Neutral Aggregation in Finite Length Genotype space

Bahram Houchmandzadeh

Neutral Aggregation in Finite Length Genotype space.

Bahram Houchmandzadeh
CNRS, LIPHY, F-38000 Grenoble, France
Univ. Grenoble Alpes, LIPHY,
F-38000 Grenoble, France

The advent of modern genome sequencing techniques allows for a more stringent test of the neutrality hypothesis of Darwinian evolution, where all individuals have the same fitness. Using the individual based model of Wright and Fisher, we compute the amplitude of *neutral* aggregation in the genome space, *i.e.,* the probability of finding two individuals at genetic (hamming) distance k as a function of genome size L, population size N and mutation probability per base ν. In well mixed populations, we show that for $N\nu < 1/L$, neutral aggregation is the dominant force and most individuals are found at short genetic distances from each other. For $N\nu > 1$ on the contrary, individuals are randomly dispersed in genome space. The results are extended to geographically dispersed population, where the controlling parameter is shown to be a combination of mutation and migration probability. The theory we develop can be used to test the neutrality hypothesis in various ecological and evolutionary systems.

I. INTRODUCTION.

Aggregation of individuals is a common observation in evolutionary and ecological systems. By aggregation, we mean the observation that some areas of *space* contain large numbers of individuals while other parts contain relatively few. To be more precise, the variance of population distribution is much larger than its mean. Consider for example ecological communities where observations are made in the *real* space. Since the seminal work of Taylor et al.,[1] who surveyed around 4000 samples from 100 species across different kingdoms, it has been established that all species tend to spatially aggregate.

There are many causes for aggregation and determining these factors is the fundamental subject of ecological and evolutionary theories and of our understanding of the natural world. One cause that is often disregarded is encoded in the very nature of life: individuals appear by birth close to their parents, but can die anywhere. Therefore, each birth event enriches the spatial pair correlation function at short distances, while a death event depletes all distances. Of course, individuals (or their seed for plants) move randomly in space and the diffusion phenomena tends to counteract and dilute the effect of correlation created at short distance by birth. It can be shown however [2–5]that diffusion is not enough to efficiently dilute the correlation creation at short distances for *spatial dimensions* $d \leq 3$ *and* large system size extension. This phenomena is called neutral clustering (see [6] for a review); it has been demonstrated experimentally[7] and the main concept has been applied to other systems such as neutrons in nuclear reactors [8, 9] and evolution of bimolecular networks[10].

The same arguments can be applied to genome space. Consider individuals characterized by the sequence of their genome, of length L. Each duplication event can give rise to a new individual due to mutations at one position in the sequence. The genetic distance k between two individuals being defined as the number of differences between their sequence (the hamming distance), we see that birth events again tend to enrich correlations at short distances, under the assumptions that all mutations are *neutral*, *i.e.,* don't affect the fitness of the individual. There is however a marked difference between diffusion in real space due to random movements and diffusion in genotype space due to mutations. The former happens in a low dimensional space ($d \leq 3$) and large extension; the latter occurs in high dimensional space ($d = L$) but small extension (the number of values such as A,T,C,G that a position along the sequence can take). Therefore, it is not clear *a priori* which effect (neutral aggregation or diffusion) is dominant. The purpose of this article is to weight these factors precisely in genotype space.

In the evolutionary field, neutral models were first proposed by Kimura [11]as the main driving force in evolution. In the ecology field, Hubbell [12] used a special version of the neutral model (UNTB), called the infinite allele model [13], to explain the pattern of biodiversity in nature in terms of neutral mutations . Both theories are passionately debated in the literature (see [14] for a review of UNTB, [15] for a critical review of both theories and their interconnections, or [16] defending the importance of neutral theories in ecology).

The advent of modern gene sequencing tools has opened the possibility for more stringent tests of the neutral hypothesis in ecological communities, combining the standard measurement of abundances of species with the histogram of genetic distances between species. Jeraldo et al.[17] for example measured the genetic distance between OTUs of six gastrointestinal microbiomes of different mammals and found that the histogram of distances is sharply peaked toward short distances. They thus concluded that neutral processes play a negligible role in these communities and selection is the dominant force. This argument of Jeraldo et al. is however problematic, as they equate *neutral* with "drawn at random", ignoring the importance of neutral forces discussed above. This point was raised by D'Andrea and Ostling[18] who developed a simple, interaction free model to demonstrate that neutral causes for this model will lead to genetic

distance histograms that are peaked at short distances.

We will show here, by exactly computing the distance pair correlation function (normalized histogram) under the neutral hypothesis, that the picture is more nuanced and depends crucially on the mutation number, *i.e.* the product of mutation probability and the population size. We show that in well mixed populations, for small mutation number, neutral aggregation is the dominant force and the distribution of distances is nearly geometric, showing a sharp peak at the origin. In this limit, most individuals are very close to each other in genotype space. For large mutation numbers however, the distribution of distances is more binomial-like and shows a peak at $\sim L/2$, as was assumed by Jeraldo et al. (figure 5).

A further complication is that real populations are geographically distributed and spatial migration also plays an important role. Histogram measurements have thus to take into account this factor in order to weight the effects of neutral factors.

The aim of this article is to compute precisely the genetic pair correlation function under the neutral hypothesis, using an individual based model. This quantity is of prime importance in population genetics[19] and is nowadays investigated mainly by coalescent theory[20]. However, application of coalescent theory to *finite* sequences and to geographically dispersed populations is rather difficult. In contrast, the method we develop below uses only straightforward mathematical tools and the results are derived by simple means. The computations presented below are confirmed by individual based numerical simulations.

This article is organized as follow: in the next section, we define the model, the pair correlation function u_k and summarize the main findings of the article. The next three sections are devoted to the mathematical derivations of these results: in section III, we use the well known "infinite site" model to illustrate how the equations governing u_k can be obtained directly from the model and be solved. Section IV generalizes this approach to finite sequences when back mutations are explicitly taken into account and the u_k are found by their probability generating function. This section constitutes the heart of this article. Section V generalizes this approach to take into account the geographic extension of populations and the influence of migrations. The final section is devoted to placing this work in perspective and to the concluding remarks.

These computations should constitute a useful tool in assessing the importance of neutral phenomena in ecological communities in general, where results analogous to that obtained by Jeraldo et al. can be evaluated.

II. MODEL DEFINITION AND MAIN RESULTS.

In this section, we define the model and summarize the main results of the manuscripts. The mathematical derivations of these results are given in the following sections.

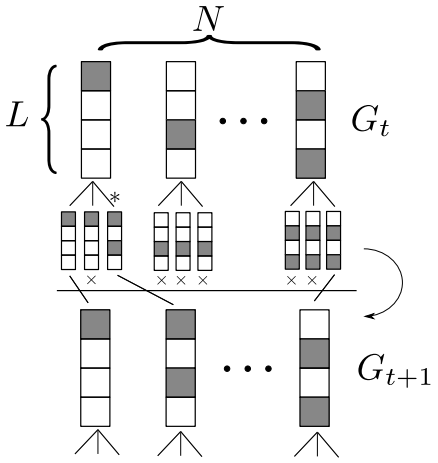

Figure 1. Neutral WF scheme for a well mixed population. The habitat contains N haploid individuals with various alleles of a gene of length L. At generation G_t, each individual produces q progeny, some of whom may contain a mutation with respect to their parent (marked by a star). N individuals are selected among the Nq progeny to constitute the G_{t+1} generation.

The two fundamental individual based models of population genetics are the non-overlapping generations model of Wright-Fisher (WF)[21] and the continuous time model of Moran[22]. These models are equivalent and obey the same diffusion equation in the large population limit[23]. Their use is dictated by the respective ease by which relevant quantities are obtained. For the present computations, we use the WF model which allows for the straightforward computation of the pair correlation function and its extension to the spatial case.

Consider a habitat containing N haploid individuals at each generation (figure 1). The individuals are characterized by the sequence of their genome of length L. Without loss of generality, we will assume that each base can take only two values, 0 and 1: When there are 2^n possible values for each base (such as A, T, C, G where $n = 2$ and $2^n = 4$), we can map the problem to a binary system with sequence length nL.

At each generation, each individual produces $q > 1$ progeny (the model can be trivially generalized to a random number of progeny). The progeny are then sampled at random to constitute the N individuals of the next generation (figure 1). We assume the system to be *neutral*: all individuals have equal fitness, and each progeny, regardless of its genome, has the same probability of getting to the next generation.

