A Markov chain description of the stepwise mutation model: Local and global behaviour of the Allele process

Amke Caliebe, Arne Jochens, Michael Krawczak, Uwe Rösler

www.elsevier.com/locate/yjtbi

A Markov chain description of the stepwise mutation model: Local and global behaviour of the Allele process

Amke Caliebe, Arne Jochens, Michael Krawczak, Uwe Rösler

PII: S0022-5193(10)00324-3
DOI: doi:10.1016/j.jtbi.2010.06.033
Reference: YJTBI 6053

To appear in: *Journal of Theoretical Biology*

Received date: 2 February 2010
Revised date: 24 June 2010
Accepted date: 24 June 2010

Cite this article as: Amke Caliebe, Arne Jochens, Michael Krawczak and Uwe Rösler, A Markov chain description of the stepwise mutation model: Local and global behaviour of the Allele process, *Journal of Theoretical Biology*, doi:10.1016/j.jtbi.2010.06.033

A Markov Chain Description of the Stepwise Mutation Model: Local and Global Behaviour of the Allele Process

Amke Caliebe [a,b,*] Arne Jochens [a,b,c] Michael Krawczak [b]

Uwe Rösler [c]

[a] *equal contribution of authors*

[b] *Institut für Medizinische Informatik und Statistik, Haus 31, Christian-Albrechts-Universität Kiel, Arnold-Heller-Str. 3, 24105 Kiel, Germany, caliebe@medinfo.uni-kiel.de, jochens@medinfo.uni-kiel.de, krawczak@medinfo.uni-kiel.de*

[c] *Mathematisches Seminar, Ludewig-Meyn-Str. 4, Christian-Albrechts-Universität Kiel, 24098 Kiel, Germany, roesler@math.uni-kiel.de*

Abstract

The stepwise mutation model (SMM) is a simple, widely used model to describe the evolutionary behaviour of microsatellites. We apply a Markov chain description of the SMM and derive the marginal and joint properties of this process. In addition to the standard SMM, we also consider the normalised allele process. In contrast to the standard process, the normalised process converges to a stationary distribution. We show that the marginal stationary distribution is unimodal. The standard and normalised processes capture the global and the local behaviour of the SMM, respectively.

Key words: microsatellite, evolution, population genetics, asymptotic behaviour

* Corresponding author. Address: Dr. Amke Caliebe, Institut für Medizinische Informatik und Statistik, Haus 31, Christian-Albrechts-Universität Kiel, Arnold-Heller-Str. 3, 24105 Kiel, Germany, Phone: +49 (431) 597 - 3199, Fax: +49 (431) 597 - 3193, E-Mail: caliebe@medinfo.uni-kiel.de

1 Introduction

Microsatellites are successive iterations of a given short DNA sequence motif (usually 2-6 nucleotides long) that is repeated 5-100 times (Tautz, 1993; Chambers and MacAvoy, 2000). The number of iterations (the "repeat number") serves to identify a given microsatellite allele. Microsatellites are abundant in many species and have very high mutation rates (up to 10^{-2} per generation, Li et al., 2002). Owing to their high degree of variability, microsatellites are frequently used as markers in population genetics (Goldstein et al., 1999; Kashi and King, 2006), DNA fingerprinting (Cassidy and Gonzales, 2005; Bindu et al., 2007), whole genome mapping (Weissenbach et al., 1992) and genetic epidemiology (Thibodeau et al., 1993; Ashley and Warren, 1995).

The stepwise mutation model (SMM) was first introduced by Ohta and Kimura (1973) to describe the behaviour of electrophoretically detectable alleles in a population. Since then, the SMM has been widely used for modelling microsatellite mutation and evolution (Tishkoff et al., 1996; Zhivotovsky et al., 2003; De Iorio et al., 2005; Vardo and Schall, 2007). The SMM assumes that, in one generation, the repeat number can only increase or decrease by at most one, usually with equal probability. More refined models have been proposed that include mutations of greater length, mutation rates that depend upon repeat number, or the additional introduction of point mutations (Di Rienzo et al., 1994; Garza et al., 1995; Feldman et al., 1997; Zhivotovsky et al., 1997;

Kruglyak et al., 1998; Durrett and Kruglyak, 1999; Falush and Iwasa, 1999; Calabrese et al., 2001); for an overview, see Watkins (2007) or Calabrese and Sainudiin (2005). As yet, however, it has remained controversial to what extent these models approximate the reality (Chambers and MacAvoy, 2000; Whittaker et al., 2003; Sainudiin et al., 2004; Cornuet et al., 2006).

In the following, we will consider the classical SMM. In 1975, Moran discovered that the distribution of the absolute frequencies $n_i(t)$ of alleles (as identified by their repeat number i) at time t does not converge, but has bounded variance. He subsequently conjectured that the distribution "remains in a bunch" and characterised its behaviour as "wandering", without being more specific as to the existence of a limiting distribution (Moran, 1975). To investigate convergence, Moran considered quantities $C_k(t) := N^{-2} \sum_i n_i(t) n_{i+k}(t)$, where N is the population size. For $k = 0$, this is the "effective number of neutral alleles in the population" of Ohta and Kimura (1973). Moran was able to show that "unlike most problems in population genetics that have been discussed in the past, we do not obtain a limiting distribution or convergence in probability [of $C_k(t)$]." (Moran, 1975). Shortly after Moran's publication, Kingman investigated the normalised Markov chain of the SMM, given by the repeat number difference to the allele of the N-th (or any other) individual in each generation (Kingman, 1976). Using characteristic functions, he could prove exponentially fast convergence in distribution for a generalised model. He also obtained results about the limiting distribution of samples from a population when the

population size tends to infinity conditioned that a certain relationship between time and population size holds.

Here, we will give a detailed analysis of the behaviour of the allele process under the SMM, where our focus will be upon the resulting Markov chain. Markov processes have been applied before to the characterisation of microsatellite mutation models by Watkins (2007). In contrast to Kingman (1976), who used the analytic tool of characteristic functions, we will apply the stochastic method of recurrence of Markov chains. In Section 2 we will introduce the stepwise mutation model which is the basis for all subsequent results. In Section 3 the allele process X is investigated. We will make use of the fact that every population which does not die out, such as under a Wright-Fisher model, contains a genealogical lifeline that does not die out. Adding independent mutations to the genealogy generates an inherent random walk, and thereby results for the marginal distribution of X. In the second subsection, we will show that X is an irreducible, aperiodic and null recurrent Markov chain. The behaviour of X represents the global aspect of the SMM. The normalised allele process V is analysed in Section 4 characterising the local view of the SMM. Again, marginal results such as moments and exponential moments will be given in the first subsection. Then, it will be proven that V is a positive recurrent Markov chain with exponentially fast convergence to the invariant distribution. A central result is provided in the third subsection where it will be shown that the marginal invariant distribution is

unimodal. Finally, some simulation results for this distribution are given.

