A Haploid Wright-Fisher Mutation Model with Three Alleles

This paper was downloaded from TechRxiv (<https://www.techrxiv.org>).

LICENSE

CC BY 4.0

SUBMISSION DATE / POSTED DATE

10-01-2022 / 19-01-2022

CITATION

Tyvand, Peder A.; Nøland, Jonas Kristiansen (2022): A Haploid Wright-Fisher Mutation Model with Three Alleles. TechRxiv. Preprint. <https://doi.org/10.36227/techrxiv.18100745.v1>

DOI

[10.36227/techrxiv.18100745.v1](https://doi.org/10.36227/techrxiv.18100745.v1)

A Haploid Wright-Fisher Mutation Model with Three Alleles

Peder A. Tyvand and Jonas Kristiansen Nøland, *Member, IEEE*

Abstract—Mainstream population genetics uses diffusion models with approximative solutions rather than encompassing the exact stochastic process. This practice was necessary due to practical computing limitations for large populations. The present paper develops the exact Markov chain algebra (MCA) for a discrete haploid tri-allelic Wright-Fisher model (TA-WFM) with a full mutation matrix to address this challenge. Nonzero mutations between all three alleles give a tri-allelic model irreducible to previous bi-allelic models. The gamete frequencies for asymptotic equilibrium are calculated analytically. The exact time-dependent Markov model is evaluated numerically and presented concisely in terms of diffusion variables. The convergence with increasing population size to a diffusion limit is demonstrated for the population composition distribution. There will never be exact (irreversible) extinction when there are nonzero mutation rates into each allele. Moreover, there will never be an exact (irreversible) fixation when there are nonzero mutation rates out from each allele. From our general haploid triallelic mutation model, we only present results where there is no exact extinction and no exact fixation. We present some computations for the Markov process in full detail, exposing the behavior near the boundaries for the compositional domains, which are non-singular boundaries according to diffusion theory.

Index Terms—Bottlenecks, Markov chain, Mutations, Tri-allelic Wright-Fisher model (TA-WFM).

1 INTRODUCTION

A CLASSICAL model of population genetics is called the Wright-Fisher model (WFM). This is a discrete model without memory, where each generation t is restarted with a new distribution over the available compositions of parental individuals performing random mating. This gives rise to an offspring generation $t+1$, whereupon the same process continues by turning this offspring generation into a parental population for the next offspring generation $t+2$.

The WFM in itself has no remembrance since each parental generation performs a repetitive random mating for its immediately following offspring generation. Thus, the stochastic WFM represents a Markov process. Nevertheless, there is a memorizing effect in an actual computation of the WFM. This is the step-wise broadening of the probability distribution when time increases, representing an expanding multinomial distribution. Without mutations, irreversible fixation accumulates at absorbing boundaries for the genetic probability composition distribution. Consequently, mutation with finite probability away from a state of fixation may restore some of the genetic diversity that would otherwise get lost.

The present work is devoted to studying a full matrix of mutations in a triallelic haploid population. Random mutations from a given genotype to another genotype are sharply defined mathematically, as far as they are equipped with given probability during one generation. However, biological disadvantages hide behind this apparent mathematical clarity, letting mutations appear as more well-ordered and

Fig. 1. Sketch of the present haploid model. The three different alleles are color coded, and the six mutation rates are marked by arrows, e.g., μ_{ij} denotes $\mu_{i \leftarrow j}$ from j to i . The letter-to-number equivalences for the notation of the alleles are $A \leftrightarrow 1, a \leftrightarrow 2, \alpha \leftrightarrow 3$, which means: Allele A (red color) has index 1. Allele a (blue color) has index 2. Allele α (white color) has index 3.

controllable phenomena than they really are.

Selection will be neglected in the present work for several reasons. A fully predictive description of selection is not available since one cannot distinguish between the biological causes of selection and the biological effects of selection. It is not possible to condense the different stages of selection during one generation into one single selection parameter. However, the nature of these dilemmas for the selection parameters does not change when we proceed from a biallelic to a triallelic WFM.

- J. K. Nøland are with the Norwegian University of Science and Technology (NTNU).
E-mail: jonas.k.noland@ntnu.no
- P. A. Tyvand are with the Norwegian University of Life Sciences (NMBU).

In the present paper, we will focus on mutations that are nonzero both into each allele and away from each allele. Thereby, there is no exact fixation and no exact extinction, and the tri-allelic model will stand out as principally different from any bi-allelic model. We allow a full mutation matrix where there are, in total, six independent mutation rates. However, for the sake of simplicity, ring mutations will be preferred. We define ring mutations by having only three nonzero mutation rates, with one mutation away from one allele and one mutation into each allele. As a consequence, any irreversible fixation or extinction is avoided. All three alleles contribute on the same level to the full process, as all boundaries states are non-absorbing and remain in two-way interactions with their neighboring states.

There are three established approaches to the Markov process of genetic drift by random mating in a finite population. First the diffusion model, pioneered by Wright [1] and Kimura [2]. The second approach (in spite of being more basic), is the exact algebra of the discrete Markov process, which was carried out at the haploid level by Ewens [3], who also made comparisons with the diffusion model. The third approach is Monte Carlo sampling, carried out by various authors, see Baxter *et al.* [4]. Tyvand [5] performed it on the diploid level, demonstrating good agreement with the exact algebra of a trinomial Markov chain, where bottlenecks were incorporated.

The three above-mentioned approaches, i.e., diffusion theory, Markov chain algebra (MCA), and Monte Carlo sampling, are able to offer benchmarking and mutual confirmations. Moreover, they have complementary advantages and disadvantages. The exact Markov chain is slow to compute for large populations. Interesting hybrid methods have been suggested by Franzke and Kosko [6], [7], [8] to speed up the convergence by adding noise. Consistent diffusion models are highly regarded and are useful due to their high precision for large populations and great numbers of steps in time. It should not be forgotten that a diffusion model is usually a continuous approximation to an exact discrete stochastic process. The accuracy of diffusion approximations to a discrete stochastic process is unevenly distributed over the domains of variation for the diffusion variables, which will be demonstrated in the present work.

Diffusion equations for population genetics have parallels to the Schrödinger equation in quantum mechanics, which have been discussed by Waxman [9]. For the eigenfunctions of quantum mechanics, their macroscopic energy eigenvalues are available for empirical testing. On the contrary, precise empirical testing methods are not available for the local solutions of the diffusion equations of population genetics. From these solutions, some macroscopic quantities can be extracted, like average frequencies, fixation, and extinction. A standing challenge is that verification of diffusion models from empirical data is hardly achievable. The order of validation is rather the other way round: Assuming that properly chosen diffusion models are valid, they can be used for mutation parameter estimation from real data, as exemplified by Schrepff and Hobolth [10].

The continuum diffusion models of population genetics have the advantage that they rest on an underlying lower level, which is the discrete Markov process of random mating. This is very different from quantum mechanics,

which does not allow for such lower-level testing due to the fact that the continuum level of wave functions is the lowest level for prediction. This is because any discretized lower level is blocked by Heisenberg uncertainty.

The discrete level of a Markov chain is an underlying level of verification for the continuum description in population genetics, as a contrast to quantum mechanics. However, this Markov-chain test for a diffusion model is not an empirical test. It is a merely theoretical test within the random mating paradigm, scrutinizing the consistency of diffusion theory.

The haploid WFM involves a type of neighbor interactions with similarity to random walk, and it inspired Fisher [11] to attempt a diffusion approximation. Wright [12] carried these asymptotic results further, later completing his diffusion theory [1]. The breakthrough for the diffusion approximation came with Kimura [2]. Among the papers confirming Kimura's diffusion theory, we mention Ewens [3], and McKane and Waxman [13]. The latter authors showed how to include Dirac singularities for representing fixation boundaries in diffusion variables.

Karlin and McGregor [14] showed that the Wright-Fisher model WFM and the Moran model share a common diffusion approximation. Tyvand and Thorvaldsen [15] checked these approximations against exact their respective Markov chains. The Moran model gives a more evenly distributed approximation to diffusion theory, since it represents a nearest-neighbor interaction. The WFM cannot avoid a compromised convergence to diffusion theory at its sample points closest to the fixation boundaries. However, it is important to remember that the discrete Markov chain for a WFM with a given population size is the exact stochastic process, while its diffusion model is only a smoothed approximation. The compromised convergence to diffusion theory for finite-population Wright-Fisher Markov-chains towards the fixation boundaries are therefore shortcomings for the diffusion models, not for the underlying discrete Markov models of random mating.

