetics. 1971.

9. Wilkins JF, Wakeley J. The coalescent in a continuous, finite, linear population. Genetics. 2002;161: 873–888.

10. Maruyama T. On the rate of decrease of heterozygosity in circular stepping stone models of populations. Theor Popul Biol. 1970;1: 101–119.

11. Kingman JFC. The coalescent. Stochastic Process Appl. 1982;13: 235–248.

12. Fisher RA. On the dominance ratio. Bulletin of Mathematical Biology. 1990. pp. 297–318. doi:10.1007/bf02459576

13. Wakeley J. Coalescent Theory: An Introduction. Macmillan Learning; 2016.

14. Wakeley J. Nonequilibrium migration in human history. Genetics. 1999;153: 1863–1871.

15. Leviyang S. The distribution of Fst and other genetic statistics for a class of population structure models. J Math Biol. 2011;62: 203–289.

16. Kelleher J, Etheridge AM, McVean G. Efficient Coalescent Simulation and

Genealogical Analysis for Large Sample Sizes. PLoS Comput Biol. 2016;12: e1004842.

17. Arani A. Honors. GitHub repository. GitHub; 2021. Available: https://github.com/AkashArani/Honors,

18. Ryabov AB, Blasius B. Population Growth and Persistence in a Heterogeneous Environment: the Role of Diffusion and Advection. Math Model Nat Phenom. 2008;3: 42–86.

19. Weissman DB, Feldman MW, Fisher DS. The rate of fitness-valley crossing in sexual populations. Genetics. 2010;186: 1389–1410.

20. Bomba L, Walter K, Soranzo N. The impact of rare and low-frequency genetic variants in common disease. Genome Biol. 2017;18: 77.

21. Felsenstein, Joseph. A Pain in the Torus: Some Difficulties with Models of Isolation by Distance. The American Naturalist, vol. 109, no. 967, 1975, pp. 359–368., doi:10.1086/283003

22. Maruyama, Takeo. Analysis of Population Structure: II. Two-Dimensional Stepping Sone Models of Finite Length and Other Geographically Structured Populations. Annals of Human Genetics, vol. 35, no. 2, 1971, pp. 179–196., doi:10.1111/j.1469-1809.1956.tb01391.x.