2 Wright-Fisher Model with Stepwise Mutations

The microsatellite allele process under neutral evolution will be studied using a Wright-Fisher model with stepwise mutation. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be the underlying probability space, and let $N \in \mathbb{N} := \{1, 2, 3, ...\}$ be the constant population size. A microsatellite allele will be represented by the number of iterations of the sequence motif, the repeat number. Alleles are normalised such that allele 0 corresponds to a particular basic state $m \in \mathbb{N}$, e.g. the most commonly observed repeat number. For simplicity in the classical SMM, which we apply here, there are no length restrictions on the allele size and even negative repeat numbers are theoretically possible. Thus, the set of possible alleles equals $\mathbb{Z}$, and an allele $z \in \mathbb{Z}$ then has repeat number $m + z$.

a) Genealogy

The genealogy is assumed to be given by a Wright-Fisher model. For ease of notation and terminology, we will consider only haploid individuals. However, our results can easily be transferred to diploid individuals by regarding each of their two alleles separately.

Let $Y_n(i)$ be the direct ancestor of the ith individual in the nth generation, $i \in \{1, ..., N\}$, $n \in \mathbb{N}$. Clearly, $Y_n(i)$ is an individual of the $(n-1)$th generation. According to the Wright-Fisher model, $Y_n : \Omega \to \{1, ..., N\}^N$ satisfies

$\mathbb{P}(Y_n(i) = j) = \frac{1}{N}$ for all $i, j \in \{1, ..., N\}$ and the $(Y_n(i))_{i\in\{1,...,N\},n\in\mathbb{N}}$ are independent.

b) Mutation process

Let $Z_n(i)$ be the mutational event preceding inheritance, from $Y_n(i)$, of the allele of the ith individual in the nth generation. We only consider mutation events that either increase or decrease the repeat number by 1, or leave the repeat number unchanged, i.e. $Z_n(i) \in \{-1, 0, 1\}$. Let $\mu \in (0, 1)$ be the mutation rate, i.e. the probability of a change in repeat number per generation and per individual. Then, $Z_n : \Omega \to \{0, 1, -1\}^N$ is assumed to satisfy $\mathbb{P}(Z_n(i) = 0) = 1 - \mu$, $\mathbb{P}(Z_n(i) = 1) = \mathbb{P}(Z_n(i) = -1) = \mu/2$.

As usual, we assume an *Independence Property* for the genealogical and mutational processes, namely that

$$\text{the whole family } Z_n(i), Y_n(i), n \in \mathbb{N}, i \in \{1, ..., N\} \text{ is independent.} \quad (1)$$

c) Allele process

Let $X_n(i)$ denote the allele of the ith individual in the nth generation. For all $n \in \mathbb{N}_0$, $X_n : \Omega \to \mathbb{Z}^N$ can be written as

$$X_n(\omega)(i) := X_{n-1}(\omega)(Y_n(\omega)(i)) + Z_n(\omega)(i) \qquad \text{with} \qquad X_0 \equiv 0\,. \quad (2)$$

$X := (X_n)_{n\in\mathbb{N}_0}$ is called the *allele process* of the Wright-Fisher SMM. The distribution of the initial states X_0 is arbitrary and does not influence the asymptotic behaviour. For the sake of simplicity, we assume $X_0 \equiv 0$ which means that all alleles have the same repeat number m at time 0.

d) Fundamental properties of the allele process X

Let $\mathcal{A}_n := \sigma(Y_1, ..., Y_n, Z_1, ..., Z_n)$ be the σ algebra generated by $Y_1, ..., Y_n, Z_1, ..., Z_n$. Then the following property follows directly from the definition of X.

Proposition 1

(i) *X_n is $\mathcal{A}_n$-measurable for all $n \in \mathbb{N}$.*

(ii) *For all $n \in \mathbb{N}$ the family $(X_n(i))_{i\in\{1,...,N\}}$ is exchangeable.* (Exchangeability Property)

3 Global Behaviour: the Allele Process X

3.1 Marginal Properties of X

To investigate the marginal distribution of the allele process X, we will use an immanent random walk. This is generated by the "lifeline" of the genealogy, i.e. the line of descent that never dies out. J_n is the index, in generation n, of the (unique) member of the lifeline.

Proposition 2

(i) *There exists an almost surely unique $J : \Omega \to \{1, ..., N\}^{\mathbb{N}}$ such that $Y_n(J_n) = J_{n-1}$ for all $n \in \mathbb{N}$, and J_n is $\sigma\left(Y_k, k \in \mathbb{N}, k > n\right)$ measurable. Furthermore, for $n \in \mathbb{N}$, $X_n(J_n)$ has the same distribution as $X_n(1)$.*

(ii) *$(X_n(J_n))_{n\in\mathbb{N}_0}$ is a random walk. For $k \in \mathbb{Z}$, the transition probabilities are $\mathbb{P}(X_n(J_n) = k|X_{n-1}(J_{n-1}) = k) = 1-\mu$ and $\mathbb{P}(X_n(J_n) = k-1|X_{n-1}(J_{n-1}) =$*

$k) = \mathbb{P}(X_n(J_n) = k+1 | X_{n-1}(J_{n-1}) = k) = \mu/2.$

Proof. (ii) follows from the definition of X once the existence of J has been established. Let τ_n be the first generation (after n) in which all individuals have a common ancestor in generation n, i.e.

$$\tau_n := \inf\{k > n : \forall i, j \in \{1, ..., N\}\ Y_{n+1} \circ ... \circ Y_k(i) = Y_{n+1} \circ ... \circ Y_k(j)\}\ .$$

τ_n is $\sigma\left(Y_k, k \in \mathbb{N}, k > n\right)$ measurable and almost surely finite.

Then, for $n \in \mathbb{N}_0$, define on $\tau_n < \infty$

$$J_n := Y_{n+1} \circ Y_{n+2} \circ ... \circ Y_{\tau_n - 1} \circ Y_{\tau_n}(1)\ .$$

J_n is almost surely well defined. For $\tau_n = \tau_{n-1}$ the equality $Y_n(J_n) = J_{n-1}$ is clear. For $\tau_n > \tau_{n-1}$ define $Z := Y_{\tau_{n-1}+1} \circ ... \circ Y_{\tau_n}(1)$. Then

$$Y_n(J_n) = Y_n \circ Y_{n+1} \circ ... \circ Y_{\tau_{n-1}}(Z) = J_{n-1}\ .$$

Hence J_n satisfies the required properties. □

The first and second moment of the marginal distribution of X and a recurrence equation follow immediately from this proposition and from Prop. 1(i). A limit result for the first absolute moment can be derived by applying the central limit theorem to the random walk $(X_n(J_n))_{n \in \mathbb{N}_0}$. For all $n \in \mathbb{N}$, $z \in \mathbb{Z}$ define

$$\rho_n(z) := \mathbb{P}(X_n(i) = z)\ .$$

Note that, owing to the exchangeability property of Prop. 1(ii), $\rho_n(z)$ is independent of the choice of $i \in \{1, ..., N\}$.

Lemma 3 *For any* $i \in \{1, ..., N\}, n \in \mathbb{N}$

(i) $\mathbb{E}\,(X_n(i)) = 0$

(ii) $\mathrm{Var}\,(X_n(i)) = \mu\, n$

(iii) $\rho_n(z) = (1-\mu)\rho_{n-1}(z) + \mu/2\,(\rho_{n-1}(z-1) + \rho_{n-1}(z+1))$

(iv) $\lim\limits_{m\to\infty} \mathbb{E}\,|X_m(i)| = \infty$

Note that $\lim_{n\to\infty} \mathrm{Var}\,(X_n(i)) = \infty$.