The first diffusion theory for a tri-allelic locus was presented by Kimura [16], but he did not include mutations. In the present work, we will confirm the convenience of diffusion variables for the tri-allelic model by computations for the exact Markov-type stochastic process. Tier [17] introduced a tri-allelic diffusion model with mutations comparable with our work; see also Zeng *et al.* [18]. The present work elaborates a tri-allelic Wright-Fisher model TA-WFM where a full representation of mutation is incorporated, defining a mutation matrix with six elements. We will not benchmark our results with diffusion theory, but we will present our results within a frame of diffusion variables and argue for its consistency.

The paper is organized as follows. First, the stochastic model is formulated in Section 2. In Section 3 the combinatorics for the discrete Markov process is developed before Section 4 transforms the variables of the Markov chain into normalized diffusion variables. In Section 5, the analytical solution for the asymptotic steady-state average frequencies is derived. In Section 6 we present a number of exact Markov chain computations for the average frequencies. Section 7 exemplifies the detailed Markov process in compositional coordinates before Section 8 concludes the

findings of this paper.

2 FORMULATION OF THE STOCHASTIC MODEL

We will, in this section, formulate the stochastic model. The present population genetics model is a WFM of discrete non-overlapping generations of a given size. The founder generation $t = 0$ is a sharply defined collection of individuals that compose the initial parental mating pool at the basic level of haploid gametes. The following discrete generations $t = 1, 2, 3, \dots$ may, in general, have different sizes, allowing bottlenecks to be studied.

This discrete stochastic model considers haploid random mating at one locus. There are three different alleles as candidates for occupying the single locus that we consider, and these are denoted by A , a , and α .

- 1) The first alternative for the allele is termed A , i.e., red color-coding.
- 2) The second alternative for the allele is termed a , i.e., blue color-coding.
- 3) The third alternative for the allele is termed α , i.e., white color-coding.

These three colors are basically tags for visualizing the present model and thereby distinguish it better from other models of population genetics. The three colors are carefully chosen to be different from one another, avoiding the fatal misunderstanding that one of the three colors may appear as a mixture of the other two. The model is depicted in Fig. 1.

The mating is assumed to be bisexual (monoecious), so we can think of the individuals as plants with different flower colors. In the stochastic process, we do not consider the genotype level at all, but only a haploid level where the mating can be considered as picking randomly one gamete at a time, independently from the others. There is a crucial restriction for the model to operate entirely and consistently within the haploid gamete pool. The gamete is picked for reproduction and is put back into the gamete pool after this reproductive mating.

Let us imagine for a moment that there exists a superior diploid level of genotypes: Six different genotypes will then exist, AA , Aa , $A\alpha$, aa , $a\alpha$ and $\alpha\alpha$. For the mating to operate exclusively on the lower gamete level, the probability of self-fertilization must be equal to cross-mating probability. This is necessary to have the full mating process occurring on the gamete level, where we consider only the three different alleles A , a , α and not the six different genotypes listed above. The three colors that we have introduced are considered the haploid coding for red, blue, and white, not the phenotype appearance of the six different genotypes. Let us imagine that the three heterozygote diploid genotypes represent mixed colors so that the flower Aa is violet, $A\alpha$ is pink, and $a\alpha$ is pale blue. These three mixed colors of violet, pink and pale blue do not mean anything since the mating in the present model takes place exclusively at the lowest haploid level. The haploid mating pool knows only the three pure colors of red, blue, and white.

From now on, we will develop the model only at the haploid level. Therefore, the size of the population will be counted in terms of haploid individuals in the gamete pool. Suppose we want to interpret the model as representing a

diploid population. In that case, the population size will have to be given as an even number, not an odd number. This is because there must always be an even number of gametes in a gamete pool for mating when diploid organisms perform the mating.

We will follow the combinatorics introduced by Tyvand [5] since the scheme for the present model runs in parallel with the previous model, also having three different mating elements. The interpretation of these overlapping mathematical concepts is nevertheless different because the present model is haploid while the model by Tyvand [5] was diploid. The distinct, non-overlapping generations are denoted by the non-negative integer t . The founder generation is $t = 0$, and the population at $t = 0$ is to be chosen as a precise population composition, not a weighted distribution over different compositions. However, the stochastic process, as time (generation number) proceeds, will already from $t = 1$ be a distribution over different population compositions. This is the essence of the MCA that we will develop below.

We denote the population size in each generation by $n(t)$, and it is given *a priori*, preferably as even numbers, to allow the interpretation in terms of diploid individuals, even though the model operates only on the haploid level. Each specific parental population of size $n(t)$ has n_1 gametes of the first type A , it has n_2 gametes of the second type a , and it has n_3 gametes of the third type α . A sharply defined founder population is numbered as generation number zero ($t = 0$), and a precise set of values for n_1 , n_2 , and n_3 must be chosen for allowing the present model to evolve with time.

2.1 Definitions for the Composition of a Population

By definition, the number $n(t)$ counts the number of gametes in the gamete pool at generation t , implying

$$n_1 + n_2 + n_3 = n(t), \quad (1)$$

summing up the numbers of haploid individuals with each of the three alleles. The present model allows arbitrary bottlenecks, where we are free to define the number of gametes $n(t)$ in each generation. We introduce p_1 , p_2 , p_3 for the respective gamete frequencies of A , a , α in a specific parental population of gametes. These frequencies are the relative numbers of individual gametes of each type

$$(p_1, p_2, p_3) = \frac{(n_1, n_2, n_3)}{n}. \quad (2)$$

The corresponding probabilities for each offspring gamete of the three types A , a , α are denoted by P_1 , P_2 and P_3 , respectively. By definition these probabilities add up to one, yielding

$$P_1 + P_2 + P_3 = 1. \quad (3)$$

A specific offspring population of size $N(t) = n(t + 1)$ has N_1 gametes of the first type A , it has N_2 gametes of the second type a , and it has N_3 gametes of the third type α .

3 COMBINATORICS FOR THE MARKOV CHAIN

In this section, the combinatorics for the Markov chain of the stochastic model will be derived. We will first determine the probabilities for the three different offspring gametes of a

given parental population with frequencies p_1, p_2, p_3 , in the absence of mutations. The formula for these probabilities is strikingly simple, where

$$P_k = p_k, \quad k = 1, 2, 3, \quad (4)$$

since the probability for each random picking remains the same as the corresponding frequency. There is a drawing with a replacement from the gamete pool. This result is valid only for random mating without mutations and without selection. We will now generalize it to account for the six possible mutation rates that may exist.

We will disregard selection in the present model.

3.1 Inclusion of Mutations in the Random Mating

We will now modify the simple formulas (4) to account for mutations, see again the definition sketch in Fig. 1. Our model takes into account six mutation rates μ_{ij} back and forth between the three haploid gametes A (allele number 1), a (allele number 2), α (allele number 3). Our chosen notation for the mutation rates can appropriately be written

$$\mu_{i \leftarrow j} \text{ to be written with shorthand as } \mu_{ij}$$

denoting mutation rate from the allele j to the allele i . This ordering of the indexes gives some advantages in our book-keeping of the MCA.

The first formula contained in eq. (4) with $k = 1$ gives the probability of an offspring allele A , rewritten as

$$P_1 = p_1 + 0 \cdot p_2 + 0 \cdot p_3, \quad (5)$$

where factors of zero highlight the vanishing contributions from parental alleles a and α . This formula changes by including four mutation rates that involve the first allele A . Two mutations go out from allele A , changing the first term in eq. (5) as follows: $p_1 \rightarrow (1 - \mu_{21} - \mu_{31})p_1$. Similarly, the second (vanishing) term transforms into $0p_2 \rightarrow \mu_{12}p_2$. The transformation for the third term in eq. (5) is $0p_3 \rightarrow \mu_{13}p_3$. These three contributions add up to the formula

$$P_1 = p_1(1 - \mu_{21} - \mu_{31}) + \mu_{12}p_2 + \mu_{13}p_3, \quad (6)$$

representing the probability for an offspring allele A (first type). The similar probabilities for the other offspring gametes are

$$P_2 = p_2(1 - \mu_{32} - \mu_{12}) + \mu_{21}p_1 + \mu_{23}p_3, \quad (7)$$

for allele a and

$$P_3 = p_3(1 - \mu_{13} - \mu_{23}) + \mu_{31}p_1 + \mu_{32}p_2 \quad (8)$$

for allele α . The relationship $P_1 + P_2 + P_3 = 1$ from eq. (4) does not change. Eqs. (6) - (8) are reformulated in matrix form, yielding

$$\begin{bmatrix} P_1 \\ P_2 \\ P_3 \end{bmatrix} = \begin{bmatrix} 1 - \mu_{21} - \mu_{31} & \mu_{12} & \mu_{13} \\ \mu_{21} & 1 - \mu_{12} - \mu_{32} & \mu_{23} \\ \mu_{31} & \mu_{32} & 1 - \mu_{13} - \mu_{23} \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix}. \quad (9)$$

3.2 The Compact Algebra of Macrostates

A population composition corresponds to a macrostate in statistical physics. The number of microstates in each macrostate is called multiplicity. It represents the number of permutations for a given population composition, given by the trinomial distribution. The transition probability matrix for random mating from the parental generation t to its offspring generation $t + 1$ is given as

$$M_{ji}^{(t,t+1)} = \frac{N!}{N_1!N_2!N_3!} P_1^{N_1} P_2^{N_2} P_3^{N_3}, \quad (10)$$

according to the trinomial distribution. Capital N refers to the offspring population, which will be represented as the output vector of the MCA.