Each progeny can differ from its parent because of mutations. In the following, we will assume the mutation probability per base ν to be small and neglect the probability of having two mutations on the same progeny. This limitation can be relaxed if needed. We further assume ν to be the same for all bases of the genome sequence, regardless of their position in the sequence and the value

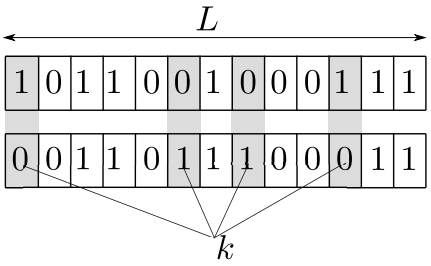

Figure 2. Two alleles of a gene of length L, differing in k of their bases (grayed area). A mutation in the grayed area will decrease the difference, while a mutation in the white area will increase it.

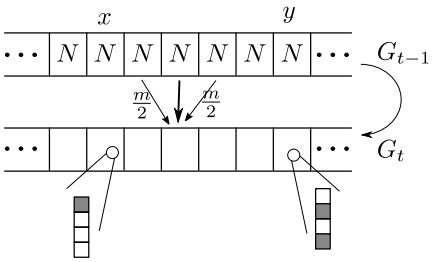

Figure 3. Scheme of the spatial WF model for finite sequence length. Space is divided into patches of N individuals. In each patch, the WF stochastic process described in fig. 1 takes place, with the additional condition that a progeny can come, with probability $m/2$ from one of the neighboring patches. $u(x, y; k; t)$ is the probability that two individuals drawn at random in patches x and y at time t are at genetic distance k.

of neighboring sites. This mutation model, known as K80, was first proposed by Kimura[24] and various refinements of this model are widely used in the field of molecular evolution, for example to deduce phylogenetic trees from sequence data[25].

The *genetic distance* between two individuals is defined as the number of bases in which they differ (hamming distance). For example, the two individuals shown in figure 2 have a hamming distance of $k = 4$.

The quantity we compute in the following sections is u_k, *i.e.,* the probability that, for large times, two individuals drawn at random are at distance k from each other. The computations are done in the limit of small mutation probability per sequence $\lambda = L\nu \ll 1$. For a gene of length $L = 1000$ base pair and mutation probability per base of $\nu = 10^{-10}$, this is a reasonable hypothesis. On the other hand, for viruses with high mutation probabilities such as $\nu = 10^{-4}$, this approximation is more problematic and the model has to be extended to take into account higher order perturbations.

If the distribution u_k is peaked at $k = 0$ and is a fast decreasing function of k, the system is clustered: there is one dominant sequence in the population, and most other sequences are at short distances from the dominant one. On the other hand, if sequences were distributed totally randomly, we would expect u_k to have a binomial distribution

$$u_k = 2^{-L} \left(\begin{array}{c} L \\ k \end{array} \right) \tag{1}$$

and display a peak at $k = L/2$.

We find that the actual distribution u_k depends only on two parameters: the sequence length L and the mutation number *per base*

$$\Omega = 2\nu N \tag{2}$$

The mutation number combines in a single number the contribution of mutations and the population size. We find that for small mutation numbers $\Omega < 1/L$, the population is clustered, u_k is peaked at $k = 0$ and decreases nearly exponentially with k. On the other hand, for $\Omega > 1$, the distribution of distances u_k tends toward a binomial one, showing peaks at $k > 1$ (Figure 5).

To measure experimentally u_k necessitates a large quantity of data. More robust quantities are the moments of the distribution such as the mean distance $\langle k \rangle = \sum k u_k$ and the variance $V = \sum k^2 u_k - \langle k \rangle^2$. The moments of the distribution contain the same information as the distribution itself. The results for the moments are particularly simple. For example, the mean is given by

$$\frac{\langle k \rangle}{L} \approx \frac{\Omega}{1 + 2\Omega} \tag{3}$$

and all the other moments are deduced from the mean by simple recurrence (equation 33). The probability distribution u_k can be determined from these moments (equation 36).

We see again from expression (3) that for $\Omega < 1/L$, $\langle k \rangle < 1$ and that most sequences are located at very short distances from the dominant one. On the other hand, for $\Omega \gg 1$, $\langle k \rangle \approx L/2$, a result which is expected from the binomial distribution of u_k (relation 1) and which denotes a random distribution of sequences.

We observe that the distribution depends crucially on the size of the population N. For large populations, the hypothesis of well mixed populations ceases to be valid. We must explicitly take into account physical space and migrations in order to determine the effective size of the population.

For this purpose, we resort to the classical stepping stone model (see [6] for a review), dividing the space into demes (or patches) of N individuals, and denote by m the migration probability, *i.e.* the probability that an individual in a deme at generation $t+1$ has its parent in a neighboring deme at generation t (figure 3). The *migration number* is defined as $M = Nm$ and combines into a single number the size of the deme and the migration probability.

Let y be the number of patches between two demes. We compute $u(k, y)$, the probability that two individuals drawn at random in two demes at physical distance y are at genetic distance k. We restrict the computation to the

mean hamming distance

$$\langle k(y)\rangle = \sum_{k=0}^{L} k u(k,y)$$

We show that the mean hamming distance in the same patch is similar to expression (3)

$$\frac{\langle k(0)\rangle}{L} = \frac{\Omega_e}{1+2\Omega_e} \tag{4}$$

Where the effective mutation number Ω_e is defined as

$$\Omega_e = \sqrt{\Omega^2 + 2\Omega M}$$

The amount of neutral clustering depends crucially on competition between migration and mutations. Moreover, $\langle k(y)\rangle$ tends exponentially in y toward $L/2$ (relation 47).

In the following sections, we detail the mathematical derivations of the above results and provide a rough estimation of the expected clustering for gut bacteria.

III. INFINITE SEQUENCE MODEL.

Consider a well mixed population of N individuals, where each individual is characterized by the sequence of its genome of length L; the population follows a Wright-Fisher competition (figure 1). At each reproduction event, a base (a position along the sequence) can be mutated with probability ν. We define the mutation probability per *sequence* as

$$\lambda = L\nu$$

and we suppose $\lambda \ll 1$ in this article (see below for the limit of validity of this hypothesis).

We wish to compute the probability $u_k(t+1)$ that two individuals drawn at random at generation G_{t+1} are at (genetic) distance k, *i.e.* they differ in exactly k bases.

Before doing the full computation, let us consider the classical "infinite site" model[19, 26], where the main mathematical concepts are straightforward to introduce. The infinite site model assumes that each mutation gives rise to a new individual that does not already exist in the habitat; this consists in assuming $L \to \infty$ while λ remains finite.

We denote

$$a = \frac{q-1}{Nq-1} \tag{5}$$

as the probability that two individuals drawn at random are from the same parent, where q is the number of progeny of each individual. Note that for $q \gg 1$, $a \approx 1/N$.

The probability that two individuals are at distance $k = 0$ from each other is :

1. The two individuals are from the same parent (probability a) AND no mutation has taken place in them (probability $(1-\lambda)^2$) OR
2. The two individuals are from different parents which themselves are genetically identical (probability $(1-a)u_0(t)$) AND no mutation has taken place in them (probability $(1-\lambda)^2$).

In other words,

$$u_0(t+1) = (1-\lambda)^2\left[a + (1-a)u_0(t)\right] \tag{6}$$

Following the same line of argument (see appendix A), we can write generally the linear system:

$$u_0(t+1) = (1-a)Bu_0(t) + b_0 \tag{7}$$
$$u_1(t+1) = (1-a)\left(Au_0(t) + Bu_1(t)\right) + b_1 \tag{8}$$
$$u_k(t+1) = (1-a)\left(Au_{k-1}(t) + Bu_k(t)\right) \quad (k>1) \tag{9}$$

where $A = 2\lambda(1-\lambda)$, $B = (1-\lambda)^2$, $b_0 = (1-\lambda)^2 a$ and $b_1 = 2\lambda(1-\lambda)a$. Note that the probability that two individuals from the same parent are at distance $k = 2$ is $a\lambda^2$ and therefore it has been neglected in relation (9). It should also be noted that because we suppose $\lambda \ll 1$, $A/B = b_1/b_0 \sim 2\lambda \ll 1$. The coefficient A captures mutation events that *increase* the genetic distances, while B relates to events that maintain the genetic distance.