Lemma 4 *For any* $i, j \in \{1, ..., N\}, i \neq j, n \in \mathbb{N}$

$$\mathrm{Cov}\,(X_n(i), X_n(j)) = \mu\left(n + \frac{(N-1)^n - N^n}{N^{n-1}}\right) .$$

A proof of Lemma 4 is given in the appendix.

3.2 Characterisation of X as a Markov chain

The following theorem shows that, in our new representation as a Markov chain, the allele process X is null recurrent (see Breiman (1992), p. 140, for the definition of null recurrent). Therefore, no asymptotic distribution exists. In the following, we will write 0_N for $(0, ..., 0) \in \mathbb{Z}^N$. For the definition of $\mathcal{A}_n$, see Prop. 1.

Theorem 5

(i) The allele process X is an irreducible, aperiodic Markov chain on $\mathbb{Z}^N$ with respect to $(\mathcal{A}_n)_{n\in\mathbb{N}_0}$.

(ii) The allele process X is null recurrent.

Proof. (i) follows directly from the definition of X. For the proof of the recurrence, it suffices to verify recurrence for state $0_N \in \mathbb{Z}^N$ because of irreducibility. Remember that $X_0 \equiv 0_N$. We will prove the criterion $\sum_{n=1}^{\infty} P(X_n = 0_N) = \infty$ (Chung, 1967, p.23, Theorem 4). One possibility for process X to get from state 0_N at time 0 to state 0_N at time $2n+1$, is that $X_{2n}(1) = 0$, $Y_{2n+1}(i) = 1$ and $Z_{2n+1}(i) = 0$ for all $i \in \{1, ..., N\}$. Therefore,

$$P(X_{2n+1} = 0_N) \geq P(X_{2n}(1) = 0) \left(\frac{1}{N}\right)^N (1-\mu)^N \ .$$

Now choose J according to Prop. 2(i). Then

$$\sum_{n=1}^{\infty} P(X_n = 0_N) \geq \sum_{n=1}^{\infty} P(X_{2n}(J_{2n}) = 0) \left(\frac{1}{N}\right)^N (1-\mu)^N = \infty$$

since the random walk of Prop. 2(ii) is known to be recurrent.

Let $\tau := \inf\{n \in \mathbb{N} : X_n = 0_N\}$. For null recurrence, it remains to be shown that $\mathbb{E}(\tau) = \infty$. This follows from $\tau \geq \inf\{n \in \mathbb{N} : X_n(J_n) = 0\}$ and from the fact that the random walk of Prop. 2(ii) is null recurrent. □

4 Local Behaviour: the Normalised Allele Process V

Since no asymptotic distribution exists for the allele process X, we will now consider the normalised allele process V, corresponding to the differences between the repeat numbers of each allele and the allele of the N-th individual in each generation. Note that because of the exchangeability, any other individual may take the place of the N-th individual.

Definition 6 *The process* $V := (V_n)_{n\in\mathbb{N}_0}$, *defined by*

$$V_n : \Omega \to \mathbb{Z}^{N-1} \qquad \textit{with} \qquad V_n(i) := X_n(i) - X_n(N) \ ,$$

is called the normalised allele process.

4.1 Marginal properties of V

In this subsection several marginal properties of V are derived. The proofs are given in the appendix.

The first and second moments of the marginal distribution of V can be calculated directly from the corresponding moments of X (see appendix). Because of the exchangeability property, the distribution of $V_n(i)$ is symmetric around zero.

Lemma 7 *For any* $i, j \in \{1, ..., N-1\}, i \neq j, n \in \mathbb{N}$

(i) $\mathbb{E}\left(V_n(i)\right) = 0,$

(ii) $\mathrm{Var}\left(V_n(i)\right) = 2\mu N\left(1 - \left(1 - \frac{1}{N}\right)^n\right),$

(iii) $\mathrm{Cov}\left(V_n(i), V_n(j)\right) = \frac{1}{2}\mathrm{Var}\left(V_n(i)\right).$

Note that, in contrast to the behaviour of process X (see Lemma 3), $\lim_{n\to\infty} \mathrm{Var}\left(V_n(i)\right) = 2\mu N$ is finite.

We now derive a recursion for the marginal distribution of V. Note that, because of the exchangeability property of Proposition 1(ii), the distribution of $V_n(i)$ is independent of i for $i \leq N-1$. Thus, define

$$\eta_n(z) := \mathbb{P}(V_n(i) = z) \quad \text{for all } n \in \mathbb{N}_0,\ z \in \mathbb{Z}\ . \tag{3}$$

$$\text{For } k \in \mathbb{Z} \text{ let} \quad r(k) := \mathbb{P}\left(Z_n(1) - Z_n(2) = k\right)\ . \tag{4}$$

Obviously r does not depend on n, $r(0) = 1 - 2\mu + \frac{3}{2}\mu^2$, $r(1) = r(-1) = \mu - \mu^2$, $r(2) = r(-2) = \frac{1}{4}\mu^2$ and $r(k) = 0$ for any other k.

Lemma 8 *For any* $n \in \mathbb{N}, z \in \mathbb{Z}$

$$\eta_n(z) = \frac{N-1}{N} \sum_{k=-2}^{2} r(k)\ \eta_{n-1}(z-k) + \frac{1}{N}\ r(z)\ .$$

The next lemma provides a recursion for the higher moments and allows determination of the exponential moments of $V_n(i)$. For $\lambda > 0$ and $i \leq N-1$, define $c(\lambda) := \mathbb{E}\left(\exp\left(\lambda\left(Z_n(i) - Z_n(N)\right)\right)\right)$. Then

$c(\lambda) = r(0) + \left(\exp(\lambda) + \exp(-\lambda)\right) r(1) + \left(\exp(2\lambda) + \exp(-2\lambda)\right) r(2).$

Lemma 9 *Let $i \in \{1, ..., N-1\}$, $n, m \in \mathbb{N}$, $\lambda > 0$.*

(i) All moments of $V_n(i)$ are finite and emerge from the following recursion:

$$\mathbb{E}\,(V_n(i))^m = 0 \text{ for odd } m.$$

$$\mathbb{E}\,(V_n(i))^m = \frac{1}{N}\left(2\mu + \mu^2\left(2^{m-1} - 2\right)\right)$$

$$+\left(1 - \frac{1}{N}\right)\sum_{\substack{k=0 \\ k \text{ even}}}^{m}\binom{m}{k}\left(2\mu + \mu^2\left(2^{m-k-1} - 2\right)\right)\mathbb{E}\left(V_{n-1}(i)^k\right) \text{ for even } m.$$

(ii) All exponential moments of $V_n(i)$ are finite and are given by

$$\begin{aligned}\mathbb{E}\exp\left(\lambda V_n(i)\right) &= \frac{1}{N}\sum_{k=0}^{n-1}\left(1 - \frac{1}{N}\right)^k c(\lambda)^{k+1} + \left(1 - \frac{1}{N}\right)^n c(\lambda)^n \\ &= \frac{c(\lambda)}{N}\,\frac{1 - \left(1 - \frac{1}{N}\right)^n c(\lambda)^n}{1 - \left(1 - \frac{1}{N}\right)c(\lambda)} + \left(1 - \frac{1}{N}\right)^n c(\lambda)^n\,.\end{aligned}$$

The following corollary is straightforward and reveals the behaviour of the moments of $V_n(i)$ for $n \to \infty$.