The indices i and j represent the one-dimensional numbering schemes for the parental and the offspring populations, given by

$$i = 1 + n_1 + \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2}, \quad (11)$$

$$j = 1 + N_1 + \frac{(N_1 + N_2)(N_1 + N_2 + 1)}{2}, \quad (12)$$

respectively. This is an efficient book-keeping that maps the two-dimensional parameter space (n_1, n_2) to a one-dimensional numbering parameter i . Because n_3 is known as the difference $n_3 = n(t) - n_1 - n_2$, the parameter space for the three gamete numbers is 2D and not 3D. The 1D index parameter i for the input (parental) population runs from 1 to $d(n)$. The analogous 1D index parameter j for the output (offspring) population runs from 1 to $d(N)$, where $d(N)$ is the number of different population compositions for a given N , which will be further defined in the next section [see eq. (24)]. The important achievement of these 1D numbering parameters for input (i) and output (j) population compositions is that the probability distribution vector $\vec{V}$ for the population composition becomes a 1D vector. Thereby the transition probability matrix $M_{ji}^{(t,t+1)}$ is confirmed to be a 2D matrix.

The input (parental) population composition vector for generation t is

$$V_i^{(t)}, \quad (i = 1, 2, \dots, d(n)), \quad (13)$$

while the output (offspring) version of this vector is

$$V_j^{(t+1)}, \quad (j = 1, 2, \dots, d(N)), \quad (14)$$

determined by the Markov chain relationship

$$V_j^{(t+1)} = M_{ji}^{(t,t+1)} V_i^{(t)}. \quad (15)$$

Note that the generation number for the offspring vector is taken to be $t + 1$ to let it serve as the input (parental) vector of the following generation $t + 1$.

4 INTRODUCING DIFFUSION VARIABLES

This section provides the diffusion variables necessary to study the stochastic process. We will not go into detail with a tri-allelic diffusion theory here, noting the developments by Tier [17], Zeng *et al.* [18] and Baxter *et al.* [4]. For our purposes, we adapt the known diffusion variables from bi-allelic theory for establishing diffusion variables for our tri-allelic model.

4.1 Diffusion Variables for a Bi-Allelic Model with Mutations

McKane and Waxman [13] provided the relevant diffusion variables for the bi-allelic haploid model. Tyvand and Thorvaldsen [15] validated their results thoroughly, both for the present Wright-Fisher model WFM and for the Moran model.

We will first recall the diffusion variables for the bi-allelic haploid model with the two alleles A and a . With n individuals, the index i counts the number of alleles A , ranging from $i = 1$ to $i = n + 1$, so that the index i is given by $i = n_1 + 1$ where n_1 is the number of individuals of type A in the gamete pool. The compositional coordinate x in the biallelic model is thus

$$x = \frac{n_1}{n} = \frac{i - 1}{n}. \quad (16)$$

The compositional coordinate x is thus established in two steps: First the different populations is counted successively, with their numbering subtracted by one. Next these numbers are scaled so that x covers the domain from 0 to 1. Note that the other allele a in the biallelic model could be described by a compositional coordinate y defined by $y = n_2/n$. Since $n_1 + n_2 = n$, we have $y = 1 - x$, which means that y is the same coordinate as x but with an opposite direction over the unit interval. The same definition for y is a useful diffusion variable for the tri-allelic model.

The time unit τ for the diffusion variables is

$$\tau = \frac{t}{n}, \quad (17)$$

where n is the number of haploid individuals in the gamete pool.

The vector element V_i of the population composition vector in the bi-allelic Markov chain is conventionally transformed as follows to diffusion variables

$$v_n(x, \tau) = nV_i, \quad (18)$$

see Tyvand and Thorvaldsen [15], recalling our present convention that n denotes the number of haploid gametes. This gives the correct asymptotic representation of diffusion for $n \gg 1$. However, for small n there are in fact distinct alternatives for definitions of $v_n(x, \tau)$: Without mutations one might alternatively use the definition

$$v_n^*(x, \tau) = (n + 1)V_i, \quad (19)$$

thereby including all population composition numbers between $i = 1$ and $i = n + 1$ into the continuous diffusion description. The number of different bi-allelic population compositions is one higher than the population size, since $n_1 = 0$ counts as one population composition side by side with those where n_1 takes all values from 1 to n .

With the entire stochastic process continuously distributed over the unit domain $0 \leq x \leq 1$, we have to count $n + 1$ different populations into the scaling. In the asymptotic limit $n \gg 1$, the leading-order contribution in orders of n will, of course, reproduce the standard diffusion variable (18).

A mutation rate μ has the corresponding diffusion variable

$$\beta = 1 - 2n\mu, \quad (20)$$

see Hössjer *et al.* [19].

4.2 Diffusion Variables for the Present Tri-Allelic Model with Mutations

We will now introduce a set of diffusion coordinates x, y, τ for the exact stochastic process. The compositional coordinates x, y serve as spatial coordinates. We introduce three compositional coordinates defined as

$$x = \frac{n_1}{n}, \quad y = \frac{n_2}{n}, \quad z = \frac{n_3}{n}, \quad (21)$$

as an extension of the definition (16) to three variables. These three coordinates represent the frequency (compositional fraction) for each of the three alleles A, a and α in any population composition. Their sum is identically one

$$x + y + z = 1, \quad (22)$$

which means that two of these coordinates are sufficient for describing the evolution. We start with all the three compositional coordinates, but will later concentrate on x and y as our primary compositional coordinates. The time variable in the diffusion representation for our tri-allelic model is $\tau = t/n$, coinciding with eq. (17) for the bi-allelic model. Here t is the discrete generation starting with a given founder population at $t = 0$. In the present paper, diffusion variables will be used only when the given population size $n(t)$ is constant.

The diffusion variable for the distribution over populations composition in our tri-allelic model is denoted by $v_n(x, y, \tau)$. A standard definition for the discrete distribution in terms of diffusion variable is

$$v_n(x, y, \tau) = \frac{n^2}{2} V_i^{(t)}, \quad (23)$$

valid to the leading order of n when $n \gg 1$. However, this simple formula is not obvious. It can be established through a more general formula similar to eq. (19) for the bi-allelic model. We need the number of different population compositions $d(n)$ with n haploid individuals in the gamete pool, which is given by

$$d(n) = \frac{(n + 2)(n + 1)}{2}. \quad (24)$$

We established a more precise version of the diffusion variable for small bi-allelic populations (19) by arguing that all different population compositions should be included in the scaled diffusion variable. The same argument repeated for the tri-allelic model leads to the following diffusion-scaled definition of the discrete variable for the population composition

$$v_n^*(x, y, \tau) = \frac{(n + 1)(n + 2)}{2} V_i^{(t)}, \quad (25)$$

from which eq. (23) emerges as the leading order asymptotic term for $n \gg 1$. In the following we will use the notation $v_n(x, y, \tau)$ for the discrete variable approaching the continuous diffusion variable $v(x, y, \tau) = v_\infty(x, y, \tau)$, by the definition (23) as the standard diffusion variable. For comparison, we will also show calculations based on the alternative diffusion variable $v_n^*(x, y, \tau)$ defined by eq. (25). The alternatives $v_n(x, y, \tau)$ and $v_n^*(x, y, \tau)$ are two ways of presenting the same exact Markov chain computations to see which of these are best adapted to the continuous diffusion approximation.

If the diffusion description is valid, the entire stochastic process will be represented by a diffusion equation for the function $v(x, y, \tau) = v_\infty(x, y, \tau)$. The diffusion description needs an initial condition $v_\infty(x, y, 0)$ to be given, and a Dirac delta function will represent a sharply defined initial composition.

Each of the six mutation rates μ_{ij} is transformed to a matrix β_{ij} in the diffusion model, by the relationship

$$\mu_{ij} = \frac{1 - \beta_{ij}}{2n}, \quad (26)$$

or equivalently

$$\beta_{ij} = 1 - 2n\mu_{ij}. \quad (27)$$

Time dependence of the mutation rates is allowed in this algebra. There is no selection in the present model.

We will now develop the MCA for the exact stochastic process of random mating. The key element in the MCA is the distribution vector $\vec{V}$ covering all possible population compositions. The founder population at $t = 0$ is sharply defined, which means that the population distribution vector has only one nonzero element at $t = 0$.