We define

$$\Phi = \frac{(1-a)A}{1-(1-a)B} \approx \frac{2\lambda}{a+2\lambda} = \frac{\Theta}{1+\Theta} \quad \in [0,1)$$

where

$$\Theta = 2\lambda/a \approx 2\lambda N \tag{10}$$

is the *per sequence* mutation number. The stationary solution u_k for large times of relations (7-9), which is a one term recurrence relation, is:

$$u_k = \Phi^k(u_0 + \frac{a}{1-a}) \ (k \geq 1) \tag{11}$$
$$\approx \Phi^k(u_0 + \frac{1}{N}) \tag{12}$$

where

$$u_0 = \frac{b_0}{1-(1-a)B} \approx \frac{a}{a+2\lambda} = \frac{1}{1+\Theta} \tag{13}$$

Note that for the infinite allele model, the autocorrelation function is always peaked at $k = 0$ (or at $k = 1$ for $\Theta \gg 1$). In the low mutation regime ($\Theta \ll 1$), nearly all individuals are identical ($u_0 \approx 1-\Theta$), and u_k drops sharply as a function of k ($\Phi \ll 1$). On the other hand, in the high mutation regime $\Theta \gg 1$, the distribution is nearly flat for a wide range of k, and many different genomes coexist. The mean genetic distance for this model is :

$$\langle k\rangle = \sum_{k=1}^{\infty} k u_k \approx \Theta(1+a\Theta)$$

In the above computation, we have assumed $\lambda \ll 1$ and therefore have neglected terms in λ^2 . We can evaluate the accuracy of this approximation by estimating $S(t) = \sum_k u_k(t)$ and its deviation from unity. Summing the lines of relations (7-9), we have

$$\begin{aligned} S(t+1) &= (1-a)(A+B)S(t) + (b_0+b_1) \\ &= (1-\lambda^2)\left[(1-a)S(t)+a\right] \end{aligned}$$

and for large times, the stationary value of S is

$$S \approx \frac{a}{a+\lambda^2} \approx 1 - \frac{\lambda^2}{a} \approx 1 - N\lambda^2$$

Therefore the approximation we are making is valid when $N\lambda^2 \ll 1$. For gene sequences of length $L \sim 10^3$ and mutation probability per base $\nu \approx 10^{-9}$, the limit of validity of this computation is $N \ll 10^{12}$. The approximation is therefore valid for large populations, especially when we consider spatially distributed populations as in section V.

IV. FINITE SEQUENCE LENGTH.

We now investigate the case of finite sequence length L, which is the main point of this article. Note that if sequences were generated randomly with equal probabilities for 0 and 1 values of the bases, the probabilities u_k would follow the binomial distribution

$$u_k = 2^{-L} \left(\begin{array}{c} L \\ k \end{array} \right) \tag{14}$$

where the mean distance is $\langle k \rangle = L/2$ and the variance $V = L/4$. This is the implicit hypothesis used by Jeraldo et al. to exclude neutral processes. However sequences are not generated randomly, but are inherited from parents with a small probability for mutations. This fact radically changes the expected distribution of distances as we show below.

A. The evolution equation.

In the infinite site model considered in the preceding section, we captured the basic equations (7-9) in terms of events that increase the genetic distance (coefficient A) and events that maintain the genetic distance (coefficient B). For finite sequences, the picture is more complicated. First, mutations can also *decrease* genetic distances if they appear in bases which are already different between two sequences (back mutations); we would capture these events in a coefficient C. Second, the coefficients are not constant but depend on the number of bases in which two individuals already differ.

Consider two sequences of length L that differ in k bases (figure 2). The probability that an event increases their difference by one unit is that of one mutation (probability $2\lambda(1-\lambda)$) occurring in one of their identical bases (probability $(L-k)/L$) :

$$A_k = 2\lambda(1-\lambda)(1-\frac{k}{L}) = 2\lambda(1-\frac{k}{L}) + O(\lambda^2) \tag{15}$$

By the same token, the probability of decreasing their distance by one unit is

$$C_k = 2\lambda(1-\lambda)\frac{k}{L} = 2\lambda\frac{k}{L} + O(\lambda^2) \tag{16}$$

The probability that their distance is conserved is

$$B_k = (1-\lambda)^2 + 2\lambda^2\frac{k(L-k)}{L^2} = 1 - 2\lambda + O(\lambda^2) \tag{17}$$

and to the first order in λ, $A_k + B_k + C_k = 1$. Now, following the same line of argument as in the preceding section, the probability $u_k(t+1)$ of finding two individuals at distance k at time $t+1$ is

$$u_0(t+1) = (1-a)\left(B_0 u_0(t) + C_1 u_1(t)\right) + b_0 \tag{18}$$

$$u_1(t+1) = (1-a)\left(A_0 u_0(t) + B_1 u_1(t) + C_2 u_2(t)\right) + b_1 \tag{19}$$

$$u_k(t+1) = (1-a)\left(A_{k-1}u_{k-1} + B_k u_k + C_{k+1}u_{k+1}\right) \tag{20}$$

$$u_L(t+1) = (1-a)(A_{L-1}u_{L-1} + B_L u_L) \tag{21}$$

where b_0 and b_1 are defined as before and $1 < k < L$ in equation (20). Note that obviously, $u_k = 0$ for $k > L$. The above set of linear equations can be written in vectorial notation as

$$|u(t+1)\rangle = (1-a)Q\,|u(t)\rangle + |b\rangle \tag{22}$$

where $|u(t)\rangle = (u_0(t), u_1(t), ..., u_L(t))^T$,

$$Q = \begin{pmatrix} B_0 & C_1 & 0 & 0 & \cdots & 0 \\ A_0 & B_1 & C_2 & 0 & \cdots & 0 \\ 0 & A_1 & B_2 & C_3 & \ddots & \vdots \\ 0 & 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & C_L \\ 0 & & \cdots & 0 & A_{L-1} & B_L \end{pmatrix} \tag{23}$$

and $|b\rangle = (b_0, b_1, 0, \ldots, 0)^T$. The stationary probabilities $|u\rangle$ for large times are obtained from

$$(I - (1-a)Q)\,|u\rangle = |b\rangle \tag{24}$$

where I is the identity matrix. The above relation is a $(L+1)\times(L+1)$ linear system that can be solved numerically for practical purposes. We are of course interested in its analytical solution.

B. The probability generating function.

The non-homogeneous linear system (22) is a two term recurrence relation where the coefficients are not constant. The stationary solution u_k for large time however

is surprisingly simple if we use the probability generating function (PGF)

$$\phi(z) = \sum_{k=0}^{L} u_k z^k \tag{25}$$

The PGF contains the most complete information on the system; the probabilities and their moments are obtained from the derivatives of ϕ at either $z = 0$ or $z = 1$. Let

$$\left\langle k_{(n)} \right\rangle = \langle k(k-1)...(k-n+1) \rangle \tag{26}$$

be the factorial moment of order n. For example, the usual mean distance is $\langle k \rangle = \left\langle k_{(1)} \right\rangle$ and the variance is $V = \left\langle k_{(2)} \right\rangle + \langle k \rangle - \langle k \rangle^2$ and so on. The factorial moments are given by

$$\left\langle k_{(n)} \right\rangle = \left. \frac{d^n \phi}{dz^n} \right|_{z=1} \tag{27}$$

The moments, specially the lower ones, are the most robust quantities that we can estimate from real data analysis. Below, we shall also use the normalized factorial moments $\mu_n = \left\langle k_{(n)} \right\rangle / n!$. On the other hand, the probabilities are

$$u_k = \frac{1}{k!} \left. \frac{d^k \phi}{dz^k} \right|_{z=0} \tag{28}$$

although experimentally, their estimations necessitates much more data than what is needed for (lower) moment estimation.

It can be shown that the PGF obeys a simple first order differential equation (see appendix B):

$$(1-z^2)\phi' + \left(Lz - (1 + \frac{1}{\Theta})L \right) \phi = -aL(z-1) - \frac{L}{\Theta} \tag{29}$$

The term $aL \approx L/N$ weights the relative importance of the sequence length compared to the population size. The above differential equation can be exactly solved in terms of the hypergeometric function. However, as we are interested only in the probabilities u_k and their moments, we don't even need to solve equation (29) and we can extract all the moments from simple arguments as discussed below.