Corollary 10 *Let $i \in \{1, ..., N-1\}$, $m \in \mathbb{N}$, $\lambda > 0$.*

(i) $\lim_{n\to\infty} \mathbb{E}\,(V_n(i))^m < \infty$.

(ii)
$$\lim_{n\to\infty} \mathbb{E}\exp\left(\lambda V_n(i)\right) = \begin{cases} < \infty \quad \text{if } \left(1 - \frac{1}{N}\right)c(\lambda) < 1 \\ \\ = \infty \quad \text{if } \left(1 - \frac{1}{N}\right)c(\lambda) \geq 1 \end{cases}.$$

4.2 Characterisation of V as a Markov chain

Like the original allele process X, the normalised process V is a Markov chain. Contrary to X, however, V can be shown to be positive recurrent (see below; for the definition of positive recurrent see Breiman (1992), p. 140). Therefore, there is an invariant distribution that characterises the asymptotic behaviour of V, and V can even be shown to converge to this distribution exponentially fast. It should be pointed out that, whereas our Markov chain characterisation of the normalised allele process V is new, the convergence result was already obtained by Kingman, using characteristic functions (Kingman, 1976).

Theorem 11

(i) V is an irreducible, aperiodic Markov chain on $\mathbb{Z}^{N-1}$, with respect to $(\mathcal{A}_n)_{n\in\mathbb{N}_0}$.

(ii) V is positive recurrent.

(iii) V converges exponentially fast to the unique invariant distribution.

Proof. Using Eq. (2), section (i) follows from the fact that

$$\begin{aligned} V_n(i) &= X_n(i) - X_n(N) = X_{n-1}(Y_n(i)) + Z_n(i) - X_{n-1}(Y_n(N)) - Z_n(N) \\ &= V_{n-1}(Y_n(i)) - V_{n-1}(Y_n(N)) + Z_n(i) - Z_n(N)\ . \end{aligned}$$

For the proof of (ii) and (iii), write 0_{N-1} for $(0,...,0) \in \mathbb{Z}^{N-1}$ and note that, for every $z \in \mathbb{Z}^{N-1}$,

$$\begin{aligned} &\mathbb{P}\left(V_n = 0_{N-1} \,|V_{n-1} = z\right) \geq \\ &\quad \mathbb{P}\left(\forall i,j \in \{1,...,N\} :\ Y_n(i) = Y_n(j), Z_n(i) = Z_n(j)\right) > 0\ . \end{aligned}$$

Thus, process V fulfills the Doeblin condition and sections (ii) and (iii) follow (see, e.g. Doob (1953), pp. 192 ff., case b). □

4.3 Unimodality of the asymptotic marginal distribution of V

Theorem 11 implies that the distribution η_n of $V_n(i)$ (see Eq. (3)) converges in distribution as $n \to \infty$. Let $\eta = \lim_{n\to\infty} \eta_n$. We will now show that η is a unimodal discrete distribution, which is one of our main novel results.

Following Keilson and Gerber (1971), we call a distribution p on $\mathbb{Z}$ *unimodal*, if at least one $M \in \mathbb{Z}$ exists such that

$$p(z) \geq p(z-1) \qquad \text{for all } z \leq M$$
$$p(z+1) \leq p(z) \qquad \text{for all } z \geq M\ .$$

For proving the unimodality of η we need the following preparatory lemma, the proof of which can be found in the appendix.

Lemma 12 *Let $\mathbb{R}^+$ denote the set of strictly positive real numbers and $\mathbb{R}_0^+$ the set of positive real numbers including zero. If*

$$M := \{\nu : \mathbb{Z} \to \mathbb{R}_0^+ \quad |\exists\, n \in \mathbb{N}; a_1, \ldots, a_n \in \mathbb{R}^+; b_1, \ldots, b_n \in \mathbb{N}_0 :$$
$$\nu = \sum_{i=1}^{n} a_i \cdot \mathbf{1}_{\{-b_i, -b_i+1, \ldots, b_i\}}\},$$

then M is closed under convolution.

With this, we can show that η is unimodal. The critical assumption of the following theorem, namely that $\mu \leq 0.8$, can safely be assumed for microsatellites.

Theorem 13

(i) If $r : \mathbb{Z} \to \mathbb{R}$ is defined as in Eq. (4), then $\eta_1 = r$ and for all $n \in \mathbb{N}\backslash\{1\}$

$$\eta_n = \frac{1}{N}\, r * \left(\sum_{i=0}^{n-2} \left(\frac{N-1}{N} \right)^i r^i \right) + \left(\frac{N-1}{N} \right)^{n-1} r^n, \tag{5}$$

where $*$ denotes the convolution of two functions and r^i the ith convolution of r.

(ii) If $\mu \leq 0.8$, then η is unimodal and symmetric around zero.

Proof. Recalling that $X_0 \equiv 0_N$, it follows that, for all $z \in \mathbb{Z}$,

$$\eta_1(z) = \mathbb{P}(V_1(1) = z) = \mathbb{P}(X_1(1) - X_1(N) = z) = \mathbb{P}(Z_1(1) - Z_1(N) = z) = r(z),$$

according to the definition of r, see Eq. (4). Since by definition $\sum_{k\in\mathbb{Z}} r(k)\eta_{n-1}(z-k) = (r * \eta_{n-1})(z)$, we can reformulate the recursive equation in Lemma 8 as follows:

$$\eta_n = \frac{N-1}{N} r * \eta_{n-1} + \frac{1}{N} r. \tag{6}$$

We will now prove Eq. (5) by induction. For $n = 2$, Eq. (5) follows from Eq.

(6). Now, let $n \in \mathbb{N}\backslash\{1,2\}$. Assuming that Eq. (5) holds for $n-1$, we have

$$\begin{aligned}
\eta_n &= \frac{N-1}{N} r * \eta_{n-1} + \frac{1}{N} r \\
&= \frac{1}{N} r + \frac{N-1}{N} r * \left(\frac{1}{N} r * \sum_{i=0}^{n-3} \left(\frac{N-1}{N} \right)^i r^i + \left(\frac{N-1}{N} \right)^{n-2} r^{n-1} \right) \\
&= \frac{1}{N} r * \left(1 + \sum_{i=0}^{n-3} \left(\frac{N-1}{N} \right)^{i+1} r^{i+1} \right) + \left(\frac{N-1}{N} \right)^{n-1} r^n \\
&= \frac{1}{N} r * \sum_{i=0}^{n-2} \left(\frac{N-1}{N} \right)^i r^i + \left(\frac{N-1}{N} \right)^{n-1} r^n.
\end{aligned}$$

To prove section (ii), we will first show that η_n is unimodal for all $n \in \mathbb{N}$. Since $0 < \mu \leq 0.8$, the following inequalities hold:

$$1 - 2\mu + \frac{3}{2}\mu^2 > \mu - \mu^2 \geq \frac{1}{4}\mu^2.$$

Thus, in the notation of Lemma 12, $r \in M$ with $n = 3$, $b_1 = 0$, $b_2 = 1$, $b_3 = 2$, $a_1 = 1 - 2\mu + \frac{3}{2}\mu^2 - (\mu - \mu^2)$, $a_2 = -\frac{1}{4}\mu^2 + (\mu - \mu^2)$, $a_3 = \frac{1}{4}\mu^2$. Now, considering Eq. (5), Lemma 12 implies that $\eta_n \in M$ for all $n \in \mathbb{N}$, and all elements of M are clearly unimodal. Unimodality of η follows from the fact that the limit of a convergent sequence of unimodal discrete distributions is itself unimodal (see "Statement 4" in Keilson and Gerber (1971)). □

4.4 Simulation of the asymptotic marginal distribution η

Lemma 8 can be used to simulate the marginal distribution of $V_n(i)$. From Theorem 11, we know that $V_n(i)$ converges in distribution as $n \to \infty$. Figs. 1 and 2 show the behaviour in time of the distribution of $V_n(i)$, assuming $\mu = 0.01$ and either $N = 100$ or $N = 1000$, respectively. For $N = 100$, the

distribution of $V_n(i)$ is close to the stationary distribution at $n = 100$, and the domain is mainly concentrated in the interval $[-7, 7]$. For $N = 1000$, convergence is slower and the domain is larger. The distribution of $V_n(i)$ is close to the stationary distribution at $n = 1000$, and the domain is mainly concentrated in $[-13, 13]$.