4.3 Calculation of the Average Frequencies

The mean frequencies ($\bar{p}_1$, $\bar{p}_2$ and $\bar{p}_3$) can be found by summing up the contributions of N_1 , N_2 and N_3 in each vector element, weighted by being multiplied with each corresponding probability, yielding the formulas (compacted into one) for the average frequency $\bar{p}_k$ of allele $k = 1, 2, 3$:

$$n(t)\bar{p}_k^{(t)} = [n_{k,1} \quad n_{k,2} \quad n_{k,3} \quad \dots \quad n_{k,d(n)}]^{(t)} \cdot \begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ \dots \\ V_{d(n)} \end{bmatrix}^{(t)}. \quad (28)$$

Here we have introduced the notation $n_{k,i}$ where $k = 1, 2, 3$ denotes the allele number, while the index i runs over all different population compositions, from $i = 1$ to $i = d(n)$. The one-to-one correspondence between the parameter i and the two allele numbers n_1 (representing A) and n_2 (representing a) is given by eq. (11). Note that the third allele α is implicitly represented in this book-keeping, by the identity $n_3 = n(t) - n_1 - n_2$.

4.4 Calculation Process

This MCA is to be repeated for each generation with the previous output vector serving as the new input vector. However, this is not an inductive iteration by the same formula since the population size is allowed to vary with t , and the transition probability matrix is not necessarily a square matrix. We start the Markov chain with a specified founder population (not a distribution), which means that the first input vector $V_i^{(0)}$ to the Markov chain has only one non-zero component, with the value of 1. Excluding some trivial cases, all the following generations $t \geq 1$ will have a probability distribution for the population composition, where more than one element in the population composition vector $V_i^{(t)}$ is non-zero. The sum of all the elements in the population composition vector must be one because there is no extinction in the present model.

5 ASYMPTOTIC STEADY-STATE SOLUTION

This section determines the stationary solution to the stochastic process, but only as far as the average frequencies are concerned. This analytical averaging approach is useful for giving an overview of the effects of different mutation rates as well as different population sizes for the exact Markov process of genetic drift.

A steady-state solution with constant average frequencies will emerge asymptotically as $t \gg 1$. It is given with all mutations in mutual equilibrium is given by the following relations

$$P_1 = p_1, \quad P_2 = p_2, \quad P_3 = p_3, \quad (29)$$

implying that the probability of each offspring gamete is identical to its respective parental frequency. Utilizing the three formulas (6) - (8), we get the equations for the three steady gamete frequencies, determined uniquely by the six mutation rates.

$$p_1 = \frac{\mu_{12}}{\mu_{21} + \mu_{31}}p_2 + \frac{\mu_{13}}{\mu_{21} + \mu_{31}}p_3, \quad (30)$$

$$p_2 = \frac{\mu_{23}}{\mu_{32} + \mu_{12}}p_3 + \frac{\mu_{21}}{\mu_{32} + \mu_{12}}p_1, \quad (31)$$

$$p_3 = \frac{\mu_{31}}{\mu_{13} + \mu_{23}}p_1 + \frac{\mu_{32}}{\mu_{13} + \mu_{23}}p_2. \quad (32)$$

Eqs. (30), (31) and (32) are further reformulated in matrix form

$$\begin{bmatrix} 0 & a_{12} & a_{13} \\ a_{21} & 0 & a_{23} \\ a_{31} & a_{32} & 0 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \quad (33)$$

The coefficients in the matrix are defined below, and we also insert the diffusion variables according to eq. (26).

$$a_{12} = \frac{\mu_{12}}{\mu_{21} + \mu_{31}} + 1 = \frac{2n - \beta_{12}}{4n - \beta_{21} - \beta_{31}} + 1 \quad (34)$$

$$a_{13} = \frac{\mu_{13}}{\mu_{21} + \mu_{31}} + 1 = \frac{2n - \beta_{13}}{4n - \beta_{21} - \beta_{31}} + 1 \quad (35)$$

$$a_{21} = \frac{\mu_{21}}{\mu_{12} + \mu_{32}} + 1 = \frac{2n - \beta_{21}}{4n - \beta_{12} - \beta_{32}} + 1 \quad (36)$$

$$a_{23} = \frac{\mu_{23}}{\mu_{12} + \mu_{32}} + 1 = \frac{2n - \beta_{23}}{4n - \beta_{12} - \beta_{32}} + 1 \quad (37)$$

$$a_{31} = \frac{\mu_{31}}{\mu_{13} + \mu_{23}} + 1 = \frac{2n - \beta_{31}}{4n - \beta_{13} - \beta_{23}} + 1 \quad (38)$$

$$a_{32} = \frac{\mu_{32}}{\mu_{13} + \mu_{23}} + 1 = \frac{2n - \beta_{32}}{4n - \beta_{13} - \beta_{23}} + 1 \quad (39)$$

The solution of eq. (33) is as follows.

$$p_1 = \frac{a_{12}a_{23} - a_{23}a_{32} + a_{13}a_{32}}{a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32}} \quad (40)$$

$$p_2 = \frac{a_{13}a_{21} - a_{13}a_{31} + a_{23}a_{31}}{a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32}} \quad (41)$$

$$p_3 = \frac{-a_{12}a_{21} + a_{32}a_{21} + a_{12}a_{31}}{a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32}} \quad (42)$$

The coefficients in eqs. (35)-(40) can then be inserted into eqs. (41)-(43), arriving at the final solutions presented in Table 1. These are exact analytical formulas for asymptotic

TABLE 1
Analytical Solutions to the Asymptotic Steady-State.

$$\begin{aligned}
 p_1 &= \frac{\left(\frac{\mu_{12}}{\mu_{21}+\mu_{31}}+1\right)\left(\frac{\mu_{23}}{\mu_{12}+\mu_{32}}+1\right)-\left(\frac{\mu_{23}}{\mu_{12}+\mu_{32}}+1\right)\left(\frac{\mu_{32}}{\mu_{13}+\mu_{23}}+1\right)+\left(\frac{\mu_{13}}{\mu_{21}+\mu_{31}}+1\right)\left(\frac{\mu_{32}}{\mu_{13}+\mu_{23}}+1\right)}{\left(\frac{\mu_{12}}{\mu_{21}+\mu_{31}}+1\right)\left(\frac{\mu_{23}}{\mu_{12}+\mu_{32}}+1\right)\left(\frac{\mu_{31}}{\mu_{13}+\mu_{23}}+1\right)+\left(\frac{\mu_{13}}{\mu_{21}+\mu_{31}}+1\right)\left(\frac{\mu_{21}}{\mu_{12}+\mu_{32}}+1\right)\left(\frac{\mu_{32}}{\mu_{13}+\mu_{23}}+1\right)} \\
 p_2 &= \frac{\left(\frac{\mu_{13}}{\mu_{21}+\mu_{31}}+1\right)\left(\frac{\mu_{21}}{\mu_{12}+\mu_{32}}+1\right)-\left(\frac{\mu_{13}}{\mu_{21}+\mu_{31}}+1\right)\left(\frac{\mu_{31}}{\mu_{13}+\mu_{23}}+1\right)+\left(\frac{\mu_{23}}{\mu_{12}+\mu_{32}}+1\right)\left(\frac{\mu_{31}}{\mu_{13}+\mu_{23}}+1\right)}{\left(\frac{\mu_{12}}{\mu_{21}+\mu_{31}}+1\right)\left(\frac{\mu_{23}}{\mu_{12}+\mu_{32}}+1\right)\left(\frac{\mu_{31}}{\mu_{13}+\mu_{23}}+1\right)+\left(\frac{\mu_{13}}{\mu_{21}+\mu_{31}}+1\right)\left(\frac{\mu_{21}}{\mu_{12}+\mu_{32}}+1\right)\left(\frac{\mu_{32}}{\mu_{13}+\mu_{23}}+1\right)} \\
 p_3 &= \frac{-\left(\frac{\mu_{12}}{\mu_{21}+\mu_{31}}+1\right)\left(\frac{\mu_{21}}{\mu_{12}+\mu_{32}}+1\right)+\left(\frac{\mu_{32}}{\mu_{13}+\mu_{23}}+1\right)\left(\frac{\mu_{21}}{\mu_{12}+\mu_{32}}+1\right)+\left(\frac{\mu_{12}}{\mu_{21}+\mu_{31}}+1\right)\left(\frac{\mu_{31}}{\mu_{13}+\mu_{23}}+1\right)}{\left(\frac{\mu_{12}}{\mu_{21}+\mu_{31}}+1\right)\left(\frac{\mu_{23}}{\mu_{12}+\mu_{32}}+1\right)\left(\frac{\mu_{31}}{\mu_{13}+\mu_{23}}+1\right)+\left(\frac{\mu_{13}}{\mu_{21}+\mu_{31}}+1\right)\left(\frac{\mu_{21}}{\mu_{12}+\mu_{32}}+1\right)\left(\frac{\mu_{32}}{\mu_{13}+\mu_{23}}+1\right)}
 \end{aligned}$$

equilibrium after many generations with a full mutation matrix for triallelic populations.