Before the full discussion, note that $N \to \infty$ implies $\Theta \to \infty$ and $a \to 0$. In this high mutation number regime, equation (29) becomes

$$(1 - z^2)\phi' + L(z-1)\phi = 0$$

with the obvious solution

$$\phi(z) = 2^{-L}(z+1)^L$$

satisfying the initial condition $\phi(1) = 1$. This is the PGF for the binomial distribution, which is expected if sequences were drawn at random. Therefore, in the very high mutation number regime ($\Theta \gg L$), the distribution of distances indeed becomes binomial.

On the other hand, the infinite sites results can be recovered from equation (29) by letting $L \to \infty$. In this limit, the ϕ' term becomes negligible in equation (29) and the PGF is simply

$$\phi(z) = \frac{1 - a\Theta + a\Theta z}{1 + \Theta - \Theta z} \approx \frac{1 + a\Theta z}{1 + \Theta - \Theta z} \tag{30}$$

where $a\Theta = 2\lambda$ has been neglected compared to one, as we suppose $\lambda \ll 1$. Relation (30) is the PGF of the probabilities u_k computed for infinite sites model in the preceding section (equation 12). This expression was first computed by Watterson[19] for the infinite allele model.

C. Finding the moments and probabilities.

Let us first consider the point $z = 1$ in equation (29). The first term vanishes at this point and we have trivially $\phi(1) = 1$ which just states that the sum of the probabilities is unity:

$$\phi(1) = \sum_{k=0}^{L} u_k = 1$$

We see here that computing $\phi(1)$ does not require a knowledge of $\phi'(1)$. This is a general feature of the PGF equation (29): $\phi^{(n)}(1)$ does not depend on the $\phi^{(n+1)}(1)$ and we can deduce all the moments from a hierarchical structure. To see this, we differentiate equation (29) in respect to z:

$$(1-z^2)\phi" + \left((L-2)z - (1 + \frac{1}{\Theta})L \right) \phi' + L\phi = -aL \tag{31}$$

As before, the higher order term vanishes at $z = 1$ and therefore the mean distance between individuals is

$$\langle k \rangle = \phi'(1) = \frac{L\Theta(a+1)}{L + 2\Theta} \approx \frac{L\Theta}{L + 2\Theta} \tag{32}$$

we see here that for mutation numbers $\Theta < 1$, the mean distance $\langle k \rangle < 1$ and most individuals are clustered very close to each other in genetic space. Only for high mutation numbers $\Theta = O(L)$, does the mean distance between individuals become appreciable. For example, $\Theta = L/2$ results in $\langle k \rangle = L/4$. For very large mutation numbers $\Theta \gg L$ we reach $\langle k \rangle = L/2$, as in the binomial distribution.

Obtaining higher moments follows the same procedure. Applying d/dz to equation (31) and computing $\phi''(1)$, we find

$$\left\langle k_{(2)} \right\rangle = \frac{2L(L-1)\Theta^2}{(L + 2\Theta)(L + 4\Theta)}$$

For the low mutation regime $\Theta \lesssim 1$ and long sequences $L \gg 1$, $\left\langle k_{(2)} \right\rangle \approx 2\Theta^2$ and the distance distribution is

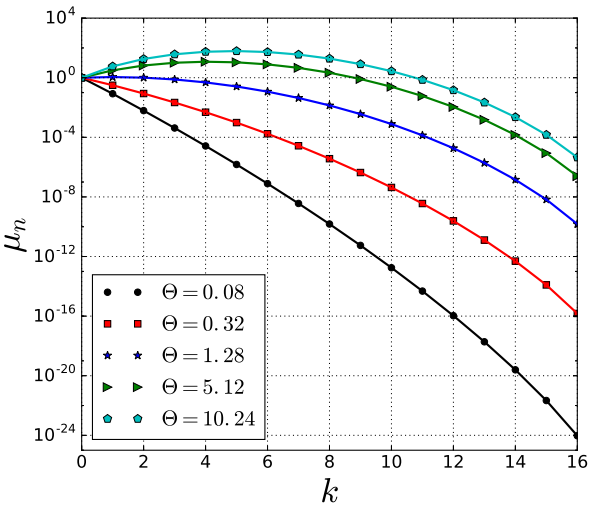

Figure 4. (Color online) The normalized factorial moments μ_n as a function of n for different values of Θ. $L = 16$, $\lambda = 2.5 \times 10^{-3}$ and $a = 1/N = 2\lambda/\Theta$. Solid lines represent theoretical values given by relation (35). Symbols are obtained by solving the linear system (24) for stationary probabilities and then computing their factorial moments from equation (26).

indeed sharply centered around the origin. For the high mutation regime $\Theta \gg L$, we find $\langle k_{(2)} \rangle = L(L-1)/4$, as expected from the binomial distribution.

Successive differentiation allows us to obtain all the moments. For $n \geq 2$

$$\langle k_{(n)} \rangle = \frac{n(L-n+1)\Theta}{L+2\Theta n} \langle k_{(n-1)} \rangle \tag{33}$$

which leads to

$$\mu_n = \frac{1}{n!} \langle k_{(n)} \rangle = 2^{-n} \frac{L(L-1)...(L-n+1)}{(\gamma+1)(\gamma+2)...(\gamma+n)} \tag{34}$$

$$= 2^{-n} \frac{(L)_{(n)}}{(\gamma+n)_{(n)}} \qquad (n \geq 1) \tag{35}$$

where $\gamma = L/(2\Theta)$ and $(r)_{(n)}$ is the descending Pochhammer symbol

$$(r)_{(n)} = r(r-1)...(r-n+1) = \frac{\Gamma(r+1)}{\Gamma(r+1-n)}$$

The expression (34,35) for $\langle k_{(n)} \rangle$ should be corrected by the factor $(1+a)$ for very small populations, as in relation (32).

Figure 4 shows the results for normalized factorial moments μ_n obtained from relation (35) and their excellent agreement with the moments obtained by numerical resolution of the linear set of equations (24).

To find the probabilities u_k, we can rearrange the PGF function

$$\phi(z) = \sum_{n=0}^{L} u_n z^n = \sum_{n=0}^{L} \mu_n (z-1)^n$$

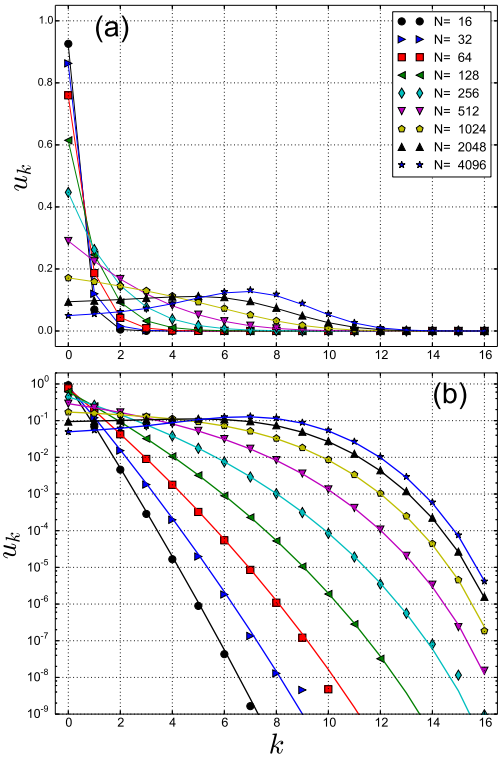

Figure 5. (Color online) Individual based stochastic simulation (symbols) of the finite sequence neutral WF model and its comparison with the theoretical predictions (solid lines). Figures show the probabilities u_k as a function of k for different values of N (and hence Θ) in (a) linear, (b) logarithmic representations. $L = 16$, $\lambda = 2.5 \times 10^{-3}$, $N \in [16, 4096]$, $\Theta \in [0.08, 20.48]$. Simulation values are obtained by generating between 10^6(for large N) to 10^8(for smaller N) paths for $10N$ generations (see appendix D).

Expanding $(z-1)^n$ and identifying the corresponding powers of z in both sums, we find

$$u_\ell = \sum_{n=\ell}^{L} (-1)^{n-\ell} \mu_n \binom{n}{\ell} \tag{36}$$

Defining the matrix C such that $(C)_\ell^n = (-1)^{n-\ell} \binom{n}{\ell}$, the solution in vectorial notation is

$$|u\rangle = C\,|\mu\rangle \tag{37}$$

Figure 5 shows the excellent agreement between the above theoretical results and individual based stochastic simulations of the neutral model (see appendix D).