5 Discussion

We have shown that the allele process of the stepwise mutation model is characterised by two different types of behaviour. The expectation of the absolute value of the repeat number of a given individual converges to infinity. This signifies the global behaviour, where no convergence occurs. However, when the allelic state of an individual is chosen as a reference point for the other individuals of the population, then a limiting invariant distribution of the resulting allele difference process emerges. This is the local behaviour of the allele process, which implies that the alleles stay "clumped together" during convergence to infinity. These results confirm Moran's notion of the term "wandering distributions" (Moran, 1975).

The convergence of the allelic differences is exponentially fast, as was already noted by Kingman (Kingman, 1976). This is reassuring because it means that estimates or test statistics obtained from the allele differences not only approach a limiting distribution, but do so very quickly. As we showed, the resulting limiting marginal distribution is unimodal.

It should be noted that the SMM is a very simple model of microsatellite mutation. In some cases, it would be reasonable to assume not only mutations that change the repeat number by one unit but to allow a wider range of mutations (Huang et al., 2002). Kingman also considered generalised forms of mutations (Kingman, 1976). As long as the individual mutation events $Z_n(i)$ remain independent, which is biologically plausible, central Theorem 11 of this paper will hold true. If the random walk corresponding to the mutation process Z is null recurrent, Theorem 5 will apply. Another limitation of the SMM is the unboundedness of the state space whereas, in reality, negative repeat numbers cannot occur. Also, very large repeat numbers can result in physically unstable microsatellites and stop the evolutionary process at certain thresholds. One way to account for these limitations would be to restrict the state space of the allele process X by reflecting boundaries. The result would be a Markov chain with finite state space, and convergence to an invariant distribution would follow even for the non-normalised process X. However, differences between the normalised and non-normalised behaviour of the allele process remain possible, for instance, in the form of different convergence rates or different shapes of the invariant distribution. Because of the Markov structure and the assumed one-unit-up-or-down mutations, the process would only "realise" the existence of boundaries when it would be very close to them. Most of the time, the process would stay away from the boundaries and behave according to the stationary distribution of the normalised process V, as if no boundaries would exist.

Regarding the simplicity of the SMM, our results are only a first step towards a better understanding of the real-life situation, and investigations of how the allele process behaves under more realistic models incorporating, for example, variable mutation rates or migration, are warranted.

Acknowledgments

This work was partly funded by the German Ministry of Science and Education (BMBF) through an NGFN SMP-GEM grant to Michael Krawczak (01GS0426).

Appendix

In order to calculate the respective covariances of Lemma 4 from Eq. (2), we need the following little lemma that can be proven by induction.

Lemma 14 *Let $a, b \in \mathbb{R}$, $b \neq 1$. Then, for the real valued sequence $(x_n)_{n\in\mathbb{N}_0}$ defined by $x_n = (n-1)a + bx_{n-1}$ and $x_0 = 0$,*

$$x_n = a\,\frac{b^n + n(1-b) - 1}{(1-b)^2}\,.$$

Proof of Lemma 4

Note that the exchangeability property implies that $\mathbb{P}(X_n(i) = y, X_n(j) = z)$ is independent of $i, j \in \{1, ..., N\}$ as long as $i \neq j$. We can calculate a recursion for the covariances using Eq. (2), Prop. 1(i), Lemma 3(i), (ii) and the independence property (1):

$$\begin{aligned}
\mathrm{Cov}\left(X_n(i), X_n(j)\right) &= \mathrm{Cov}\left(X_{n-1}(Y_n(i)), X_{n-1}(Y_n(j))\right) \\
&= \sum_{k,l=1}^{N} \mathrm{Cov}\left(X_{n-1}(k), X_{n-1}(l)\right) \mathbb{P}(Y_n(i) = k, Y_n(j) = l) \\
&= \frac{1}{N^2} \sum_{k,l=1}^{N} \sum_{y,z\in\mathbb{Z}} yz\ \mathbb{P}(X_{n-1}(k) = y, X_{n-1}(l) = z) \\
&= \frac{1}{N^2} \sum_{k=1}^{N} \sum_{y\in\mathbb{Z}} y^2\ \mathbb{P}(X_{n-1}(k) = y) \\
&\quad + \frac{1}{N^2} \sum_{\substack{k,l=1 \\ k\neq l}}^{N} \sum_{y,z\in\mathbb{Z}} yz\ \mathbb{P}(X_{n-1}(k) = y, X_{n-1}(l) = z) \\
&= \frac{1}{N}\mathrm{Var}(X_{n-1}(1)) + \frac{N-1}{N}\mathrm{Cov}(X_{n-1}(i), X_{n-1}(j)) \\
&= \frac{1}{N}\mu(n-1) + \frac{N-1}{N}\mathrm{Cov}(X_{n-1}(i), X_{n-1}(j))\,.
\end{aligned}$$

Therefore, from Lemma 14,

$$\begin{aligned}\mathrm{Cov}\left(X_n(i), X_n(j)\right) &= \frac{\mu}{N}\,\frac{\left(\frac{N-1}{N}\right)^n + n(1-\frac{N-1}{N}) - 1}{(1-\frac{N-1}{N})^2} \\ &= \mu\left(n + \frac{(N-1)^n - N^n}{N^{n-1}}\right). \quad \square\end{aligned}$$

Proof of Lemma 7

(i) follows directly from the definition of V. Variance and covariance can be derived using Lemmata 3 and 4:

$$\begin{aligned}\mathrm{Var}\left(V_n(i)\right) &= \mathrm{Var}\left(X_n(i)\right) + \mathrm{Var}\left(X_n(N)\right) - 2\mathrm{Cov}\left(X_n(i), X_n(N)\right) \\ &= 2\mu n - 2\mu\left(n + \frac{(N-1)^n - N^n}{N^{n-1}}\right) \\ \mathrm{Cov}\left(V_n(i), V_n(j)\right) &= \mathrm{Var}\left(X_n(N)\right) - \mathrm{Cov}\left(X_n(N), X_n(j)\right) \\ &\quad - \mathrm{Cov}\left(X_n(i), X_n(N)\right) + \mathrm{Cov}\left(X_n(i), X_n(j)\right) \\ &= \mathrm{Var}\left(X_n(N)\right) - \mathrm{Cov}\left(X_n(N), X_n(j)\right) \quad \square\end{aligned}$$