A simple test for these formulas is that all six mutation rates are equal: $\mu = \mu_{12} = \mu_{21} = \mu_{13} = \mu_{31} = \mu_{23} = \mu_{32}$). Then the analytical solutions in Table 1 predicts that $p_1 = p_2 = p_3 = 1/3$, which is obvious because of the full symmetry between the three alleles.

In Fig. 2 we display the variation of the steady average frequencies with the mutation rates. We consider the case where three mutation rates vanish ($\mu_{12} = \mu_{23} = \mu_{31} = 0$), while the other three mutation rates are arbitrary. We choose to plot this ring of one-way mutations because it gives the strongest effects, while any two-ways mutations gives a tendency of leveling out the differences between average frequencies.

The presentation of Fig. 2 is made using μ_{21} as scaling unit for the other two mutation rates μ_{13} and μ_{32} . The diagrams are shown with logarithmic units for μ_{32}/μ_{21} and μ_{13}/μ_{21} . Cyclic permutations between indexes 2 and 3 can be performed to show why subfigures (b) and (c) are mirror images of one another. Fig. 2 shows that the asymptotic limits for p_k , ($k = 1, 2, 3$) as μ_{32}/μ_{21} and μ_{13}/μ_{21} tend to zero or infinity have three possible values: zero, $1/2$ or 1 .

Fig. 2. Surface plots of stationary mean allele frequencies as a function of μ_{13}/μ_{21} and μ_{32}/μ_{21} , with $\mu_{12} = \mu_{31} = \mu_{23} = 0$. **a)** Frequency p_1 for allele A. **b)** Frequency p_2 for allele a. **c)** Frequency p_3 for allele α .

6 MARKOV CHAIN COMPUTATIONS FOR THE AVERAGE FREQUENCIES

In this section, we will classify and exemplify basic cases of different mutation rates, by computing the MCA of the exact stochastic process. We will not extract the distribution of population compositions, but only the mean frequencies of each haploid gamete. The evolution of these mean frequencies will be studied as they approach the asymptotic steady-state values that we calculated in the previous section.

In these computations, all mutation rates will be kept constant, and we will consider only constant population sizes. The results will be presented in diffusion variables, and thereby we provide an overall verification of the asymptotic convergence of the exact Markov process towards a diffusion limit with increasing population size. We will study how the initial state will gradually be wiped out as the Markov process converges to a final steady state, which is determined by the mutation rates alone.

The unit diffusion time ($\tau = 1$) represents a time scale for the convergence process towards the final steady state. We expect the standard value ($\beta = 0$) for the mutation parameter in diffusion variables to be the case where $\tau = 1$ is the characteristic time for the initial state to decay, after which the mutation rates will set the mean values for the alleles.

6.1 All Mutation Rates Equal

We start recalling the case where all six mutation rates are equal, $\mu_{12} = \mu_{21} = \mu_{13} = \mu_{31} = \mu_{23} = \mu_{32}$, as mentioned in Section 5. Then the final steady-state becomes very simple

$$(p_1, p_2, p_3) \rightarrow \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) \text{ as } \tau \rightarrow \infty, \quad (43)$$

which is obvious because of the complete symmetry of the mutation process.

This is further emphasized in Fig. 3, where we study how the average frequencies are approaching the common asymptotic value of $1/3$ with increasing diffusion time τ for two different mutation parameters, expressed in diffusion variables. Our chosen founder population is inhomogeneous, with the frequencies $0.75, 0.25, 0$ for A, a and α , respectively. Thereby, we expose how the initial distribution is gradually wiped out until a final steady-state with equal distribution of the three gametes is obtained. The first subfigure (a) is the standard case $\beta = 0$ for the common mutation parameter. As expected, this standard mutation rate has the unit diffusion time $\tau = 1$ as time scale for decay of its initial state. The second subfigure (b) represents strong mutation rates where all six mutation parameters are chosen as $\beta = -1$. It has a shorter time scale because of its strong mutation rate.

Fig. 3 also demonstrates the convergence of the diffusion process by displaying different population sizes, choosing $n = 4$ as the smallest and $n = 64$ as the largest population size. It provides an assessment of the convergence of the diffusion approximation, even though we do not present a diffusion theory. In these plots, we first represent the smallest population size $n = 4$, where the discrete time steps of the exact Markov process are visible in the plots. Next, we double this size $n = 4$ four times up to $n = 64$ as

the highest included value. Each time step of one generation is shortened by one half for each doubled population size. There is a gradual convergence to an asymptotic limit with the diffusion time scale. The convergence for the allele A with increasing n is shown in a magnified frame with more detail.

Fig. 3. Mean allele frequencies as functions of the diffusion time ($\tau = t/n$), for different populations; $n = 4, n = 8, n = 16, n = 32$ and $n = 64$, respectively (where $\mu = 0.5(1-\beta)/n$). The start populations are $n_1 = 0.75n, n_2 = 0.25n$ and $n_3 = 0$, respectively. All 6 mutation rates are equal, yielding $\beta = \beta_{13} = \beta_{12} = \beta_{21} = \beta_{13} = \beta_{31} = \beta_{23} = \beta_{32}$, i.e., $\mu = \mu_{13} = \mu_{12} = \mu_{21} = \mu_{13} = \mu_{31} = \mu_{23} = \mu_{32}$. **a):** $\beta = 0$ **b):** $\beta = -1$.

6.2 Equal Two-Way Mutations

A basic case slightly more complicated than all six mutation rates equal, is given by

$$\mu_{12} = \mu_{21}, \quad \mu_{13} = \mu_{31}, \quad \mu_{23} = \mu_{32}. \quad (44)$$

There are three mutation rates: three sets of mutually equal mutations back and forth between any pair of alleles. The red flower A mutates to the blue flower a with the same mutation rate as the blue flower mutates back to the red

flower, and similarly for red/white flower and blue/white flower mutations.

This case is depicted in Fig. 4, where we show two examples of this type, with the small population size of $n = 4$. In comparison with the previous Fig. 3, we note the basic difference that there is no longer one single mutation rate that sets the time scale of decay from the initial state to the final steady state. The two subfigures of Fig. 4 show: (a) one rapid ($\beta = -1$) and two slow mutation rates ($\beta = 0$); and (b): two rapid and one slow mutation rate. In both cases, we have a population size of four haploid gametes. As before, the initial parental population has the distribution $(n_1, n_2, n_3) = (0.75n, 0.25n, 0)$ for the three gametes. These two plots do not differ much from one another, indicating that the slowest mutation dominates in the process. These two time-scales of decay are intermediates between the two cases shown in Fig. 3, where all mutations are either slow (a) or rapid (b).

Fig. 4. Mean allele frequencies as functions of the diffusion time ($\tau = t/n$) for a population size of $n = 4$ (where $\mu_{ij} = 0.5(1 - \beta_{ij})/n$). Here there are different mutation rates, but the mutual rates are equal ($\mu_{ij} = \mu_{ji}$) between each pair of alleles i and j . The parental population is given by $n_1 = 0.75n$, $n_2 = 0.25n$ and $n_3 = 0$. **a)**: $\beta_{13} = \beta_{31} = -1$ and $\beta_{12} = \beta_{21} = \beta_{23} = \beta_{32} = 0$ **b)**: $\beta_{13} = \beta_{31} = \beta_{23} = \beta_{32} = -1$ and $\beta_{12} = \beta_{21} = 0$.

6.3 A One-Way Ring of Mutations

We have just studied two-way mutations in the last subsection, where all six mutation rates were nonzero. Moreover, we have seen that the final steady-state will be symmetric in all three alleles (with $n_1 = n_2 = n_3$). It occurs whenever the rates of mutual mutations between two alleles are equal in both directions, back and forth.

Therefore, a particularly interesting case is the extreme opposite case when the mutation is only one way between any pair of alleles. This case has already been plotted Fig. 2. It can be called ring mutations, where all mutations go only one way in a ring: Gamete A delivers mutations only to a , while a delivers only to α and α delivers mutations only to A . The three nonzero mutation rates are $\mu_{21}, \mu_{32}, \mu_{13}$. The remaining mutation rates are zero

$$\mu_{12} = \mu_{23} = \mu_{31} = 0. \quad (45)$$

As a business interpretation of this ring mutation process, we may bear in mind a closed economy game between three partners. Partner A pays only to a , partner a pays only to α , and partner α pays only to A . The payment rate is proportional to what each individual owns, and it is proportional to the given mutation rate. The amount that each partner owns in the next time step is given by what he had plus his net gain (positive or negative) from the mutations in and out.