It can be observed that as Θ increases , u_k transforms from a sharply decreasing distribution with its peak at zero toward a binomial distribution with the peak approaching $L/2$. The curves cease to decrease monotonically as a function of the genetic distance k for $\Theta \approx L/2$.

For small mutation numbers $\Theta \ll 1$, μ_n given by expression (35) can be approximated by

$$\mu_n = (1+a)\left(\frac{\Theta}{L}\right)^n (L)_{(n)} \quad n \geq 1 \tag{38}$$

which decreases faster than exponentially. In this case, only the first term contributes to the sum (36) and we have

$$u_0 = 1 - (1+a)\Theta$$
$$u_k = \mu_k \quad (k \geq 1)$$

It has been checked numerically that this approximation of the probabilities u_k is very good for $\Theta < 0.1$.

To summarize this section, we have obtained the exact solution for the moments μ_n (relation 35) and probabilities u_k (relation 37) of the finite sequence neutral WF model; we have obtained various limiting expressions (large and small Θ, large L) and have shown the accuracy of these expressions by comparing them with numerical simulations. The parameter Ω that controls the behavior of the system is the ratio of mutation number to sequence length

$$\Omega = \frac{\Theta}{L} = 2N\nu \tag{39}$$

i.e., the *per base* mutation number. The expressions for moments could have been derived by using this number instead. For example, the relative mean distance is

$$\frac{\langle k \rangle}{L} = \frac{\Omega}{1+2\Omega} \tag{40}$$

For bacteria such as *E.Coli* present at 10^{10}/ml in the human gut for example[27] and overall mutation probability at $\approx 10^{-10}$ per *base*[28], the mutation number is $\Theta \sim 2L$ for a population contained in 1ml. In this regime, we expect to find a large distribution of distances with a peak around $L/2$, as was indeed supposed by Jeraldo et al.[17]. On the other hand, the 1 ml volume choice is arbitrary, and if we had considered a 1μl volume instead, we would be in the low mutation regime where neutral aggregation dominates. In order to determine the effective mutation number Ω_e needed to test the neutral hypothesis, we must explicitly take into account the dispersion of individuals in the *real* space. The next section deals with this problem.

V. SPATIALLY DISTRIBUTED POPULATIONS.

Consider a population distributed into patches, each site containing N individuals. In each site, the same WF stochastic process as discussed above takes place, with the additional condition that a progeny can descend from a parent in a neighboring site with probability $m/2d$, where d is the dimension of real space. To take space into account, we need to compute $u(x, y; k; t)$, the probability of finding two individuals drawn at random in sites x and y at time t at genetic distance k.

Let us first consider a one dimensional real space $d = 1$. We assume translational invariance:

$$u(x, x'; k; t) = u(|x - x'|; k; t)$$

and will compute $u(y; k; t)$ where y is the absolute discrete distance between two patches. The evolution equation (22) for $d = 0$ of the previous section was derived by the usual combination of AND,OR, distinguishing the cases where two individuals descend from the same parent or not. For the spatial case, we must also distinguish the cases where the parent can be from the same patch or not. Again, we collect $u(y; k; t)$ $(0 \leq k \leq L)$ into a vector

$$|u(y;t)\rangle = u\left((y;0;t), u(y;1;t), ..., u(y;L;t)\right)^T$$

Following the same line of arguments as the preceding section, the evolution equation (22) is generalized for the spatial case :

$$|u(0;t)\rangle = (1-2m)\left\{(1-a)Q\,|u(0;t)\rangle + |b\rangle\right\} + 2mQ\,|u(1;t)\rangle \tag{41}$$

$$|u(1;t)\rangle = m\left\{(1-a)Q\,|u(0;t)\rangle + |b\rangle\right\} + (1-2m)Q\,|u(1;t)\rangle + mQ\,|u(2;t)\rangle \tag{42}$$

$$|u(y;t)\rangle = mQ\,|u(y-1;t)\rangle + (1-2m)Q\,|u(y;t)\rangle + mQ\,|u(y+1;t)\rangle \quad (y \geq 2) \tag{43}$$

As before, the stationary solution of these equations for large times can be obtained by using the spatial PGF function

$$\phi(y;z) = \sum_{k=0}^{L} u(y;k) z^k \tag{44}$$

By its very definition, $\phi(y;1) = 1 \ \forall y$. An evolution equation can be obtained for the PGF which has the same property as in the preceding section: $\left(\partial^n/\partial_{z^n}\right)\phi(y,z)\big|_{z=1}$ depends only on the lower derivatives, and all the factorial moments can be obtained from a hierarchical structure. For example, the mean

$$\langle k(y) \rangle = \sum_{k=0}^{L} k u(y,k) = \left.\frac{\partial \phi(y,z)}{\partial z}\right|_{z=1}$$

is shown to obey the equation

$$(-r + m\Delta)\langle k(y)\rangle = -\frac{Lr}{2} + a\left\{(1-2m)\delta_{y,0} + m\delta_{y,1}\right\}\langle k(0)\rangle \tag{45}$$

where Δ is the discrete Laplacian operator

$$\Delta f(y) = f(y-1) - 2f(y) + f(y+1)$$

and

$$r = \frac{2a\Theta}{L - 2a\Theta} \approx 4\frac{\lambda}{L} = 4\nu$$

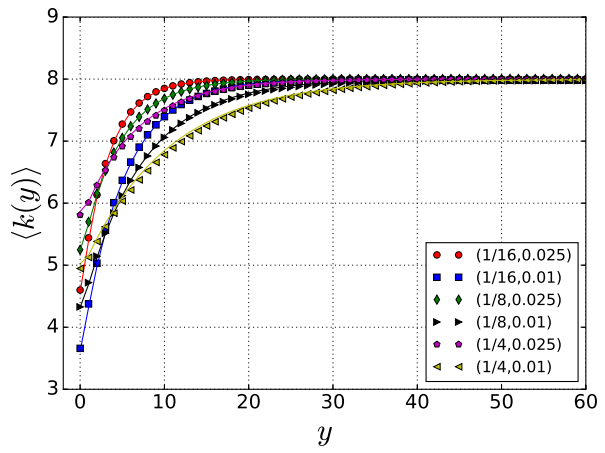

Figure 6. (Color online) Comparison between the theoretical result for $\langle k(y)\rangle$ (solid line) and stochastic simulations of the spatial, neutral, finite sequence length of WF model (symbols) for various value of (m, λ). There are $Q = 128$ one dimensional, circular sites with $N = 32$ individual per site and $L = 16$.

Equation (45) is a non-homogeneous two terms recurrence equation, which can easily be solved. The solution depends on the per sequence mutation number $\Omega = 2N\nu = \Theta/L$ and the migration number

$$M = Nm \tag{46}$$

and is written

$$\frac{\langle k(y)\rangle}{L} = \frac{1}{2} - \left[\frac{1}{2} - (1-a)\frac{\langle k(0\rangle}{L}\right] s^y \quad (y > 0) \tag{47}$$

where s is the less than unity solution of

$$s^2 - 2(1 + \frac{\Omega}{M})s + 1 = 0 \tag{48}$$

and

$$\frac{\langle k(0)\rangle}{L} = \frac{\Omega + M(1-s)}{1 + 2\Omega + 2M(1-s)} \tag{49}$$

in the regime where $a \ll 1$ and $m \ll 1$ (see appendix C for the full expression). Defining the effective mutation number as

$$\Omega_e = \Omega + M(1-s) \tag{50}$$

$$= \sqrt{\Omega^2 + 2\Omega M} \tag{51}$$

we see that the expression for $\langle k(0)/L\rangle$ is similar to the case of a well mixed population (equation 40) where Ω has been replaced by Ω_e. Figure 6 shows the excellent agreement between the above theoretical expressions and the results from individual based numerical solution of the spatial, finite sequence length WF model.

We first note that for large separations $y \gg 1$ between sites, $\langle k(y)\rangle \approx L/2$. This result is expected, as far-distant sites evolve independently of each other and we expect to recover the binomial distribution for $u(y, k)$ when $y \gg y^* = -1/\log(s)$.

On the other hand, $\langle k(0)\rangle$ is given by a balance between mutation number Ω and migration number M, and their relative amplitude with respect to unity. Even when $\Omega \ll 1$, $\langle k(0)\rangle$ can be substantially greater than unity if the migration number is not too small. In particular, note that $\Omega/M = 2\nu/m$; when $\nu \ll m$, the effective mutation number is

$$\Omega_e = \sqrt{\Omega^2 + 2\Omega M} \approx \sqrt{2\Omega M} = 2N\sqrt{\nu m}$$

and the typical distance under which the mean distance notably deviates from $L/2$ is

$$y^* = -1/\log s \approx \sqrt{m/(4\nu)}$$

Higher moments can be found by the same method if needed, although the algebra becomes increasingly cumbersome. The extension to higher dimensions is trivial and necessitates only the redefinition of the discrete Laplacian operator Δ (see appendix C).

We can now try to make a rough estimation of the order of magnitude of the neutral clustering for the bacterial communities studied by Jeraldo et al. Let us suppose that individuals (or their seed) diffuse in one generation according to a dispersal law which for simplicity we take to be a Gaussian

$$p_s(x) = (2\pi\sigma^2)^{-1/2} e^{-x^2/2\sigma^2}$$

The dispersion length σ determines both N and m. We can model the space as divided into patches of size $\ell = 3\sigma$. The exact size of the patch is not crucial, we have to choose it in a way that insures that (i) there is negligible migration to the next nearest neighbor and (ii) every progeny has a non-negligible probability of descending from every parent inside the same patch. $\ell = 3\sigma$ is a good compromise. For patches of this size, the migration probability between patches is $m \approx 0.27$ and the number of individual inside each patch is $N = \ell c_0$, where c_0 is the bacterial concentration. Note that the migration number Nm is not sensitive to the choice of ℓ.

We can, as an order of magnitude, choose 1 μm as the bacterial size; in very dense bacterial concentrations, bacterial movements are reduced and we may broadly estimate σ as some 10 times the bacterial size and therefore $N = 30$. With this choice of N and m, and $\nu = 10^{-10}$, $\Omega_e \approx 3\times10^{-4}$: this number corresponds to low mutation regime ; $\langle k(0)\rangle / L$ is very small and individuals inside the same patch are at small genetic distance of each other. The number of patches over which this clustering is observable is $y^* \approx 2.5 \times 10^4$ patches or ≈0.8m in physical dimension! If we suppose that samples were taken from few μL of bacterial intestinal residues, we see that we can expect large neutral clustering in these samples.

The above rough estimation shows that in principle, Jeraldo et al. data cannot exclude the neutral hypothesis for these communities.

VI. DISCUSSION AND CONCLUSION.

A fundamental question in population genetics is the number of segregating sites in the genome of individuals of a given population submitted to neutral mutations. We have derived in this article the probability u_k that two individuals, drawn at random in a population of size N, differ at k sites of a given gene of length L. This information is usually derived in the framework of coalescent theory, which is more general that the approach developed here. For example, using coalescent theory, one derives the probability that n individuals drawn at random have k segregating sites, while the approach here is restricted to $n = 2$, *i.e.,* the *pair* correlation function. On the other hand, application of coalescent theory to *finite* sequence lengths is rather difficult and it is even harder to take into account structured populations like geographically dispersed individuals (see [20] p45 for a detailed discussion of these difficulties or [29] for recent developments of coalescent theory). The approach we have developed here tackles these problems rather easily, making it interesting for analyzing more realistic systems.

The main assumptions of the approach we have developed is that the population is subject to *neutral* selection and mutation. In particular, we have supposed that all sites have the *same* mutation probability. This constraint can be relaxed to some extent and allow for rate heterogeneity, a subject of intense development in the field of molecular evolution[25, 30]. Consider for example a sequence of length $L = L_1 + L_2$, where sites belonging to the $i-$th subsequence have mutation number $\Theta_i = L_i\Omega_i$. Then the PGF for the whole sequence is the product of the PGF for each sub-sequence

$$\phi(z) = \phi_1(z)\phi_2(z)$$

and therefore all the moments of the whole sequence can be deduced from the moments of the subsequences. In particular, the mean is simply given by

$$\begin{aligned}\langle k\rangle &= \langle k_1\rangle + \langle k_2\rangle \\ &= L_1\frac{\Omega_1}{1+\Omega_1} + L_2\frac{\Omega_2}{1+\Omega_2}\end{aligned}$$

when $\Omega_i \ll 1$, the system behaves as having an effective mutation rate per base equal to the weighted mean of the subsequences.

Experimental measurement of u_k constitutes a stringent test of neutral theories in ecology (UNTB) introduced by Hubbell[12]. It has been shown by many authors[16] that UNTB predicts abundance distributions that are indeed observed in nature. On the other hand, other scientists have argued that abundance distribution is not a very selective criterion and the observed abundance patterns in natural communities can be predicted just as well by other competing theories [15]. The advent of modern gene sequencing tools allows one to develop more selective criteria. O'Dwyer et al [31] for example have observed that the reconstructed phylogeny of microbes from various habitats does not correspond to that predicted by UNTB. Jeraldo et al[17] have used a variant of u_k measurement to argue against UNTB for various microbial habitats. Their main argument that, as u_k is a sharply decreasing function of k and is therefore in contradiction with UNTB, is however weak. The weakness of this argument was pointed out by D'Andrea and Ostling[18] who showed, by combining theoretical modeling and numerical simulation, that a neutral theory gives rise to a u_k function that is sharply peaked at the origin. In this article, we have derived the explicit expression for u_k and we show that the shape of the u_k function depends critically on the mutation number : it can be either peaked around the origin or far from it, depending on the value of the mutation *and* the migration number.

For the microbial communities discussed by Jeraldo et al[17], our rough estimation suggests that the UNTB hypothesis cannot be ruled out. Of course, the current state of DNA sequencing does not allow for the measurement of the *spatial* correlation function $u(y;k)$ for these communities. But even if we assume that the samples used by Jeraldo et al have been mixed over few millimeters, the amplitude of neutral clustering remains important.

In this article, we have used the neutral WF model of competition where all individuals compete against all the others with equal strength. The genetic distance between individuals does not appear explicitly in this competition. Other models can be formulated where the genetic distance can modify the competition. In kin selection theory for example, competition is reduced for individuals closely related to each other[32]. On the other hand, Biancalani et al[33], have investigated the problem of competitive exclusion in ecology; in their approach, individuals are represented by a sequence which determines their use of available ecological niches. In their model, competition is enhanced for individuals close to each other in terms of their resource consumption, hence giving rise to patterns of abundance in the sequence space.

The approach we have developed here shares many limitations with other analytical approaches: (i) As we do not track the sequences but only pair differences, it is not possible to compute the abundance curves as in UNTB (which, it should be stressed, was derived, only for infinite allele models). (ii) Deriving more refined results beyond the pair correlation function also seems problematic for the same reasons. The simplicity of the dynamics of pair correlation function (equations 18,21) is lost when $n-$correlation functions are considered. (iii) We have considered only simple mutations. In the field of molecular evolution however, the substitution rate of a base in a sequence depends on the state of the base; for example, $A \to T$ and $A \to C$ have in general different probabilities. These different rates are contained in a substitution matrix for which many models coexist [25]. It could be possible in principle to generalize the approach developped here to include general substitution matrices, but the algebraic cost seems at present to be prohibitive.

Despite all the limitations enumerated above, we believe that the formalism developped in this article is a

step forward in the search for a better understanding of natural populations. In particular, we are convinced that pair correlation function measurement as proposed by Jeraldo et al will be extended to many more natural communities and the explicit expression derived here can be used to quantify the amplitude of neutral versus non-neutral mutations.

ACKNOWLEDGMENTS

We thanks M. Vallade, E. Geissler, O. Rivoire and I. Junier for the critical reading of the manuscript and fruitful discussions.

Appendix A: The infinite allele model.

The other terms of the linear system (7-9) are obtained by the same argument as the $u_0(t+1)$ terms. The probability $u_1(t+1)$ is obtained by considering that:

1. The two individuals are from the same parent (probability a) AND a mutation has taken place in only one of them (probability $2\lambda(1-\lambda)$).

2. The two individuals are from different $k=0$ parents (probability $1-a$) AND one mutation has taken place in one of them (probability $2\lambda(1-\lambda)$).

3. The two individuals are from different $k=1$ parents AND no mutation has taken place in any of them.

or, in other words

$$u_1(t+1) = 2\lambda(1-\lambda)\left[a+(1-a)u_0(t)\right]+(1-\lambda)^2(1-a)u_1(t) \tag{A1}$$

As we will neglect terms of order λ^2, two individuals at distance $k \geq 2$ will not be from the same parent. Therefore the probability $u_k(t)$ $(k \geq 2)$ is

1. The two individuals are from different $(k-1)$-distant parents AND one mutation has taken place in one of them

2. The two individuals are from different k-distant parents AND no mutation has taken place in either of them

or in other words

$$u_k(t+1) = (1-a)\left[2\lambda(1-\lambda)u_{k-1}(t)+(1-\lambda)^2u_k(t)\right] \tag{A2}$$

These relations are summarized in the linear system (7-9).

Appendix B: The PGF equation for $d=0$.

We derive here the stationary PGF equation (29) of the main text. All notations are identical to the main text (eqs. 22,23). Consider the linear form $\langle\eta| = (1, z, z^2, ..., z^L)$. By its very definition,

$$\langle\eta|u\rangle = \sum_{k=0}^{L} z^k u_k = \phi(z)$$

and

$$\langle\eta|b\rangle = b_0 + b_1 z \tag{B1}$$

Moreover, straightforward matrix multiplication shows that

$$\langle\eta|Q|u\rangle = z\sum_{k=0}^{L-1} A_k u_k z^k + \sum_{k=0}^{L} B_k u_k z^k + \frac{1}{z}\sum_{k=1}^{L} C_k u_k z^k \tag{B2}$$

Note that $A_L = 0$ and $C_0 = 0$; all the sums in relation (B2) can therefore be taken between the boundaries 0 and L. Recalling the definition of the coefficients A_k, B_k and C_k (eqs. 15-17), we see that relation (B2) contains only sums of the form $\sum u_k z^k$ and $\sum k u_k z^k$. On the other hand,

$$\sum_{k=0}^{L} k u_k z^k = z\phi'(z)$$

and therefore,

$$\langle\eta|Q|u\rangle = \frac{2\lambda}{L}(1-z^2)\phi'(z) + (1+2\lambda(z-1))\,\phi(z) \tag{B3}$$

The stationary solution for the probabilities $|u\rangle$ derived from equation (22) is given by

$$(I-(1-a)Q)\,|u\rangle = |b\rangle \tag{B4}$$

Applying the linear form $\langle\eta|$ to the above relation, we obtain

$$\frac{2\lambda}{L}(1-z^2)\phi'(z)+\left(1-\frac{1}{1-a}+2\lambda(z-1)\right)\phi(z) = -b_0-b_1 z \tag{B5}$$

The term

$$1-\frac{1}{1-a} = -\frac{a}{1-a} \approx -a$$

as we suppose that $a \ll 1$ and neglect terms of $O(a^2)$. Multiplying both side of the relation (B5) by $L/2\lambda$ leads to equation (29) of the main text.

Appendix C: The PGF equation and first moment for $d=1$.

The evolution equation for the spatial case (41-43) is obtained by generalizing the $d=0$ case. Consider for example two individuals in the same patch ($y=0$). Either

both of them descend from parents of the same patch (probability $(1-m)^2 \approx 1-2m$) or one of them descends from a parent in a neighboring patch (probability $2 \times m(1-m) \approx 2m$). Distinguishing these two cases leads to equation (41). Equations 42,43) are derived by following the same arguments. These equations can be written as

$$\begin{aligned} |u(y)\rangle &= (1+m\Delta)Q\,|u(y)\rangle \\ &+ a\left\{(1-2m)\delta_{y,0} + ma\delta_{y,1}\right\}\left(-Q\,|u(0)\rangle + |b\rangle\right) \end{aligned} \tag{C1}$$

The PGF is

$$\phi(y;z) = \sum_{k=0}^{L} u(y,k) z^k = \langle \eta | u(y) \rangle$$

where $\langle \eta |$ was defined in the previous appendix and is has been shown that

$$\begin{aligned} \langle \eta | Q | u(y) \rangle &= \frac{2\lambda}{L}(1-z^2)\partial_z \phi(y;z) + (1+2\lambda(z-1))\,\phi(y;z) \\ &= \mathcal{L}[\phi(y,z)] \end{aligned} \tag{C2}$$

where the operator $\mathcal{L}[\phi]$ captures the right hand side of equation (C2). As $\langle \eta |$ and Δ commute, applying $\langle \eta |$ to equation (C1) leads to

$$\begin{aligned} \phi(y;z) &= (1+m\Delta)\mathcal{L}[\phi(y,z)] \\ &+ a\left\{(1-2m)\delta_{y,0} + ma\delta_{y,1}\right\}\left(-\mathcal{L}[\phi(0,z)] + f(z)\right) \end{aligned} \tag{C3}$$

where

$$f(z) = \langle \eta | b \rangle = a\,(1+2\lambda(z-1))$$

Note that $f(1) = a$ and $f'(1) = 2\lambda a$. As $\mathcal{L}[\phi(y,1)] = \phi(y,1) = 1$, it can be checked that equation (C3) trivially verifies $\phi(y,1) = 1$.

In order to compute $\langle k(y) \rangle$, the mean genetic distance between two patches

$$\langle k(y) \rangle = \sum_{k=0}^{L} k u(y,k) = \left.\frac{\partial \phi(y;z)}{\partial z}\right|_{z=1}$$

we need to apply the projection operator $D = \partial_z|_{z=1}$ to equation (C3). Setting $r = 4\lambda/L = 4\nu$ and noting that

$$D\mathcal{L}[\phi(y,z)] = (1-r)\,\langle k(y) \rangle + 2\lambda$$

we find that $\langle k(y) \rangle$ must obey the relation

$$\begin{aligned} \langle k(y) \rangle &= (1+m\Delta)\left\{(1-r)\,\langle k(y)\rangle + 2\lambda\right\} \\ &+ a\left\{(1-2m)\delta_{y,0} + m\delta_{y,1}\right\}\left\{(1-r)\,\langle k(0)\rangle - 2\lambda(1-a)\right\} \end{aligned}$$

Grouping the terms in $\langle k(y) \rangle$, dividing by $(1-r)$ and approximating $r/(1-r) \approx r$, $(1+a) \approx 1$ finally leads to equation (45) of the main text.

The equation (45) can be solved by noting that the solution for the bulk equation ($y \geq 1$) is

$$\langle k(y) \rangle = \frac{L}{2} + C\ell^y \tag{C4}$$

where ℓ must obey the equation

$$\ell^2 - (2 + \frac{r}{m})\ell + 1 = 0 \tag{C5}$$

this equation has two positive solutions ℓ_1 and ℓ_2 for which $\ell_1 \ell_2 = 1$. As the mean distance is bounded by L, the solution $\ell > 1$ is unphysical and we consider only the root $\ell < 1$. Note also that $r/m = 2\Omega/M$, where Ω and M are per-base mutation and migration numbers (relations 39,46).

Equation (45) at $y = 0$ provides a linear relation between $\langle k(0) \rangle$ and $\langle k(1) \rangle$. Using solution (C4) at $y = 1$ provides a second linear relation between these two quantities. Solving this 2×2 linear system, we find

$$\frac{\langle k(0\rangle}{L} = \frac{\Omega + M(1-\ell)}{1 - 2m + 2\Omega + 2M\,(1-\ell(1-a))} \tag{C6}$$

which reduces to the expression (49) for $m \ll 1$ and $a \ll 1$. The coefficient C, found from the same linear system is

$$C = (1-a)\,\langle k(0) \rangle - \frac{L}{2} \tag{C7}$$

which completes the solution.

Equations for higher moments can be obtained by the same method and as in the case of $d = 0$, the moment of order n depends only on moments of order $n-1$. Their solution involves only straightforward, although cumbersome algebra. As an example, consider the second factorial moment

$$\left\langle k_{(2)}(y) \right\rangle = \sum_{k=0}^{L} k(k-1) u(y;k) = \frac{\partial^2 \phi(y,z)}{\partial z^2}$$

which is obtained by applying the operator $D^2 = \partial_z^2|_{z=1}$ to equation (C3):

$$\begin{aligned} (-2r + m\Delta)\,\langle k_2(y) \rangle &= -Lr(1+m\Delta)\,\langle k(y)\rangle \\ &+ a\left\{(1-2m)\delta_{y,0} + m\delta_{y,1}\right\} \times \\ &\left\{\left\langle k_{(2)}(0)\right\rangle + Lr\,\langle k(0)\rangle\right\} \end{aligned}$$

As $\langle k(y) \rangle$ is known (relation C4), this is again a solvable two term recurrence relation.

The extension to higher dimensions is easily obtained by redefining the discrete Laplace operator Δ. For example, for $d = 2$,

$$\begin{aligned} \Delta_2 f(x,y) &= -2f(x,y) \\ &\frac{1}{2}\left\{f(x+1,y) + f(x-1,y) + f(x,y-1) + f(x,y+1)\right\} \end{aligned}$$

and the moments $\langle k(x,y) \rangle$ are given by

$$\langle k(x,y) \rangle = \frac{L}{2} + C\ell^{x+y}$$

Appendix D: Numerical simulations.

All numerical simulations are written in C++, and data analysis is performed by the high level language Julia[34]. For numerical simulation of well mixed populations (0 dimension), individuals are represented by the binary sequence of their genome. Using sequence lengths of powers of 2 (such as 16 and 32) allows us to represent each sequence as an integer, and to use the tools of the C language to manipulate the bits directly. If for example the powers of 2 integers are stored in an array p2, flipping the $n-$th bit of the integer b is performed by the operation b^p2[n] where "^" represents the binary XOR.

For WF simulation of a population of size N, two one-dimensional integer arrays of size N are considered (present and future generations). Each element in the second array chooses randomly a parent in the first array and inherits its integer (genome), accompanied possibly by a mutation with probability λ. The process is iterated (by exchanging the role of the two arrays) T times (usually $T = 10N$). Hamming distances between all elements of the final array are computed and stored in a new array H. The process is then repeated M times ($M \sim 10^6 - 10^8$) to obtain a statistically significant array H, which represent the probabilities u_k.

For one dimensional spatial simulations, the same process is applied to two (present and future generation) two-dimensional integer arrays B of size $Q \times N$, where Q is the spatial extension of the system and $B[q][n]$ represent the genome of individual n at position q. This time, each progeny chooses its parent from a neighboring site with probability $m/2$, and from the same site with probability $(1 - m)$.

[1] L. R. Taylor, I. P. Woiwod, and J. N. Perry. The density dependence of spatial behaviour and the rarity of randomness. *J. Anim. Ecol.*, 47:383, 1978.

[2] W. R. Young, A. J. Roberts, and G. Stuhne. Reproductive pair correlations and the clustering of organisms. *Nature*, 412(6844):328–331, jul 2001.

[3] B Houchmandzadeh. Clustering of diffusing organisms. *Phys Rev E Stat Nonlin Soft Matter Phys*, 66(5 Pt 1):52902, 2002.

[4] B Houchmandzadeh and M Vallade. Clustering in neutral ecology. *Phys Rev E Stat Nonlin Soft Matter Phys*, 68(6 Pt 1):61912, 2003.

[5] Bahram Houchmandzadeh. Theory of neutral clustering for growing populations. *Physical Review E*, 80(5):051920, nov 2009.

[6] K. S. Korolev, Mikkel Avlund, Oskar Hallatschek, and David R. Nelson. Genetic demixing and evolution in linear stepping stone models. *Reviews of Modern Physics*, 82(2):1691–1718, may 2010.

[7] B Houchmandzadeh. Neutral clustering in a simple experimental ecological community. *Phys Rev Lett*, 101(7):78103, 2008.

[8] Eric Dumonteil, Fausto Malvagi, Andrea Zoia, Alain Mazzolo, Davide Artusio, Cyril Dieudonné, and Clélia De Mulatier. Particle clustering in Monte Carlo criticality simulations. *Annals of Nuclear Energy*, 63:612–618, jan 2014.

[9] B. Houchmandzadeh, E. Dumonteil, A. Mazzolo, and A. Zoia. Neutron fluctuations: The importance of being delayed. *Physical Review E*, 92(5):052114, nov 2015.

[10] R. R. Stein and H. Isambert. Logistic map analysis of biomolecular network evolution. *Physical Review E*, 84(5):051904, nov 2011.

[11] Motoo Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, 1985.

[12] Stephen P. Hubbell. *The unified neutral theory of Biodiversity and Biogeography*. Princeton University Press, 2001.

[13] M Kimura and J F Crow. The number of alleles that can be maintained in a finite population. *Genetics*, 49(4):725–38, apr 1964.

[14] Sandro Azaele, Samir Suweis, Jacopo Grilli, Igor Volkov, Jayanth R. Banavar, and Amos Maritan. Statistical mechanics of ecological systems: Neutral theory and beyond. *Reviews of Modern Physics*, 88(3):035003, jul 2016.

[15] Stefan Linquist, Karl Cottenie, Tyler A Elliott, Brent Saylor, Stefan C Kremer, and T Ryan Gregory. Applying ecological models to communities of genetic elements: the case of neutral theory. *Molecular ecology*, 24(13):3232–42, jul 2015.

[16] James Rosindell, Stephen P Hubbell, and Rampal S Etienne. The unified neutral theory of biodiversity and biogeography at age ten. *Trends in ecology & evolution*, 26(7):340–8, jul 2011.

[17] P. Jeraldo, M. Sipos, N. Chia, J. M. Brulc, A. S. Dhillon, M. E. Konkel, C. L. Larson, K. E. Nelson, A. Qu, L. B. Schook, F. Yang, B. A. White, and N. Goldenfeld. Quantification of the relative roles of niche and neutral processes in structuring gastrointestinal microbiomes. *Proceedings of the National Academy of Sciences*, 109(25):9692–9698, jun 2012.

[18] Rafael D'Andrea and Annette Ostling. Can clusterting in genotype space reveal niches ? *The American Naturalist*, 187(1):130–135, jan 2016.

[19] G.A. Watterson. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7(2):256–276, 1975.

[20] Jotun Hein, Mikkel H. Shierup, and Carsten Wiuf. *Gene Genealogies, Variation and Evolution: A primer in Coalescent Theory*. Oxford University Press, Oxford, UK, 2005.

[21] R A Fisher. *The genetical theory of natural selection, a complete variorum edition*. Oxford University Press, 1999.

[22] P A P Moran. *The Statistical processes of evolutionary theory*. Oxford University Press, 1962.

[23] B Houchmandzadeh and M Vallade. Alternative to the diffusion equation in population genetics. *Phys Rev E Stat Nonlin Soft Matter Phys*, 82(5 Pt 1):51913, 2010.

[24] Motoo Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular*

Evolution, 16(2):111–120, jun 1980.

[25] Ziheng Yang. Computational molecular evolution. *Oxford series in ecology and evolution*, 2006.

[26] Fumio Tajima. Infinite-allele model and infinite-site model in population genetics. *Journal of Genetics*, 75(1):27–31, apr 1996.

[27] S Baron. Medical Microbiology. chapter 95. Microb. University of Texas Medical Branch at Galveston, 4th edition, 1996.

[28] Sébastien Wielgoss, Jeffrey E Barrick, Olivier Tenaillon, Stéphane Cruveiller, Béatrice Chane-Woon-Ming, Claudine Médigue, Richard E Lenski, and Dominique Schneider. Mutation Rate Inferred From Synonymous Substitutions in a Long-Term Evolution Experiment With Escherichia coli. *G3 (Bethesda, Md.)*, 1(3):183–186, aug 2011.

[29] John Wakeley. Coalescent theory has many new branches. *Theoretical Population Biology*, 87:1–4, 2013.

[30] Simon Y W Ho and Sebastián Duchêne. Molecular-clock methods for estimating evolutionary rates and timescales. *Molecular ecology*, 23(24):5947–65, dec 2014.

[31] James P O'Dwyer, Steven W Kembel, and Thomas J Sharpton. Backbones of evolutionary history test biodiversity theory for microbes. *Proceedings of the National Academy of Sciences of the United States of America*, 112(27):8356–61, jul 2015.

[32] A Gardner, S A West, and G Wild. The genetical theory of kin selection. *J Evol Biol*, 24:1020–43, 2011.

[33] Tommaso Biancalani, Lee DeVille, and Nigel Goldenfeld. Framework for analyzing ecological trait-based models in multidimensional niche spaces. *Physical Review E*, 91(5):052107, may 2015.

[34] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B. Shah. Julia: A Fresh Approach to Numerical Computing. *arxiv*, page 1411.1607, nov 2014.