Proof of Lemma 8

Using recursion (2), it follows that

$$\begin{aligned}\eta_n(z) &= \mathbb{P}\left(X_n(1) - X_n(N) = z\right) = \mathbb{P}\left(X_n(1) - X_n(2) = z\right) \\ &= \mathbb{P}\left(X_{n-1}(Y_n(1)) + Z_n(1) - \left(X_{n-1}(Y_n(2)) + Z_n(2)\right) = z\right) \\ &= \mathbb{P}\left(Y_n(1) \neq Y_n(2)\right) \\ &\quad \cdot \mathbb{P}\left(X_{n-1}(Y_n(1)) + Z_n(1) - X_{n-1}(Y_n(2)) - Z_n(2) = z \,|Y_n(1) \neq Y_n(2)\right) \\ &\quad + \mathbb{P}\left(Y_n(1) = Y_n(2)\right) \cdot \mathbb{P}\left(Z_n(1) - Z_n(2) = z \,|Y_n(1) = Y_n(2)\right)\end{aligned}$$

$$
\begin{aligned}
&= \frac{N-1}{N} \sum_{k\in\mathbb{Z}} \mathbb{P}\left(Z_n(1) - Z_n(2) = k\right) \\
&\quad \cdot \mathbb{P}\Big(X_{n-1}(Y_n(1)) + Z_n(1) - X_{n-1}(Y_n(2)) - Z_n(2) = z \\
&\qquad\qquad \Big| Y_n(1) \neq Y_n(2), Z_n(1) - Z_n(2) = k\Big) \\
&\quad + \frac{1}{N}\, \mathbb{P}\left(Z_n(1) - Z_n(2) = z\right) \\
&= \frac{N-1}{N} \sum_{k\in\mathbb{Z}} r(k) \mathbb{P}\left(X_{n-1}(1) - X_{n-1}(2) = z - k\right) + \frac{1}{N}\, r(z) \\
&= \frac{N-1}{N} \sum_{k\in\mathbb{Z}} r(k) \eta_{n-1}(z-k) + \frac{1}{N}\, r(z)\,. \quad \square
\end{aligned}
$$

Proof of Lemma 9

(i): From Eq. (2), we obtain

$$
\begin{aligned}
\mathbb{E}\,(V_n(i))^m &= \mathbb{P}\left(Y_n(i) = Y_n(N)\right) \mathbb{E}\,(Z_n(i) - Z_n(N))^m + \sum_{\substack{k,l=1 \\ k\neq l}}^{N} \mathbb{P}\left(Y_n(i) = k, Y_n(N) = l\right) \\
&\quad \cdot \mathbb{E}\,(X_{n-1}(k) + Z_n(i) - X_{n-1}(l) - Z_n(N))^m \\
&= \frac{1}{N} \mathbb{E}\,((Z_n(i) - Z_n(N))^m) + \left(1 - \frac{1}{N}\right) \mathbb{E}\,((V_{n-1}(i) + Z_n(i) - Z_n(N))^m) \\
&= \frac{1}{N} \mathbb{E}\,((Z_n(i) - Z_n(N))^m) \\
&\quad + \left(1 - \frac{1}{N}\right) \sum_{k=0}^{m} \binom{m}{k} \mathbb{E}\left((Z_n(i) - Z_n(N))^{m-k}\right) \mathbb{E}\left(V_{n-1}(i)^k\right)\,.
\end{aligned}
$$

Using $\mathbb{E}\,((Z_n(i) - Z_n(N))^m) = 0$ for m odd and $\mathbb{E}\,((Z_n(i) - Z_n(N))^m) = 2\mu + \mu^2\,(2^{m-1} - 2)$ for m even, section (i) follows by induction and Lemma 7.

(ii): Treating the exponential moments in the same way yields

$$\begin{aligned} d_n &:= \mathbb{E}\left(\exp\left(\lambda V_n(i)\right)\right) \\ &= \frac{1}{N}\,\mathbb{E}\left(\exp\left(\lambda\left(Z_n(i) - Z_n(N)\right)\right)\right) \\ &\quad + \left(1 - \frac{1}{N}\right)\,\mathbb{E}\left(\exp\left(\lambda V_{n-1}\right)\right)\,\mathbb{E}\left(\exp\left(\lambda\left(Z_n(i) - Z_n(N)\right)\right)\right) \\ &= \frac{1}{N}c(\lambda) + \left(1 - \frac{1}{N}\right)c(\lambda)\,d_{n-1}\,. \quad \square \end{aligned}$$

Proof of Lemma 12

Let $*$ denote the convolution of two functions and define $f_b := \mathbf{1}_{\{-b,-b+1,\ldots,b\}}$ for $b \in \mathbb{N}_0$. First note that, for all $b \leq b' \in \mathbb{N}_0$ and for all $z \in \mathbb{Z}$, the following equation holds:

$$(f_b * f_{b'})(z) = \sum_{i\in\mathbb{Z}} f_b(i)\cdot f_{b'}(z-i) = \sum_{i=-b}^{b} f_{b'}(z-i) = \sum_{i=-b}^{b} \mathbf{1}_{\{-b'+i,\ldots,b'+i\}}(z).$$

Dropping argument z and decomposing the sum on the right-hand side,

$$\begin{aligned} f_b * f_{b'} &= f_{b'} + \sum_{i=1}^{b} \mathbf{1}_{\{-b'+i,\ldots,b'+i\}} + \sum_{i=-b}^{-1} \mathbf{1}_{\{-b'+i,\ldots,b'+i\}} \\ &= f_{b'} + \sum_{i=1}^{b} (\mathbf{1}_{\{-b'+i,\ldots,b'+i\}} + \mathbf{1}_{\{-b'-i,\ldots,b'-i\}}) \\ &= f_{b'} + \sum_{i=1}^{b} (f_{b'+i} + f_{b'-i}). \end{aligned}$$

Now let $\nu, \nu' \in M$, which can be written as $\nu = \sum_{i=1}^{n} a_i \cdot f_{b_i}$ and $\nu' = \sum_{i=1}^{n'} a'_i \cdot f_{b'_i}$. Let $m(ij) := \max\{b_i, b'_j\}$. Taking into account the distributivity and linearity

of the convolution, the above implies that

$$\begin{aligned}\nu * \nu' &= \left(\sum_{i=1}^{n} a_i \cdot f_{b_i}\right) * \left(\sum_{i=1}^{n'} a'_i \cdot f_{b'_i}\right) = \sum_{i=1}^{n}\sum_{j=1}^{n'} a_i a'_j \cdot (f_{b_i} * f_{b'_j}) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n'} a_i a'_j \cdot \left(f_{m(ij)} + \sum_{k=1}^{\min\{b'_j, b_i\}} \left(f_{m(ij)+k} + f_{m(ij)-k}\right)\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n'} a_i a'_j \cdot f_{m(ij)} + \sum_{i=1}^{n}\sum_{j=1}^{n'} \sum_{k=1}^{\min\{b'_j, b_i\}} a_i a'_j \cdot f_{m(ij)+k} \\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{n'} \sum_{k=1}^{\min\{b'_j, b_i\}} a_i a'_j \cdot f_{m(ij)-k} \ .\end{aligned}$$

Since all the characteristic functions in the last expression are symmetrical around zero, it follows that $\nu * \nu' \in M$. $\square$

References

Ashley, C.T., Warren, S.T., 1995. Trinucleotide repeat expansion and human disease. Ann. Rev. of Genet. 29, 703-728.

Bindu, G.H., Trivedi, R., Kashyap, V.K., 2007. Allele frequency distribution based on 17 STR markers in three major Dravidian linguistic populations of Andhra Pradesh, India. Forensic Sci. Int. 170, 76-85.

Breiman, L., 1992. Probability. SIAM, Philadelphia.

Calabrese, P.P., Durrett, R.T., Aquadro, C.F., 2001. Dynamics of microsatellite divergence and proportional slippage/point mutation models. Genetics 159, 839-852.

Calabrese, P.P., Sainudiin, R., 2005. Models of Microsatellite Evolution. In: Nielsen, R. (Ed.), Statistical Methods in Molecular Evolution. Springer, London, pp. 289-306.

Cassidy, B.G., Gonzales, R.A., 2005. DNA testing in animal forensics. J. Wildl. Manage. 69, 1454-1462.

Chambers, G.K., MacAvoy, E.S., 2000. Microsatellites: consensus and controversy. Comp. Biochem. Physiol. B 126, 455-476.

Chung, K.L., 1967. Markov Chains with Stationary Transition Probabilities. 2nd ed. Springer, Berlin.

Cornuet, J.M., Beaumont, M.A., Estoup, A., Solignac, M., 2006. Inference on microsatellite mutation processes in the invasive mite, Varroa destructor, using reversible jump Markov chain Monte Carlo. Theor. Popul. Biol. 69,

129-144.

De Iorio, M., Griffiths, R.C., Leblois, R., Rousset, F., 2005. Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. Theor. Popul. Biol. 68, 41-53.

Di Rienzo, A., Peterson, A.C., Garza, J.C., Valdes, A.M., Slatkin,M., Freimer, N.B., 1994. Mutational processes of simple-sequence repeat loci in human populations. Proc. Natl. Acad. Sci. USA 91, 3166-3170.

Doob, J.L., 1953. Stochastic Processes. John Wiley, New York, reprinted Wiley Classics Library Edition 1990.

Durrett, R., Kruglyak, S., 1999. A new stochastic model of microsatellite evolution. J. Appl. Probab. 36, 621-631.

Falush, D., Iwasa, Y., 1999. Size-dependent mutability and microsatellite constraints. Mol. Biol. Evol. 16, 960-966.

Feldman, M.W., Bergman, A., Pollock, D.D., Goldstein, D.B., 1997. Microsatellite genetic distances with range constraints: analytic description and problems of estimation. Genetics 145, 207-216.

Garza, J.C., Slatkin, M., Freimer, N.B., 1995. Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. Mol. Biol. Evol. 12, 594-603.

Goldstein, D.B., Roemer, G.W., Smith, D.A., Reich, D.E., Bergman, A., Wayne, R.K., 1999. The use of microsatellite variation to infer population structure and demographic history in a natural model system. Genetics 151, 797-801.

Huang, Q.-Y., Xu, F.-H., Shen, H., Deng, H.-Y., Liu, Y.-J., Liu, Y.-Z., Li, J.-L., Recker, R.R., Deng, H.-W., 2002. Mutation patterns at dinucleotide microsatellite loci in humans. Am. J. Hum. Genet. 70, 625-634.

Kashi, Y., King, D.G., 2006. Simple sequence repeats as advantageous mutators in evolution. Trends Genet. 22, 253-259.

Keilson, J., Gerber, H., 1971. Some results for discrete unimodality. JASA 66, 386-389.

Kingman, J.F.C., 1976. Coherent random walks arising in some genetical models. Proc. R. Soc. Lond. A 351, 19-31.

Kruglyak, S., Durrett, R.T., Schug, M.D., Aquadro, C.F., 1998. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc. Natl. Acad. Sci. USA 95, 10774-10778.

Moran, P.A.P., 1975. Wandering distributions and the electrophoretic profile. Theor. Popul. Biol. 8, 318-330.

Li, Y-C., Korol, A.B., Tzion, F., Beiles, A., Nevo, E., 2002. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review, Mol. Ecol. 11, 2453-2465.

Ohta, T., Kimura, M., 1973. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. 22, 201-204.

Sainudiin, R., Durrett, R.T., Aquadro, C.F., Nielsen, R., 2004. Microsatellite mutation models: insights from a comparison of humans and chimpanzees.

Genetics 168, 383-395.

Tautz, D., 1993. Notes on the definition and nomenclature of tandemly repetive DNA sequences. In: Pena, S.D.J., Chakraborty, R., Epplen, J.T., Jeffreys, A.J. (Eds.), DNA Fingerprinting: State of the Science. Birkhäuser Verlag, Basel, pp. 21-28.

Thibodeau, S.N., Bren, G., Schaid, D., 1993. Microsatellite instability in cancer of the proximal colon. Science 260, 816-819.

Tishkoff, S.A., Dietzsch, E., Speed, W., Pakstis, A.J., Kidd, J.R., Cheung, K., Bonné-Tamir, B., Santachiara-Benerecetti, A.S., Moral, P., Krings, M., Pääbo, S., Watson, E., Risch, N., Jenkins, T., Kidd, K.K., 1996. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271, 1380-1387.

Vardo, A.M., Schall, J.J., 2007. Clonal diversity of a lizard malaria parasite, Plasmodium mexicanum, in its vertebrate host, the western fence lizard: role of variation in transmission intensity over time and space. Mol. Ecol. 16, 2712-2720.

Watkins, J.C., 2007. Microsatellite evolution: Markov transition functions for a suite of models. Theor. Popul. Biol. 71, 147-159.

Weissenbach, J., Gyapay, G., Dib, C., Vignal, A., Morissette, J., Millasseau, P., Vaysseix, G., Lathrop, M., 1992. A second-generation linkage map of the human genome. Nature 359, 794-801.

Whittaker, J.C., Harbord, R.M., Boxall, N., Mackay, I., Dawson, G., Sibly, R.M., 2003. Likelihood-based estimation of microsatellite mutation rates.

Genetics 164, 781-787.

Zhivotovsky, L.A., Feldman, M.W., Grishechkin, S.A., 1997. Biased mutations and microsatellite variation. Mol. Biol. Evol. 14, 926-933.

Zhivotovsky, L.A., Rosenberg, N.A., Feldman, M.W., 2003. Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. Am. J. Hum. Genet. 72, 1171-1186.