The ring mutations with three different mutation rates will produce three different allele frequencies as final steady states. In Fig. 5, we show two examples of ring mutations, which we link to the previous examples by choosing either $\beta = 0$ or $\beta = -1$. We start with a balanced founder population with equal allele numbers $n_1 = n_2 = n_3$ at $\tau = 0$, to show the evolving differences in allele frequencies. We show only two different mutation rates represented by $\beta = 0$ (standard mutation rate) and $\beta = -1$ (high mutation rate). The allele A receives more than it gives by mutations in both subfigures, while gamete α gives more than it receives. We show the small population size $n = 3$, and we include computations for the population sizes $n = 12$ and $n = 48$, indicating the asymptotic convergence to a diffusion model, in terms of the diffusion time variable τ .

As an intermediate business partner, allele a receives mutations at the same rate as it gives. Allele a is first stingy with a small rate of giving and taking ($\beta = 0$) in Fig. 5 (a), while in Fig. 5 (b) it behaves generously with a high rate of giving and taking ($\beta = -1$). In the early process, allele a places exactly in the middle between its rich partner A and poor partner α . However, the final average frequency of α shows that it finally pays off to be stingy in this closed mutation game.

Fig. 5. Mean allele frequencies as functions of the diffusion time ($\tau = t/n$), for different populations; $n = 3, n = 12$ and $n = 48$, respectively (where $\mu_{ij} = 0.5(1 - \beta_{ij})/n$). The start populations are $n_1 = n/3$, $n_2 = n/3$ and $n_3 = n/3$, respectively. Three mutation rates are zero, yielding $\mu_{12} = \mu_{31} = \mu_{23} = 0$, i.e., $\beta_{12} = \beta_{31} = \beta_{23} = 1$. The three other mutation rates are selected individually. **a)**: $\beta_{13} = -1$ and $\beta_{21} = \beta_{32} = 0$ **b)**: $\beta_{13} = \beta_{32} = -1$ and $\beta_{21} = 0$. Analytical solutions are calculated according to Table 1. They are depicted as straight lines indicating stationary frequencies.

7 THE DETAILED MARKOV PROCESS IN COMPOSITIONAL COORDINATES

So far, we have only studied the evolution of average frequencies. Thereby, we have maintained an overview with different mutation parameters. We have not yet extracted the full information of the exact Markov chains that we have computed. In this section, we will exemplify the detailed process in terms of the compositional variables x and y , representing the alleles A and a . The function $v_n(x, y, \tau)$ gives the probability of each population distribution (with n haploid members) in diffusion variables, assuming a continuous function $v_\infty(x, y, \tau)$ to exist as an asymptotic limit as $n \rightarrow \infty$ with all other parameters constant in terms of diffusion variables. The third compositional variable is given by eq. (22) as $z = 1 - x - y$. Thus z has a fixed value

at each point in the compositional (x, y) plane, so it can be disregarded in the analysis. The legal domain of variation for the compositional variables x and y is a triangle given by

$$0 \leq x \leq 1, \quad 0 \leq y \leq 1, \quad x + y \leq 1, \quad (46)$$

as demonstrated by Tier [17] and Baxter *et al.* [4].

Several other authors refer to an equilateral triangle for describing the tri-allelic compositional distribution in a balanced way. In a three-dimensional x, y, z space, a plane defined by $x + y + z = 1$ must introduced for describing the spatially two-dimensional diffusion process. A new pair of coordinates is then needed, since there can only be two independent compositional coordinates available for a haploid tri-allelic model. We choose the two compositional coordinates x and y for presenting the two-dimensional diffusion process. This means that we take a projection of the equilateral triangle (defined on the sloping plane $x + y + z = 1$ by $0 \leq x \leq 1, 0 \leq y \leq 1, 0 \leq z \leq 1$) into a right-angle isosceles triangle in the x, y plane. Now z becomes a hidden coordinate, and its role is merely to restrict the legal domain of variation within the compositional x, y plane. Since the number of alleles α is always between all and none ($1 \geq z \geq 0$), it follows from the definition $z = 1 - x - y$ that

$$0 \leq x + y \leq 1, \quad (47)$$

which is already included in the general relationships (46). There are only compositional coordinates in a two-dimensional plane, so we chose to work graphically with x and y .

Fig. 6 shows three stages of the probability distribution $v_n(x, y, \tau)$ in a population with 16 individuals. We show contour plots in the x, y plane on the left hand side, with the corresponding surface plots on the right hand side, with snapshots at $\tau = 0.25, \tau = 1$ and $\tau = 4$. The standard value $\beta = 0$ is chosen, with all six mutation rates equal. Our plots are similar to those of Tier [17] for the same TA-WFM process, based on the diffusion model. While our Markov model is exact and discrete, its only notable difference is the discontinuity in the gradient near the boundary of the triangular domain.

The founder population has 12 gametes of allele A , and 4 gametes of allele a . It is marked as a red dot in Fig. 6, located at the boundary which represents $z = 0$, since the founder population is chosen with zero α alleles. At the small time $\tau = 0.25$ for this subfigure, we see how the probability distribution has diffused out from this initially concentrated point $(x, y) = (0.75, 0.25)$, which is still the maximal point. At a time scale smaller than the diffusion time unit, the distribution is dominated by the chosen founder population. The other subfigures show how the maximum point is drifting towards the centre of the triangle at the diffusion time scale. At $\tau = 4$ the maximum point is close to its final location at $(x, y) = (1/3, 1/3)$, where also $z = 1/3$ due to symmetric mutations between the three alleles.

Fig. 6 shows the discrete Markov chain where the surface plots in subfigure (b) has no approximations, representing all the calculated point values. The contour plots in subfigure (a) need linear interpolations for the isolines to represent constant increments. We note the steep gradients near the

Fig. 6. Evolving probability distribution contours with corresponding surface plots in the x, y plane, for the discrete function $v_{16}(x, y)$ representing the exact Markov process. The given population size is $n = 16$ with mutation rate $\mu = 0.5(1 - \beta)/n$. The parental population is chosen as $n_1 = 0.75n$, $n_2 = 0.25n$ and $n_3 = 0$. All six mutation parameters are chosen to be equal: $\beta = 0 = \beta_{13} = \beta_{12} = \beta_{21} = \beta_{13} = \beta_{31} = \beta_{23} = \beta_{32}$, which also gives equal mutation rates $\mu = \mu_{13} = \mu_{12} = \mu_{21} = \mu_{13} = \mu_{31} = \mu_{23} = \mu_{32}$. We show snapshots at three diffusion times **a)**: $\tau = 1/4$ **b)**: $\tau = 1$. **c)**: $\tau = 4$.

boundaries, which is due to the fact that the WFM does not represent a nearest-neighbor interaction. The case plotted in Fig. 6 is chosen for its similarity with the computations shown by Tier [17] for the tri-allelic diffusion model. His model was continuous and applied several asymptotic approximations, while our model is discrete and represents the exact Markov process. The continuous diffusion model [17] gives smooth continuity near the boundaries, in contrast to our discrete computations with 16 individuals. However, the convergence of our model to the diffusion limit is good for sample points that are not too close to the triangular boundary for the diffusion variables.

In Fig. 6 the initial state is chosen at the boundary of the computational domain, with $z = 0$ at $\tau = 0$. Thereby a stronger genetic drift is achieved than in the comparable diffusion model computations by Tier [17]. The final steady state is approached at our last snapshot $\tau = 4$, but it reveals peculiarities different from the diffusion model: There are isolines that intersect the boundary, which means that the boundary is not a isoline for the probability distribution.

Fig. 7. Evolving probability distribution contours with corresponding surface plots in the x, y plane, for the discrete function $v_{128}(x, y)$ representing the exact Markov process. The given population size is $n = 128$ with mutation rate $\mu = 0.5(1 - \beta)/n$. The parental population is chosen as $n_1 = 0.75n$, $n_2 = 0.25n$ and $n_3 = 0$. All six mutation parameters are chosen to be equal: $\beta = 0 = \beta_{13} = \beta_{12} = \beta_{21} = \beta_{13} = \beta_{31} = \beta_{23} = \beta_{32}$, which also gives equal mutation rates $\mu = \mu_{13} = \mu_{12} = \mu_{21} = \mu_{13} = \mu_{31} = \mu_{23} = \mu_{32}$. We show snapshots at three diffusion times **a)**: $\tau = 1/4$ **b)**: $\tau = 1$. **c)**: $\tau = 4$.

There are notable finite jumps from the boundary points to their nearest neighbors, while all internal points show a gentle variation of the probability distribution.

