PHYSICAL REVIEW E, VOLUME 65, 040901(R)

Statistics of selectively neutral genetic variation

A. Eriksson,[1] B. Haubold,[2] and B. Mehlig[1]
[1]*Physics and Engineering Physics, Chalmers/GU, Gothenburg, Sweden*
[2]*LION Bioscience AG, Waldhofer Strasse 98, 69123 Heidelberg, Germany*
(Received 27 June 2001; revised manuscript received 29 November 2001; published 4 April 2002)

Random models of evolution are instrumental in extracting rates of microscopic evolutionary mechanisms from empirical observations on genetic variation in genome sequences. In this context it is necessary to know the statistical properties of empirical observables (such as the local homozygosity, for instance). Previous work relies on numerical results or assumes Gaussian approximations for the corresponding distributions. In this paper we give an analytical derivation of the statistical properties of the local homozygosity and other empirical observables assuming selective neutrality. We find that such distributions can be very non-Gaussian.

DOI: 10.1103/PhysRevE.65.040901 PACS number(s): 87.23.Kg, 87.10.+e, 89.75.−k

For more than thirty years, microscopic random models of genetic evolution have been the focus of a substantial research effort in theoretical biology [1–4]. In the future, such microscopic models and their statistical analysis will be of yet increasing significance in this field: the amount of accurate and comprehensive data on the genetics of viruses, bacteria, and especially the human genome [5–7] has increased so considerably that it is now possible to test microscopic models of genetic evolution.

Genetic information is encoded in the linear sequence of nucleotides in DNA molecules; the four different nucleotides occurring in DNA (adenosine, cytosine, guanine, and thymine) are usually denoted by A, C, G, and T. A sequence of a few hundred or a few thousand of these forms a gene, also referred to as a locus. Mutations change individual nucleotides (e.g., from A to C) and thus create modified versions of loci. The resulting different types of loci are also known as allelic types. Because loci consist of many nucleotides—each of which can be changed by mutation independently from the others—the number of possible allelic types is typically very large. To a good approximation it can thus be assumed that every mutation creates a new allelic type. This is the defining feature of the *infinite-alleles model* [1].

Empirically, genetic variation is recorded by measuring the frequencies $\omega_a^{(l)}$ of each allelic type a at each locus l. Genetic variation reflects the microscopic processes of evolutionary dynamics. The simplest model of evolution proceeds by sampling with replacement each generation from the previous generation (at constant population size N). In addition, a number of microscopic processes take place, each happening at a constant (but generally unknown) rate. One such process is mutation, measured as $\theta=2N\mu$ where μ is the probability of mutation per locus per generation (in a haploid population). Another such process is the exchange of genetic material between individuals of a population measured as $C=2Nc$ where c is the probability of an exchange event per locus per generation [8]. C is termed the recombination rate.

The model of genetic evolution described here is called the *constant-rate neutral mutation process*, referred to as *neutral process* in the following. It is a stochastic model and assumes that no selective forces act. The neutral process is one of the most significant microscopic models of genetic evolution: not only does it provide a model for genetic variation at loci unaffected by selection, deviations between empirical observations and predictions of this neutral process allow for a qualitative characterization of selective effects (see [9]).

There is by now an overwhelming amount of work, both theoretical and empirical, on the neutral process for the infinite-alleles model. A convenient way of simulating this process on a computer is to consider genealogies of samples of a given population [3,10,11] in the limit of $N\to\infty$. Random samples are most effectively generated by creating random genealogies. In this way, statistics of empirical observables may be obtained using Monte Carlo simulations. Another possibility is to simulate Ewen's sampling formula [2] which determines the statistics of the neutral process in the limit of large C. Analytical work has mostly focused on calculating expectation values and variances of empirical observables [12]. Distributions of even the simplest empirical observables (such as the one-locus homozygosity [2]) are not known analytically. The difficulty is that moments of empirical observables are usually calculated by expanding them into a sum of identity coefficients [12]. This procedure is impractical for high moments.

At the same time, the form of such distributions is of great interest: for example, they characterize sample-to-sample fluctuations. More importantly, they can be used to establish confidence intervals for empirical observations. To date, such confidence intervals have routinely been obtained from Monte Carlo calculations [13,14]. Alternatively, it has been assumed that the distributions are well approximated by Gaussians [15].

The aim of this paper is to calculate distributions of empirical observables (such as the homozygosity) in the neutral process for the infinite-alleles model. The remainder is organized as follows: first the results for a single locus are described, and then those for two and more loci. Finally, implications of the results are discussed.

Consider the *homozygosity* F_2, the probability that a pair of alleles (in a sample of size n with m allelic types) has the same allelic type. In terms of the allelic frequencies this probability can be expressed as (large n)

$$F_2=\sum_{a=1}^{m}\omega_a^2. \tag{1}$$

The statistics of F_2 is determined by the moments of F_2,

$$\phi_k = \langle F_2^k \rangle, \tag{2}$$

where the average is over random genealogies according to the neutral process. The ϕ_k may be calculated numerically in at least two ways: by generating random genealogies [10,3,11] or by evaluating Ewen's sampling formula [16]. Obtaining an analytical estimate of the ϕ_k is complicated by the fact that the allelic frequencies ω_a in Eq. (1) are not independently distributed. For instance, they must satisfy the constraint $\Sigma_{a=1}^{m} \omega_a = 1$.

To obtain analytical results we seek an approximate representation of the neutral process in terms of *independent* random numbers. When only one locus is of interest, nonrecombination models apply, irrespective of how much gene exchange actually occurs. In this case the numbers c_a of allelic types a with given frequency ω_a are approximately independently distributed [17], albeit only for sufficiently small ω_a. Unfortunately, this result does not yield the statistics of F_2 since *all* frequencies ω_a enter in Eq. (1), and not just the small ω_a.

In the following we show how the distribution of F_2 can be determined by means of a recursion for the frequencies ω_a. Assume that there are m allelic types with frequencies $\omega_1, \omega_2, \ldots, \omega_m$, obeying the normalization condition $\Sigma_{a=1}^{m} \omega_a = 1$. Add one allelic type; the corresponding $m+1$ frequencies ω_a' are defined as follows: draw a frequency $\omega_{m+1}' = z_m$ with density $\Phi(z_m)$. To ensure normalization, define $\omega_k' = (1-z_m)\omega_k$ for $k = 1, \ldots, m$. Thus

$$\omega_a' = z_{a-1} \prod_{b=a}^{m-1} (1-z_b) \tag{3}$$

where z_a (for $a \geqslant 1$) are independent random variables with density $\Phi(z_a)$ and $z_0 = 1$. For $\Phi(z_a) = \theta(1-z_a)^{\theta-1}$ it follows from [18,19] that (for large values of n) the frequencies ω_a are distributed according to the neutral process.

The recursive definition (3) enables us to derive an explicit expression for the moments of F_2: for large n

$$F_2 \simeq \sum_{a=1}^{m} \omega_a^2, \quad F_2' \simeq \sum_{a=1}^{m+1} \omega_a'^2 = z_m^2 + (1-z_m)^2 \sum_{a=1}^{m} \omega_a^2. \tag{4}$$

Since the sum on the right-hand side does not depend on z_m, it can be averaged independently of z_m. In the limit of large n, F_2 and F_2' have the same distribution, $F_2 \sim F_2'$. Using $\langle z^k \rangle_\Phi = \Gamma(1+k)\Gamma(1+\theta)/\Gamma(1+k+\theta)$ and $\langle (1-z)^l \rangle_\Phi = \theta/(l+\theta)$,

$$\phi_k = \theta \sum_{l=0}^{k-1} \binom{k}{l} \frac{[2(k-l)]!}{2k} \, \theta \, \frac{\Gamma(2l+\theta)}{\Gamma(2k+\theta)} \phi_l. \tag{5}$$

Here and above $\Gamma(x)$ is the Gamma function. Equation (5) provides an analytical approximation for arbitrary moments of F_2, appropriate in the limit of large sample sizes n.

One could reconstruct the distribution function $P(x) = \mathrm{Prob}(F_2 = x)$ of F_2 from the moments (5). It is, however, more convenient to derive the analog of Eq. (4) for $P(x)$ itself. By definition [see Eq. (4)],

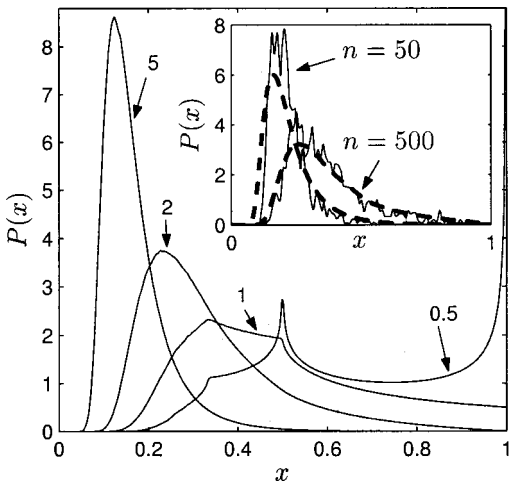

FIG. 1. $P(x) = \mathrm{Prob}(F_2 = x)$ for $L = 1$ and $\theta = 0.5, 1, 2$, and 5. Inset: analytical results for $P(x)$ compared to the Monte Carlo results of [20], for $m = 10$ and $n = 50$, $n = 500$.

$$P(x) = \int_0^1 dz \Phi(z) P[(x-z^2)/(1-z)^2] \tag{6}$$

for $0 \leqslant x \leqslant 1$ and zero otherwise. This can be rewritten as

$$P(x) = \int_0^1 dy Q(x,y) P(y) \tag{7}$$

with the kernel

$$Q(x,y) = \frac{\theta}{2a} \left[\left(\frac{1+a}{1+y} \right)^{\theta-1} + \left(\frac{1-a}{1+y} \right)^{\theta-1} H(y-x) \right] \times H\left(\frac{x}{1-x} - y \right) \tag{8}$$

where $a \equiv a(x,y) = \sqrt{x-(1-x)y}$ and $H(z)$ is the Heaviside step function. Note that $Q(x,y)$ exhibits a divergence as $x,y \to 0$. Equation (7) is solved by expanding $P(x)$ in a suitable set of basis functions on the interval $[0,1]$, resulting in an eigenvalue problem. Figure 1 shows the resulting distributions $P(x)$ for four values of θ. Clearly, the statistics of F_2 is very non-Gaussian.

The calculations summarized above are not only of interest in the case of one locus, as we show in the following (L denotes the number of loci). In the case of two loci ($L = 2$) on the same stretch of DNA, the joint distribution of allelic frequencies $\omega_a^{(l)}$ depends on the rate C of gene exchange. Consider (for large n)

$$F_2 = \frac{1}{L} \sum_{l=1}^{L} F_2^{(l)}, \quad F_2^{(l)} = \sum_a [\omega_a^{(l)}]^2. \tag{9}$$

In the limit of large C, the two genealogies for $l = 1$ and $l = 2$ are essentially independent and the frequencies $\omega_a^{(l)}$ are well approximated by Eq. (3) for each l (and large n). The

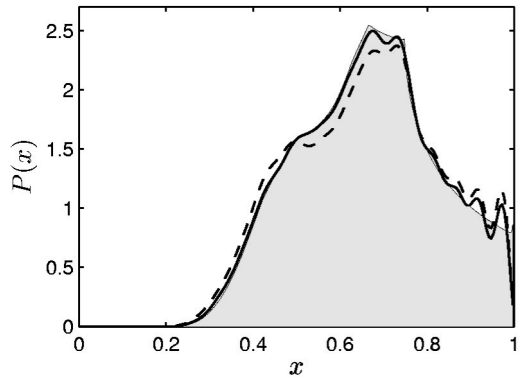

FIG. 2. $P(x)=\mathrm{Prob}(F_2=x)$ for two loci ($L=2$) and $\theta=0.5$, in the limit of large C and n (shaded). Also shown are results of Monte Carlo simulations for $n=100$ and $C=10$ (solid line) and $C=1$ (dashed line).

distribution $P(x)=\mathrm{Prob}(F_2=x)$ is thus obtained from the single-locus $P(x)$ by convolution. The resulting distribution is shown in Fig. 2. Empirically determined recombination rates are often so large that this result for $P(x)$ is a good approximation: in Fig. 2 two distributions of F_2 are shown, for $n=100$, $\theta=1/2$, and $C=1$ and 10, obtained from Monte Carlo simulations. One observes good agreement with the prediction (shaded), even for values of C as low as $C=1$. It must be emphasised that the distribution is markedly non-Gaussian. The wiggles in the Monte Carlo results are statistically significant; they are a consequence of the finite sample size ($n=100$).

When $L\gg 1$, and in the limit of large C, the distribution of F_2 [as defined in Eq. (9)] is Gaussian, and its moments are obtained as

$$\phi_k=\left[1+\binom{k}{2}\frac{2\theta}{(2+\theta)(3+\theta)}\frac{1}{L}\right](1+\theta)^{-k}. \qquad (10)$$

In an empirical data set, n (and m) are necessarily finite. It must then be asked: to what extent are the z_a independently and identically distributed for finite n (and m)? Figure 3(a) shows z_a values determined from empirical data on *C. jejuni* [21], at the locus *GltA* ($n=194$ and $m=27$), in comparison with the theory for $n=\infty$. The empirical z_a are approximately identically distributed, except at the edges where finite-size effects are observed (remember that $z_0\equiv 1$). Monte Carlo simulations for $n=194$ and $m=27$ confirm the effect of finite sample size. Figure 3(b) is a similar plot with data taken from one Monte Carlo sample. The inset of Fig. 3(b) shows that the z_a are indeed independently distributed. It can be concluded that the theory works well in the present case.

In the remainder two implications of our results are discussed. First, in practice it is necessary to decide whether empirically observed frequencies at a given locus are consistent with the neutral process. The standard statistical test (see [2], p. 263) uses the distribution of F_2 as an input (albeit with the number m of allelic types as a parameter and not θ as in the above equations). Since the distribution of F_2 was

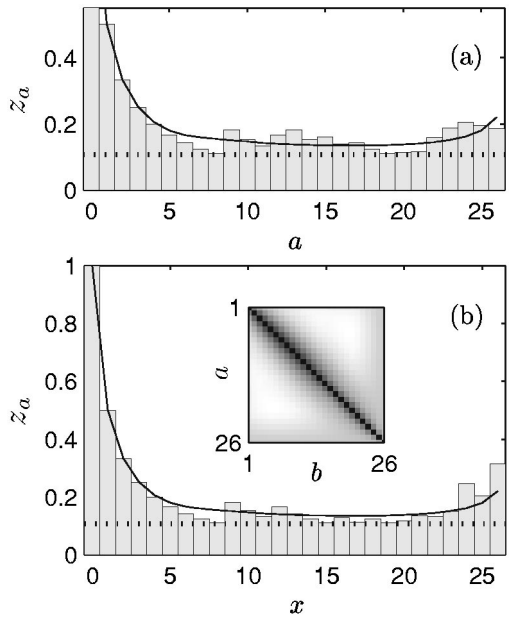

FIG. 3. (a) Frequencies z_a from empirical ω_a (locus *GltA* in *C. jejuni* [21]), compared to the neutral model for $n\to\infty$ (dashed line). Also shown are results of Monte Carlo simulations for finite $n=194$ (solid line). (b) is a similar plot with data taken from one Monte Carlo sample. The inset shows the correlation strength between z_a and z_b for $n=194$ and 27 alleles. Black corresponds to full correlation.

unknown, it was usually determined by Monte Carlo simulations. Now, however, the result (7),(8) can be used: for $m\gtrsim\log n$, Eqs. (7),(8) apply independently of whether m or θ is taken as the parameter. The corresponding distributions are compared to Monte Carlo data [20] in Fig. 1. Shown are two cases: $m=10,n=50$ and $m=10,n=500$. In both cases, the agreement between our results and those of Monte Carlo simulations is very good.

Second, many recent empirical studies (see, for instance, [13,14,22]) have analyzed the extent of gene exchange. A common measure is the variance V_D of the number of pairwise differences at all loci under consideration. In the limit

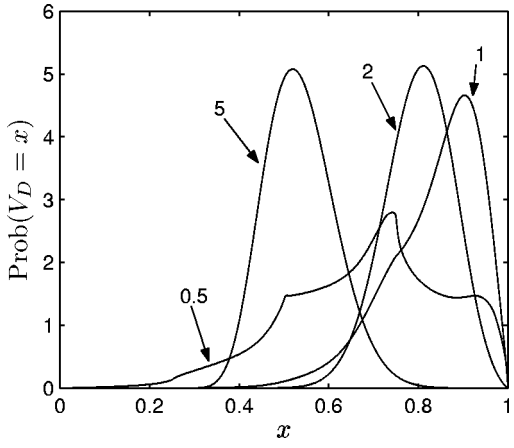

FIG. 4. Null distribution of V_D for $L=4$, and $\theta=0.5$, 1, 2, and 5 (in the limit of large C, the range of V_D is $0\leqslant V_\mathrm{D}\leqslant L/4$).

of $C\to\infty$ (linkage equilibrium) $\langle V_{\mathrm{D}}\rangle=\langle\Sigma_{l=1}^{L}(1-F_2^{(l)})F_2^{(l)}\rangle$ (for the neutral process this evaluates to $L\,\theta(4+\theta)/[(1+\theta)\times(2+\theta)(3+\theta)]$, see [12]). However, for finite values of C (linkage disequilibrium), and especially for small C, the expected value of V_{D} is larger. The empirically determined value of V_{D} can be compared to a critical value obtained under the null hypothesis that all loci are in linkage equilibrium. The corresponding null distribution is usually obtained using Monte Carlo simulations [13,14].

In cases where the neutral model applies, the null distribution of V_{D} can be determined from Eqs. (5), (7), and (8). Consider first the case of large L, where the null distribution is approximately Gaussian. Using $V_{\mathrm{D}}\sim\Sigma_{l=1}^{L}(1-F_2^{(l)})F_2^{(l)}$ for large C, one obtains

$$\mathrm{Var}[V_{\mathrm{D}}]=L[\phi_4-2\phi_3+\phi_2-(\phi_1-\phi_2)^2]=L\,\frac{2\theta(1872-420\theta-584\theta^2+229\theta^3+163\theta^4+23\theta^5+\theta^6)}{(1+\theta)^2(2+\theta)^2(3+\theta)^2(4+\theta)(5+\theta)(6+\theta)(7+\theta)}. \tag{11}$$

This variance is always larger than the corresponding quantity in a random shuffling scheme [13,14] because the latter is conditioned on the homozygosity, and not on θ. When L is small, the null distribution will be very non-Gaussian, as the above results for the distribution of F_2 show. In Fig. 4, the null distribution of V_{D} [as determined from Eqs. (7),(8)] is shown for the case of $L=4$ and for four values of θ. Note that the forms of the distributions imply large, asymmetric confidence intervals. Finally, for $m\gtrsim\ln n$, the distributions in Fig. 4 are insensitive to whether the process is conditioned on fixed θ or fixed k [23].

We have shown that distribution functions of empirical observables measuring genetic diversity in selectively neutral populations may exhibit strong non-Gaussian tails. We have found analytical approximations for these distributions, valid for large sample sizes and in the limit where gene exchange is frequent; and have discussed implications for the statistical analysis of genetic variation. It is highly desirable to extend the present results to the case where gene exchange is rare, corresponding to clonal or nearly clonal populations.

[1] M. Kimura and J. F. Crow, Genetics **49**, 725 (1964).
[2] W. J. Ewens, *Mathematical Population Genetics* (Springer, Berlin, 1979).
[3] R. R. Hudson, in *Gene Genealogies and the Coalescent Process*, Vol. 7 of *Oxford Surveys in Evolutionary Biology* (Oxford University Press, Oxford, 1990), pp. 1–44.
[4] W. J. Ewens and G. Grant, *Bioinformatics* (Springer, Berlin, 2001).
[5] http://www.mlst.net.
[6] E. S. Lander *et al.*, Nature (London) **409**, 860 (2001).
[7] J. C. Venter *et al.*, Science **291**, 1304 (2001).
[8] J. Maynard Smith, J. Evol. Biol. **7**, 525 (1994); Th. Wiehe, J. Mountain, P. Parham, and M. Slatkin, Gen. Res., Camb. **75**, 61 (2000); M. Bahlo, Theor. Popul. Biol. **56**, 265 (1999); C. Wiuf and J. Hein, Genetics **155**, 451 (2000).
[9] T. Ohta, BioEssays **18**, 673 (1996); M. Kreitman, *ibid.* **18**, 678 (1996); N. Takahata, Curr. Opin. Genet. Dev. **6**, 676 (1996).
[10] J. F. C. Kingman, Stochastic Proc. Appl. **13**, 245 (1982).
[11] S. Tavaré, *Calculating the Secrets of Life: Applications of the Mathematical Sciences to Molecular Biology* (NAP, Washington, DC, 1995).
[12] R. R. Hudson, J. Evol. Biol. **7**, 535 (1994).
[13] V. Souza *et al.*, Proc. Natl. Acad. Sci. U.S.A. **89**, 8389 (1992).
[14] B. Haubold and P. B. Rainey, Mol. Ecol. **5**, 747 (1996).
[15] A. H. D. Brown, M. W. Feldman, and E. Nevo, Genetics **96**, 523 (1980).
[16] Ewen's sampling formula [2] determines the probability of any neutral sample configuration, as well as the probabilities of these samples conditional on the number of allelic types. Numerical simulations based on this formula are described by P. A. Fuerst, R. Chakraborty, and M. Nei, Genetics **36**, 455 (1977).
[17] R. Arratia and S. Tavaré, Adv. Math. **104**, 90 (1994).
[18] G. P. Patil and C. Taillie, Bull. Int. Stat. Inst. **47**, 497 (1977).
[19] P. Donnelly, Theor. Popul. Biol. **30**, 271 (1986).
[20] G. A. Watterson, Genetics **88**, 405 (1977).
[21] K. E. Dingle *et al.*, J. Clin. Microbiol. **39**, 14 (2001).
[22] M. C. J. Maiden *et al.*, Proc. Natl. Acad. Sci. U.S.A. **95**, 3140 (1998).
[23] Note, however, that for small θ any error in its estimation may influence the distribution of V_D; see Fig. 4.