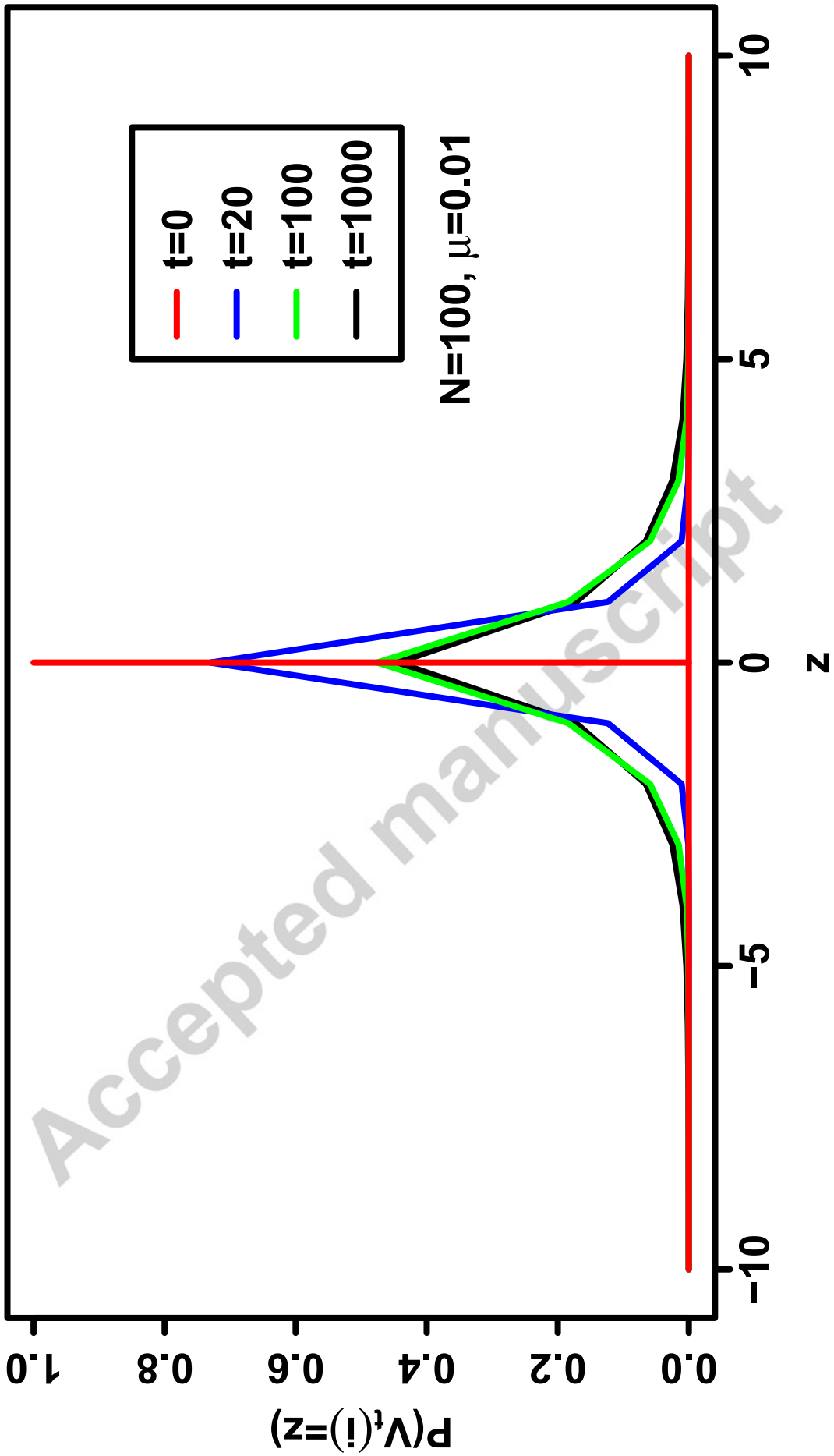
Figure legends

Figure 1: Convergence of the marginal distribution of the normalised allele process V. For illustration, the discrete probabilities $\mathbb{P}(V_t(i) = z)$ obtained for integer z are connected by lines.

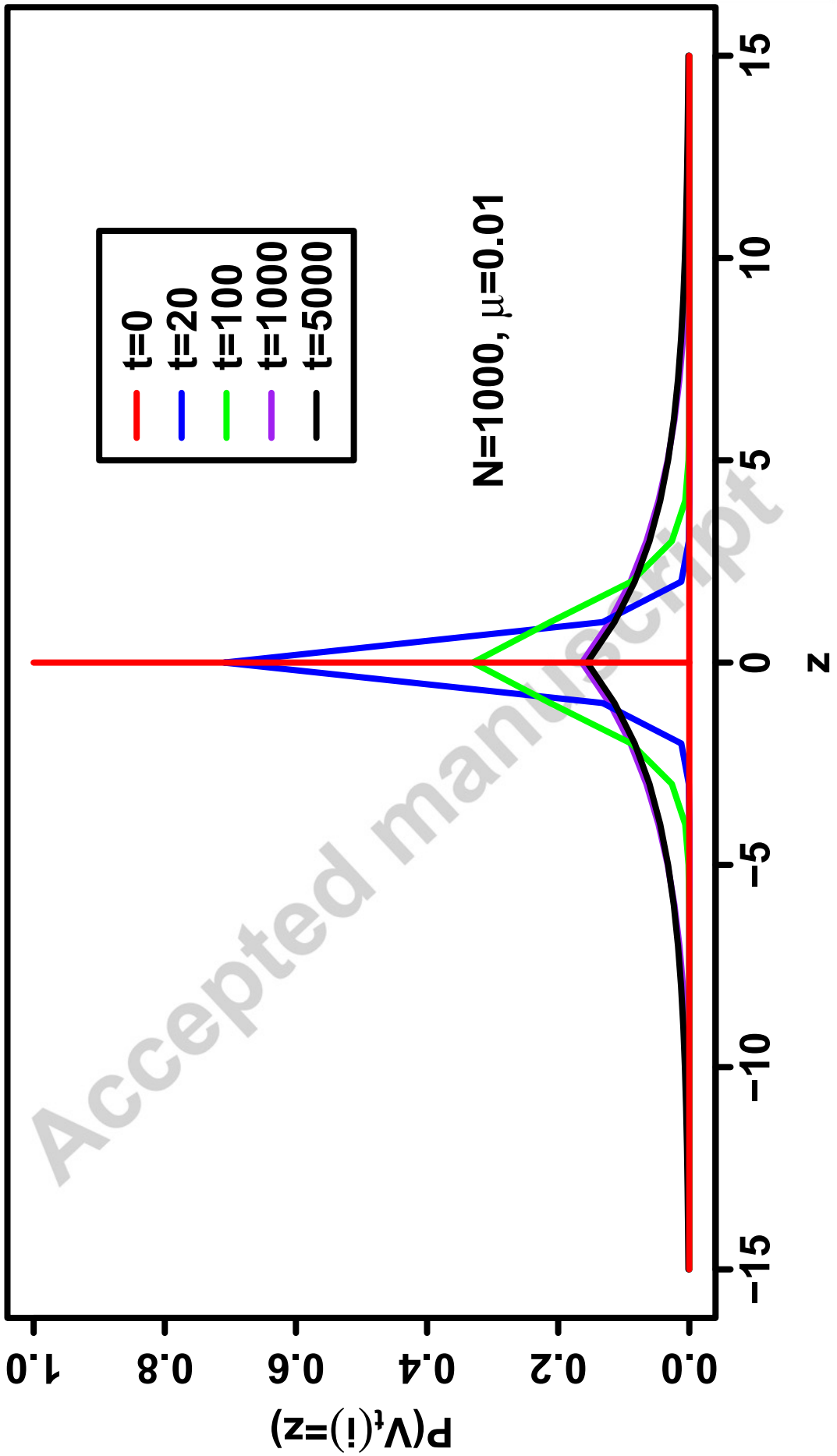
N=100, $\mu = 0.01$, t: number of generations

Figure 2: Convergence of the marginal distribution of the normalised allele process V. For illustration, the discrete probabilities $\mathbb{P}(V_t(i) = z)$ obtained for integer z are connected by lines.

N=1000, $\mu = 0.01$, t: number of generations



Caliebe et al. Fig. 1



Caliebe et al. Fig. 2