It is natural to think that the small population size $n = 16$ in Fig. 6 is to blame for the peculiarities near the boundary. For resolving this dilemma we have added Fig. 7 where the same exact MCA is carried out for the much greater population size $n = 128$, with all diffusion variables equal to Fig. 6. We realize from this figure that these peculiarities remain. Many isolines for the probability distribution will still intersect the boundary, and the finite jumps in the probability distribution from the boundary points to their nearest neighbors do not vanish. No convergence to a smooth probability distribution embracing the non-absorbing boundary seems to take place as the population size is increased. Diffusion theory seems unable to handle these finite jumps in probability distribution at a non-absorbing boundary, in a tri-allelic mutation model where these jumps show variation in magnitude along each boundary. The exact Markov chain does not fit with the

Fig. 8. Normalized probability plots in terms of z , with $x = y$, where $z = 1 - x - y$, for different population sizes (where $\mu = 0.5(1 - \beta)/n$). The start populations are $n_1 = 0.75n$, $n_2 = 0.25n$ and $n_3 = 0$, respectively. All 6 mutation rates are equal, yielding $\beta = 0 = \beta_{13} = \beta_{12} = \beta_{21} = \beta_{13} = \beta_{31} = \beta_{23} = \beta_{32}$, i.e., $\mu = \mu_{13} = \mu_{12} = \mu_{21} = \mu_{13} = \mu_{31} = \mu_{23} = \mu_{32}$. **a)**: $\tau = 1/4$ **b)**: $\tau = 1$. **c)**: $\tau = 4$. **d)**: $\tau = 16$.

known diffusion approximations, suggesting a scrutiny of the underlying assumptions and principles.

In Figs. 6 (c) and 7 (c) we show cuts along the dashed line $x = y$, as a preparation for the subsequent plots in Fig. 8, where we show in closer detail the convergence to the diffusion limit. We visualize each discrete function $v_n(x, y, \tau)$, by showing the variation with increasing n only along the line $x = y$. Thereby we track each function $v_n(z, \tau)$ in terms of a single compositional coordinate coinciding with the hidden coordinate $z = 1 - x - y = 1 - 2x$, thus representing the frequency of the allele α in the gamete pool. Fig. 8 shows the same evolving problem in diffusion variables as the previous Figs. 6 and 7, but now it is expressed in one compositional dimension as $v_n(z, \tau)$, along the line $x = y$. Moreover, we include several other population sizes, starting with $n = 4$ and ending with $n = 64$, so that the previous case of $n = 16$ appears as an intermediate size. Here we expose in detail compromised convergence to the diffusion limit for the sample points closest to the boundary. We note that the exact boundary points stick out and do not exhibit similar convergence of its neighboring points, even though the boundaries are non-singular according to diffusion theory. Diffusion theory gives Dirac singularities only at an absorbing boundary with zero mutation rates away from that boundary.

The first three subfigures show the standard diffusion variable $v_n(z, \tau) \approx v_\infty(z, \tau)$ defined by eq. (23), while in Fig. 8 (d) we introduce a modified diffusion variable $v_n^*(z, \tau)$ defined by eq. (25). This modified diffusion variable takes the exact number of different population compositions into account, which gives an obvious improvement with respect to conservation of probability as n increases. Some visual clarity is lost, however, as the approach to the diffusion limit is no longer monotonous as the population size n increases.

TABLE 2
Convergence results of the dependent diffusion variables $v_n(x, y, \tau)$ and $v_n^*(x, y, \tau)$ studied at the triangle boundaries represented by $z = 0$ and $z = 1$, along the symmetry line $x = y$, where $z = 1 - x - y$, according to the exact WFM, with diffusion variables given in eqs. (23) and (25), respectively. All 6 mutation rates are equal, with $\beta = 0$. The convergence is studied for four values for the diffusion time coordinate $\tau = 1/4$, $\tau = 1$, $\tau = 4$, and $\tau = 16$, respectively. The initial allele frequencies are $p_1 = 0.75$, $p_2 = 0.5$, and $p_3 = 0$, respectively.

n	τ	$v_n(x, x, \tau)$		$v_n^*(x, x, \tau)$	
		z=0	z=1	z=0	z=1
8	1/4	1.955603	0.000248	2.750066	0.000349
16		2.573576	0.000053	3.076227	0.000064
32		3.116841	0.000017	3.415133	0.000018
64		3.534722	0.000007	3.702139	0.000008
128		3.823457	0.000004	3.913536	0.000004
8	1	0.866002	0.180906	1.217816	0.254398
16		1.025570	0.186444	1.225876	0.222859
32		1.148361	0.193811	1.258263	0.212359
64		1.234813	0.200560	1.293298	0.210059
128		1.291559	0.205825	1.321987	0.210674
8	4	0.651904	0.413090	0.916740	0.580907
16		0.730224	0.519359	0.872846	0.620796
32		0.791365	0.611737	0.867102	0.670282
64		0.835108	0.682454	0.874662	0.714777
128		0.864279	0.731649	0.884641	0.748886
8	16	0.650572	0.414869	0.914867	0.583410
16		0.727890	0.522788	0.870056	0.624895
32		0.788236	0.616688	0.863673	0.675707
64		0.831433	0.688573	0.870813	0.721186
128		0.860261	0.738566	0.880529	0.755966

Tables 2, 3, and 4 give numerical results for the peculiar behavior near the boundary, corresponding to Fig. 8 with its choice of parameters. Values for the probability distribution are tabulated for sample points along the mid-line $z = 1 - x - y$ of the compositional triangle. All these tables present the two versions for the population distribution in diffusion variables. Tables 3 and 4 show the spatial convergence for the closest neighboring points to the boundary at each population size as the population size increases. There is convergence between these internal points, but we cannot see a convergence between the boundary point and its closest neighbor. Table 2 is devoted to the behavior of the boundary points themselves for different diffusion times.

TABLE 3

Convergence results of the dependent diffusion variables $v_n(x, y, \tau)$ and $v_n^*(x, y, \tau)$ studied for $z \rightarrow 0$ with $x = y$, where $z = 1 - x - y$, in the exact WFM, given in eqs. (23) and (25), respectively. All 6 mutation rates are equal, with $\beta = 0$. The rescaled time variable is $\tau = 4$, and the initial allele frequency were $p_1 = 0.75$, $p_2 = 0.5$, and $p_3 = 0$, respectively.

n		z=0	z= 1/64	z=1/32	z=1/16	z=1/8	z=1/4
8	$v_8(x, x, 4)$	0.6519					0.9084
16	$v_{16}(x, x, 4)$	0.7302				0.9256	0.9569
32	$v_{32}(x, x, 4)$	0.7914			0.9405	0.9636	0.9794
64	$v_{64}(x, x, 4)$	0.8351		0.9577	0.9713	0.9830	0.9904
128	$v_{128}(x, x, 4)$	0.8643	0.9730	0.9807	0.9872	0.9930	0.9959
8	$v_8^*(x, x, 4)$	0.9167					1.2774
16	$v_{16}^*(x, x, 4)$	0.8728				1.1064	1.1438
32	$v_{32}^*(x, x, 4)$	0.8671			1.0305	1.0558	1.0732
64	$v_{64}^*(x, x, 4)$	0.8747		1.0031	1.0173	1.0295	1.0374
128	$v_{128}^*(x, x, 4)$	0.8846	0.9959	1.0038	1.0105	1.0164	1.0193

TABLE 4

Convergence results of the dependent diffusion variables $v_n(x, y, \tau)$ and $v_n^*(x, y, \tau)$ studied for $z \rightarrow 1$ with $x = y$, where $z = 1 - x - y$, in the exact WFM, given in eqs. (23) and (25), respectively. All 6 mutation rates are equal, with $\beta = 0$. The rescaled time variable is $\tau = 4$, and the initial allele frequency were $p_1 = 0.75$, $p_2 = 0.5$, and $p_3 = 0$, respectively.

n		z=1-1/4	z=1-1/8	z=1-1/16	z=1-1/32	z=1-1/64	z=1
8	$v_8(x, x, 4)$	0.7685					0.4131
16	$v_{16}(x, x, 4)$	0.8607	0.7969				0.5194
32	$v_{32}(x, x, 4)$	0.9242	0.8723	0.8494			0.6117
64	$v_{64}(x, x, 4)$	0.9579	0.9293	0.9006	0.9011		0.6825
128	$v_{128}(x, x, 4)$	0.9756	0.9595	0.9444	0.9285	0.9420	0.7316
8	$v_8^*(x, x, 4)$	1.0807					0.5809
16	$v_{16}^*(x, x, 4)$	1.0288	0.9526				0.6208
32	$v_{32}^*(x, x, 4)$	1.0126	0.9558	0.9307			0.6703
64	$v_{64}^*(x, x, 4)$	1.0032	0.9734	0.9433	0.9438		0.7148
128	$v_{128}^*(x, x, 4)$	0.9986	0.9821	0.9666	0.9504	0.9642	0.7489

8 SUMMARY AND CONCLUSIONS

We have studied the discrete stochastic process of tri-allelic random mating with a full mutation matrix according to the WFM. We have performed computations of the exact Markov chain with a rectangular transition matrix that allows bottlenecks in the population size. However, we have not included computations of bottlenecks in the presented results, primarily because bottlenecks cannot be described in terms of diffusion variables.

The literature lacks exact formulas for the time-evolution of tri-allelic mutation models. Therefore we could not benchmark our exact MCA of population composition evolution with respect to results from diffusion theory. Yet, we have applied diffusion variables for the presentation of our results, strongly suggesting convergence to a diffusion limit for haploid population sizes up to 128 individuals, at least for compositional coordinates not too close to their triangular boundary.

We have performed a broad study of the evolution of average frequencies for various cases of different mutation rates. We have picked the interesting case of ring mutations where each of the three alleles have one incoming mutation and one outgoing mutation, amounting to three nonzero mutation rates in combination with three that are equal to zero. The results for the average frequencies rest on numerical computations of the underlying exact Markov chain.

Numerical evaluations of the exact Markov chain are presented in full detail only for the case where all six mutation rates are equal. The stochastic process takes place over a triangular domain for the compositional coordinates. As

expected from diffusion theory, the internal sample points for the triangle behave smoothly, but there is a discontinuity between the boundary points and its closest neighboring sample point. The convergence to diffusion theory is also compromised as we approach the boundary of the compositional domain. This not so surprising because the WFM is not a nearest-neighbor interaction. Each boundary point is then different from the internal points as it does not have neighboring points on all sides to make their influence during one time-step, from one generation to the next.

While the literature does not give exact analytical solutions for the evolution of multi-allelic diffusion models with mutations, some formulas for asymptotic steady states are known [20]. Our single contribution to the literature of exact asymptotic results is the equilibrium state of all three alleles defined by the full mutation matrix. The rest of our exact results is limited to smaller populations as it rests on the heavy numerical Markov chain algebra. A fundamental difficulty for diffusion models with mutations is to formulate valid boundary conditions to be applied at the boundaries of the compositional domains, as discussed by Tier [17] and Baxter *et al.* [4]. This dilemma does not exist in our exact Markov model, which does not have boundary conditions. At a non-absorbing boundary, this leads to qualitative differences compared with diffusion models: We do not find smooth convergence to the boundary points. Instead of smoothness, we find a finite jump discontinuity between the boundary and its nearest neighbor, which remains finite as the population size increases. A number of isolines for the population distribution will intersect the boundary, different from the diffusion models [17] where all computed isolines tend to

be closed curves within the computational domain.

We have presented the results in diffusion variables, and have demonstrated in detail that this is a consistent approach. Ironically we have not been able to find sufficient accurate results for benchmarking our MCA with diffusion models. Diffusion models suffer from the serious obstacle that boundary conditions are needed at non-absorbing boundaries. Our exact Markov model does not have boundary conditions, which indicates that benchmarking can go the other way round. The difficulties with developing consistent diffusion models indicate that they should be benchmarked with respect to the exact underlying discrete model. In statistical physics, it seems clear that an underlying microscopic discrete model should serve as verification of a continuous macroscopic model, as far as possible. Population genetics is different in the sense that diffusion models appear as more modern than the older models involving combinatorics. However, as long as the process itself is discrete, the exact MCA is always superior as far as it can be computed with numerical convergence.

With our present focus on the exact MCA and its representation in diffusion variables, we have left several important topics untouched. Selection is not considered at all in our algebraic model. Bottlenecks where the prescribed population size changes from generation to generation are contained in our general formulation, but not included in the computations. Extinction does not occur in the computations we have presented. Absorbing boundaries with exact fixation are also omitted since we have chosen nonzero mutation rates both away from and into each of the three alleles. Thereby one might expect smooth population probability distributions as the population size increases, and this is what we find at all internal sample points for any population size. However, next to the boundary, there are finite jumps between the probability at the non-absorbing boundary and its nearest neighbor, which may also indicate why it is so difficult to establish boundary conditions at a non-absorbing boundary for multi-allelic diffusion models.

REFERENCES

- [1] S. Wright, "The differential equation of the distribution of gene frequencies," *Proc. Nat. Acad. Sci. United States Am.*, vol. 31, no. 12, p. 382, 1945.
- [2] M. Kimura, "Solution of a process of random genetic drift with a continuous model," *Proc. Nat. Acad. Sci. United States Am.*, vol. 41, no. 3, p. 144, 1955.
- [3] W. J. Ewens, "Numerical results and diffusion approximations in a genetic process," *Biometrika*, vol. 50, no. 3/4, pp. 241–249, 1963.
- [4] G. J. Baxter, R. A. Blythe, and A. J. McKane, "Exact solution of the multi-allelic diffusion model," *Math. Biosci.*, vol. 209, no. 1, pp. 124–170, 2007.
- [5] P. A. Tyvand, "An exact algebraic theory of genetic drift in finite diploid populations with random mating," *J. Theoretical Biol.*, vol. 163, no. 3, pp. 315–331, 1993.
- [6] B. Franzke and B. Kosko, "Noise can speed convergence in markov chains," *Phys. Rev. E*, vol. 84, no. 4, p. 041112, 2011.
- [7] —, "Using noise to speed up markov chain monte carlo estimation," *Procedia Comp. Sci.*, vol. 53, pp. 113–120, 2015.
- [8] —, "Noise can speed markov chain monte carlo estimation and quantum annealing," *Phys. Rev. E*, vol. 100, no. 5, p. 053309, 2019.
- [9] D. Waxman, "The diffusion equation of random genetic drift—biology's analogue of the schrödinger equation?" *Contemp. Phys.*, vol. 58, no. 3, pp. 253–261, 2017.
- [10] D. Schrempf and A. Hobolth, "An alternative derivation of the stationary distribution of the multivariate neutral wright–fisher model for low mutation rates with a view to mutation rate estimation from site frequency data," *Theoretical Population Biol.*, vol. 114, pp. 88–94, 2017.
- [11] R. A. Fisher, "Xxi.—on the dominance ratio," *Proc. Royal Soc. Edinburgh*, vol. 42, pp. 321–341, 1923.
- [12] S. Wright, "Evolution in mendelian populations," *Genetics*, vol. 16, no. 2, p. 97, 1931.
- [13] A. McKane and D. Waxman, "Singular solutions of the diffusion equation of population genetics," *J. Theoretical Biol.*, vol. 247, no. 4, pp. 849–858, 2007.
- [14] S. Karlin and J. McGregor, "On a genetics model of moran," in *Math. Proc. Cambridge Phil. Soc.*, vol. 58, no. 2. Cambridge University Press, 1962, pp. 299–311.
- [15] P. A. Tyvand and S. Thorvaldsen, "Exact markov chains versus diffusion theory for haploid random mating," *Math. Biosci.*, vol. 225, no. 1, pp. 18–23, 2010.
- [16] M. Kimura, "Random genetic drift in a tri-allelic locus; exact solution with a continuous model," *Biometrics*, vol. 12, no. 1, pp. 57–66, 1956.
- [17] C. Tier, "A tri-allelic diffusion model with selection, migration, and mutation," *Math. Biosci.*, vol. 44, no. 1-2, pp. 41–60, 1979.
- [18] Z.-B. Zeng, H. Tachida, and C. C. Cockerham, "Effects of mutation on selection limits in finite populations with multiple alleles," *Genetics*, vol. 122, no. 4, pp. 977–984, 1989.
- [19] O. Hössjer, P. A. Tyvand, and T. Miloh, "Exact markov chain and approximate diffusion solution for haploid genetic drift with one-way mutation," *Math. Biosci.*, vol. 272, pp. 100–112, 2016.
- [20] S. Wright, *Evolution and the Genetics of Populations, Volume 2: Theory of gene frequencies*. University of Chicago Press, 1984, vol. 2.

Peder A. Tyvand received a Ph.D. degree in theoretical hydrodynamics from the University of Oslo in 1980 and is a professor in physics at the Norwegian University of Life Sciences since 1993. His research in fluid mechanics focuses on water waves and thermal convection. He is currently pursuing cross-disciplinary applications of deterministic and stochastic Markov processes.

Jonas Kristiansen Nøland (S'14-M'17) was born in Drammen, Norway, in 1988. He received the Ph.D. degree in engineering physics from Uppsala University, Uppsala, Sweden, in 2017. Since 2018, he has been an Associate Professor with the Norwegian University of Science and Technology. Dr. Nøland serves as an Editor for the IEEE TRANSACTIONS ON ENERGY CONVERSION and as an Associate Editor for the IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS.