Adv. Appl. Prob. **14,** 225–239 (1982)
Printed in N. Ireland
0001-8678/82/020225-15$01.75

THE NUMBER OF ALLELES AND SEGREGATING SITES IN A SAMPLE FROM THE INFINITE-ALLELES MODEL

R. C. GRIFFITHS,* *Monash University*

Abstract

A study is made of the joint distribution of the number of alleles and segregating sites in a random sample of genes from the infinite-alleles model of population genetics. Tables of the joint distribution are given for representative values of the mutation parameter. A conclusion is that each segregating site has a high probability of delineating an allele type.

INFINITE-ALLELES MODEL; INFINITE-SITES MODEL; POPULATION GENETICS

1. Introduction

In the infinite-alleles model of population genetics the probability distribution of the number of allele types, k, in a random sample of n genes from a stationary population has been derived by Ewens (1972) as

$$(1.1)\qquad (\theta^k/\theta_{(n)})\,|S_n^{(k)}|, \qquad k=1,\cdots,n$$

where $\{S_n^{(k)}\}$ are Stirling numbers of the first kind, θ is a mutation-rate parameter, and $\theta_{(n)}=\theta(\theta+1)\cdots(\theta+n-1)$. The corresponding probability generating function (p.g.f.) is

$$(1.2)\qquad (\theta w)_{(n)}/\theta_{(n)}.$$

A gene in this model can be thought of at the microscopic level as an infinite sequence of completely linked sites which reproduce together. Mutation occurs at a site not previously segregating in the population and never at the same site twice. Different sequences are then identified as different alleles. The p.g.f. of the number of segregating sites in a random sample of n from a stationary population is

$$(1.3)\qquad (n-1)!\Big/\prod_{r=2}^{n}(r+\theta(1-z)-1)$$

derived in Watterson (1975). In this paper the joint distribution of the number of alleles and segregating sites in a sample is studied. The main conclusion is

Received 8 May 1981.

* Postal address: Mathematics Department, Monash University, Clayton, VIC 3168, Australia.

that there is a very high correlation between the number of alleles and segregating sites and the numbers are asymptotically the same for large sample sizes. Table 1 (pp. 235–236) of the joint probability distribution also shows that there is a high probability that a segregating site will delineate an allele type.

A full description of the model and discussion is given in Ewens (1979). The model can be derived from a limiting K-allele Wright–Fisher model, with equal mutation rates between the K allele types, and this is the approach used here.

2. Derivation of the joint probability generating function of the number of segregating sites and alleles in a sample

In a neutral K-allele discrete-generation Wright–Fisher model with equal mutation rates between types let u be the mutation rate per gene per generation from one type to another and $2N$ the gene population size. The transition probability of moving from $(y_1, \cdots, y_K) \to (x_1, \cdots, x_K)$ genes of the different types in one generation is

$$\frac{(2N)!}{x_1! \cdots x_K!} \pi_1^{x_1} \cdots \pi_K^{x_K},$$

where

$$\pi_j = \{y_j(1-(K-1)u) + u(1-y_j)\}/2N, \qquad j = 1, \cdots, K.$$

Let there be one allele type initially and denote $P_r^\tau(n_1, \cdots, n_k; l)$ as the probability, in a random sample of r genes at generation τ, that there are $n_1, \cdots, n_K$ genes of the different types and a total of l mutations have occurred to the sample and its parents since a single common ancestor of the sample. Let

$$P_r^\tau(n_1, \cdots, n_K; z) = \sum_{l=0}^{\infty} z^l P_r^\tau(n_1, \cdots, n_K; l).$$

As $K \to \infty$ while Ku is held constant the process becomes an infinite-allele Wright–Fisher model at the macroscopic level. At a microscopic level the number of mutations which have occurred since a common ancestor of the sample is in the limit equivalent to the number of segregating sites in the sample, where a mutation begins a new segregating site in an infinite sequence of completely linked sites. Some results are first obtained for K finite.

To order N^{-1} there are r or $r-1$ parents of a sample of r genes, and a single mutation may only occur if there are r parents. The probabilities of having r or $r-1$ parents are, to order N^{-1}, $q_r = 1 - r(r-1)/4N$, $q_{r-1} = r(r-1)/4N$. This leads to the recurrence relationship

$$\begin{aligned}P_r^{\tau+1}&(n_1, \cdots, n_K; z)\\ &= q_r P_r^{\tau}(n_1, \cdots, n_K; z)(1-(K-1)u)^r\\ &\quad + q_r \sum_{i\neq j} (n_{j+1}) P_r^{\tau}(n_1, \cdots, n_i - 1, \cdots, n_j + 1, \cdots, n_K; z)(1-(K-1)u)^{r-1} zu\\ &\quad + q_{r-1} \sum_i (n_i - 1) P_{r-1}^{\tau}(n_1, \cdots, n_i - 1, \cdots, n_K; z)/(r-1) + O(N^{-2}).\end{aligned}$$

To obtain a diffusion limit let $N \to \infty$ while measuring time $t = 2N\tau$ and holding $\theta = 4Nu(K-1)$ fixed. Then in the limit, with a clear notation,

$$\begin{aligned}&\frac{2d}{dt} P_r(n_1, \cdots, n_K; z; t) + r(r-1+\theta) P_r(n_1, \cdots, n_K; z; t)\\ \text{(2.1)}\qquad &= \theta z/(K-1) \sum_{i\neq j} P_r(n_1, \cdots, n_i - 1, \cdots, n_j + 1, \cdots, n_K; z; t)(n_j + 1)\\ &\quad + r(r-1) \sum_i P_{r-1}(n_1, \cdots, n_i - 1, \cdots, n_K; z; t)(n_i - 1)/(r-1), \quad r \geqq 2.\end{aligned}$$

Terms on the right where $n_i - 1$ occurs and $n_i = 0$ are to be interpreted as 0.

If instead of a Wright–Fisher model a Moran model is used the development is very similar.

To simplify (2.1) at stationarity interpret

$$P_r(n_1, \cdots, n_K; z; \infty) = (r!/n_1! \cdots n_K!) \mathscr{L}_K(X_1^{n_1} \cdots X_K^{n_K}),$$

where $\mathscr{L}_K$ is a linear operator on the space of real K-variable polynomials. Denote $\mathbf{s}$ as a column vector of arbitrary constants $s_1, \cdots, s_K$ and $\mathbf{X}$ of the K variables $X_1, \cdots, X_K$. Multiplying the stationary version of (2.1), where

$$\lim_{t\to\infty} \frac{d}{dt} P_r(n_1, \cdots, n_K; z; t) = 0,$$

by $s_1^{n_1} \cdots s_K^{n_K}$ and summing over $n_1 + \cdots + n_K = r$,

$$\begin{aligned}&(r-1+\theta+\theta z/(K-1)) \mathscr{L}_K(\mathbf{s}'\mathbf{X})^r\\ \text{(2.2)}\qquad &= \{\theta z/(K-1)\} \sum_i s_i \mathscr{L}_K\{Y(\mathbf{s}'\mathbf{X})^{r-1}\} + (r-1)\mathscr{L}_K\{(\mathbf{s}'\mathbf{X})^{r-2} \sum_i s_i^2 X_i\},\\ &\text{where} \quad Y = X_1 + \cdots + X_K.\end{aligned}$$

The joint p.g.f. of the number of mutations which have occurred in a sample since a common ancestor, S, and the number of alleles in the sample, A, is

$$\text{(2.3)}\qquad Q_K(z, w) = \sum_{n=1}^{K} \binom{K}{n} w^n (1-w)^{K-n} \mathscr{L}_K\{(X_1 + \cdots + X_n)^r\}.$$

This is obtained in the following way. Let

$$I_j=\begin{cases}1 & \text{if allele } j \text{ is in the sample}\\ 0 & \text{otherwise.}\end{cases}$$

The joint p.g.f. is then

$$\mathscr{E}\left\{z^S\prod_{j=1}^{K}(wI_j+1-I_j)\right\}=\mathscr{E}\left\{z^S\prod_{j=1}^{K}(w+(1-w)(1-I_j))\right\}. \tag{2.4}$$

The expected value of a disjoint product $z^S(1-I_{l_1})\cdots(1-I_{l_{K-n}})$ is the probability/p.g.f. that allele types $l_1,\cdots,l_{K-n}$ do not occur in the sample; by symmetry this is $\mathscr{L}_K\{(X_1+\cdots+X_n)^r\}$. The joint p.g.f. of the number of segregating sites in a sample and the number of alleles in the infinite alleles model is

$$\begin{aligned}Q(z,w)&=\lim_{K\to\infty}Q_K(z,w)\\ &=\lim_{K\to\infty,n/K\to w}\mathscr{L}_K\{(X_1+\cdots+X_n)^r\},\end{aligned}$$

because of the functional form of Q_K.

Let $X=X_1+\cdots+X_n$ and place

$$s_i=\begin{cases}s+1, & i=1,\cdots,n\\ 1, & i=n+1,\cdots,K\end{cases}$$

in (2.2) to obtain

$$\begin{aligned}&\{r-1+\theta Kz/(K-1)+\theta(1-z)\}\mathscr{L}_K(sX+Y)^r\\ &\quad=\{\theta z(ns+K)/(K-1)\}\mathscr{L}_K\{Y(sX+Y)^{r-1}\}+(r-1)(s+2)\mathscr{L}_K(sX+Y)^{r-1}\\ &\qquad-(r-1)(s+1)\mathscr{L}_K\{Y(sX+Y)^{r-2}\}\qquad r=2,\cdots.\end{aligned} \tag{2.5}$$

Define

$$\mathscr{L}(X^aY^b)=\lim_{K\to\infty,n/K\to w}\mathscr{L}_K(X^aY^b). \tag{2.6}$$

Then

$$\begin{aligned}&(r-1+\theta)\mathscr{L}(sX+Y)^r\\ &\quad=\theta z(ws+1)\mathscr{L}\{Y(sX+Y)^{r-1}\}+(r-1)(s+2)\mathscr{L}(sX+Y)^{r-1}\\ &\qquad-(r-1)(s+1)\mathscr{L}\{Y(sX+Y)^{r-2}\}\qquad r=2,\cdots.\end{aligned} \tag{2.7}$$

Note also that (2.5) shows that the limit $\mathscr{L}_K(X^aY^b)\to\mathscr{L}(X^aY^b)$ is well defined by induction on the degree $a+b$. When $a+b=1$, $\mathscr{L}_K(Y)=1$, a sample of 1 is its own ancestor, and by symmetry $\mathscr{L}_K(X)=n/K \to w$.

The joint p.g.f. in a random sample of n is $Q(z, w)=\mathscr{L}(X^n)$. Unfortunately this does not seem to have a simple expression, but can be derived from a recurrence relationship connecting $\mathscr{L}(X^mY^{n-m})$. The marginal p.g.f.'s of S and A are respectively,

$$\mathscr{L}(Y^n)=(n-1)!\Big/\prod_{j=2}^{n}\{j+\theta(1-z)-1\}$$

and

$$\mathscr{L}(X^n)=(\theta w)_{(n)}/\theta_{(n)}, \quad \text{when} \quad z=1.$$

The following theorem is proved by equating coefficients of s^m in (2.7) and denoting $q(m, r)=\mathscr{L}(X^mY^{r-m})$. The boundary condition $q(1, r)=w\mathscr{L}(Y^r)$ is true by symmetry.

Theorem 1. The joint probability generating function of the number of segregating sites, S, and the number of alleles, A, in a random sample of n from the infinite alleles diffusion model is $Q(z, w; n)=q(n, n)$, where $q(n, n)$ is the n, nth element in a triangular $n\times n$ matrix whose elements are defined recursively by

$$\begin{aligned}&\{r(r-1+\theta)-(r-m)\theta z\}q(m, r)\\ (2.8)\qquad &\quad =(r+m-1)(r-m)q(m, r-1)\\ &\quad\quad +m\theta zwq(m-1, r)+m(m-1)q(m-1, r-1),\\ &\qquad\qquad\qquad\qquad m=2,\cdots, n; \quad r=m,\cdots, n.\end{aligned}$$

Recursion proceeds

$$m=2, \quad r=2,\cdots, n; \quad m=3, \quad r=3,\cdots, n;\cdots; m=n, \quad r=n.$$

The boundary conditions are

$$q(1, 1)=w$$

$$q(1, r)=w(r-1)!\Big/\prod_{j=2}^{r}\{j+\theta(1-z)-1\}, \qquad r=2,\cdots, n.$$

It is clear from (2.8) and an induction proof that $q(m, r)$ is a p.g.f. for each pair (m, r) and

$$q(m, r)=\begin{cases}\mathscr{L}(X^m) & \text{if} \quad z=1\\ \mathscr{L}(Y^r) & \text{if} \quad w=1.\end{cases}$$

Corollary 1. Let $p(i, j; m, r)$ be the probability associated with the p.g.f.

$g(m, r)$. Then

$$\begin{aligned} &r(r-1+\theta)p(i, j; m, r) \\ (2.9)\qquad &\quad = (r-m)\theta p(i-1, j; m, r) + (r+m-1)(r-m)p(i, j; m, r-1) \\ &\qquad + m\theta p(i-1, j-1; m-1, r) + m(m-1)p(i, j; m-1, r-1), \\ &\quad i = 0, 1, \cdots, m; \quad j = 1, 2, \cdots \quad \text{and} \quad r \geqq m \geqq 2. \end{aligned}$$

The boundary probabilities satisfy

$$\begin{aligned} p(i, j; 1, r) &= 0, & j &> 1 \\ (r+\theta-1)p(i, 1; 1, r) &= (r-1)p(i, 1; 1, r-1) + \theta p(i, 1; 1, r), & r &= 2, 3, \cdots \\ p(i, 1; 1, 1) &= \delta_{i0}. \end{aligned}$$

The probabilities $p(i, j; m, r)$ are 0 if $i+1 < j$.

Although Theorem 1 does not give an explicit formula for the p.g.f. probabilities and moments can be calculated recursively. A computing algorithm for the probabilities $\{p(i, j; n, n)\}$ is given in Section 3.

Corollary 2. Let V be a triangular matrix whose elements are indexed by two numbers,

$$\begin{aligned} V(m, r; m, r) &= -r(r+\theta-1) + (r-m)\theta z \\ V(m, r; m-1, r) &= m\theta z w \\ V(m, r; m, r-1) &= r(r-1) - m(m-1) \\ V(m, r; m-1, r-1) &= m(m-1), \qquad m = 2, \cdots, n; \quad r = m, \cdots, n. \\ V(1, r; 1, r) &= -(r + \theta(1-z) - 1) \\ V(1, r; 1, r-1) &= r-1 \\ V(1, 1; 1, 1) &= 1, \qquad r = 2, \cdots, n. \end{aligned}$$

Then,

$$q(m, r) = wV^{-1}(m, r; 1, 1). \tag{2.10}$$

Proof. A restatement of (2.8) is $V\boldsymbol{q} = w\boldsymbol{e}$, where $\boldsymbol{q}$ is the vector $\{q(m, r)\}$ and $\boldsymbol{e}$ is the vector $\{\delta_{m1}\,\delta_{r1}\}$. Being triangular, with non-zero elements on the diagonal, V can be inverted and this gives (2.10).

Theorem 2. The joint distribution of (A, S) in a random sample of n genes has the following representation.

Consider a random walk on the triangular integer lattice $r \geqq 1, m \geqq 1, m \leqq r$ (with r the horizontal axis and m the vertical axis) begun at (n, n) which moves independently at each time unit. Transitions are made according to the following scheme.

Transition	*Probability*	
$(r, m) \to (r, m)$	$\varepsilon_{rm} = (r-m)\theta/r(r+\theta-1)$	
$(r, m) \to (r-1, m)$	$\alpha_{rm} = \{r(r-1)-m(m-1)\}/r(r+\theta-1)$	
$(r, m) \to (r, m-1)$	$\gamma_{rm} = m\theta/r(r+\theta-1)$	
$(r, m) \to (r-1, m-1)$	$\beta_{rm} = m(m-1)/r(r+\theta-1),$	$m \geqq 2$
$(r, 1) \to (r, 1)$	$\varepsilon_{r1} = \theta/(r+\theta-1)$	
$(r, 1) \to (r-1, 1)$	$\alpha_{r1} = (r-1)/(r+\theta-1).$	

There is a single absorbing state at (1, 1).

Denote ξ_n as the total number of times that the random walk does not change position before absorption and η_n the total number of transitions vertically down ($\leqq n-1$). Then

$$A = 1 + \eta_n$$
$$S = \eta_n + \xi_n. \tag{2.11}$$

Proof. Rearranging (2.8),

$$\begin{aligned} q(m, r) &= z\varepsilon_{rm}q(m, r) + zw\gamma_{rm}q(m-1, r) \\ &\quad + \alpha_{rm}q(m, r-1) + \beta_{rm}q(m-1, r-1), \qquad r \geqq m \geqq 2. \end{aligned}$$

and

$$q(1, r) = z(1-\alpha_{r1})q(1, r) + \alpha_{r1}q(1, r-1).$$

The probabilistic interpretation (2.11) is now clear from these equations.

The marginal distributions of A, S are easier to interpret in the random walk than the joint distribution, which is complicated. The probability that when a transition is made from row m to row $m-1$ it is made vertically down is $\mathscr{E}\{\gamma_{Rm}/(\gamma_{Rm}+\beta_{Rm})\} = \theta/(\theta+m-1)$, where R is the random column the transition is made from. This does not depend on the columns, so η_n is a 1-dimensional random walk with $n-1$ steps and p.g.f. (1.2).

The number of segregating sites gained in the rth column has a geometric distribution with a mean of $\varepsilon_{rm} + \gamma_{rm} = \theta/(r+\theta-1)$. This is seen by noting that the probability in one transition of staying in column r from any row m is $\varepsilon_{rm} + \gamma_{rm}$, which does not depend on m. At each time unit spent in the column a segregating site is gained. Thus the p.g.f. of the number of segregating sites is (1.3), a 1-dimensional random walk with $n-1$ steps.

If the sample size is large ξ_n is small compared to η_n.

$$\mathscr{E}(\eta_n)=\mathscr{E}(A)-1=\theta\sum_{j=1}^{n-1}(j+\theta)^{-1}\approx\theta\log n$$

while

$$\begin{aligned}\mathscr{E}(\xi_n)=\mathscr{E}(S-A+1)&=\theta\sum_{j=1}^{n-1}(j^{-1}-(j+\theta)^{-1})\\&=\theta^2\sum_{j=1}^{n-1}[j(j+\theta)]^{-1}\\&\leqq 1{\cdot}65\theta^2\quad\text{for all}\quad n.\end{aligned}$$

As $n\rightarrow\infty$, $\mathscr{E}(\xi_n)\rightarrow\theta(\gamma+\psi(1+\theta))$, where $\psi(z)$ is the digamma function, tabulated in Abramowitz and Stegun (1970). The joint limit distribution of

$$\{A-\theta\log n\}/(\theta\log n)^{\frac{1}{2}},\qquad\{S-\theta\log n\}/(\theta\log n)^{\frac{1}{2}}$$

is a singular bivariate normal distribution (U, V), where $U=V$, since the standardized number of alleles converges to normality and ξ_n is bounded as $n\rightarrow\infty$. In particular,

$$\text{correlation }(A, S)\rightarrow 1\quad\text{as}\quad n\rightarrow\infty.$$

Let $(M_n=n, N_n)$, (M_{n-1}, N_{n-1}), $\cdots$, (M_2, N_2) be the entry and exit positions from columns $n,\cdots,2$. The p.g.f. of the number of segregating sites gained in the rth column conditional on (M_r, N_r) is

$$z^{M_r-N_r}\prod_{m=N_r}^{M_r}(1-\mu_{rm}(z-1))^{-1},$$

where $\mu_{rm}=\varepsilon_{rm}/(1-\varepsilon_{rm})=(r-m)\theta/\{r(r-1)+m\theta\}$, and the number of alleles gained is M_r-N_r. The joint p.g.f. is therefore

$$w\mathscr{E}\left\{\prod_{r=2}^{n}(zw)^{M_r-N_r}\prod_{m=N_r}^{M_r}(1-\mu_{rm}(z-1))^{-1}\right\}\tag{2.12}$$

which unfortunately does not seem to have a simple form.

The p.g.f. (2.12) shows the structure

$$\eta_n=\sum_{r=2}^{n}(M_r-N_r),$$

$$\xi_n=\sum_{r=2}^{n}\sum_{m=N_r}^{M_r}X_{rm},$$

where $\{X_{rm}\}$ are mutually independent geometric random variables with means $\{\mu_{rm}\}$, and are independent of the sequence $\{(M_r, N_r)\}$.

The pairs $(M_n, N_n), \cdots, (M_2, N_2)$ form a reverse Markov chain,

$$P(M_r = a, N_r = b \mid (M_{r+1} = c, N_{r+1} = d), \cdots, (M_n, N_n)) \tag{2.13}$$

$$= \frac{h(a, d)a!\,(r-1)}{b!\,[b+1+r(r-1)/\theta]_{(a-b)}[r(r-1)+b\theta](r+1)}$$

where

$$h(a, d) = \begin{cases} r(r+1) - a(a-1), & a = d \\ a(a+1), & a = d-1 \\ 0, & \text{otherwise} \end{cases}$$

and $a \geqq b \geqq 1$.

3. Numerical results

The recurrence relationship (2.9) for calculating the probability distribution of the number of alleles and segregating sites in a random sample of n is not straightforward so a computing algorithm is given here. The algorithm is written so that it is possible to have large sample sizes without needing too much storage space.

```
    Input:  $\theta$ = mutation parameter
            $n$ = sample size
            $l$ = maximum number of alleles for which the probability
                distribution will be calculated
            $k$ = maximum number of segregating sites for which the probabil-
                ity distribution will be calculated.
    Dimension:  $A(k+1)$ $B(k+1)$ $C(l)$ $D(l)$ $V(k+1, l, l+1)$
                $W(k+1, l, l+1)$
      $B(0) = 1,\ D(1) = 1,\ V(0, 1, 1) = 1$
      Do 10  $r = 2, l$
      $(r+\theta-1)V(0, 1, r) = (r-1)V(0, 1, r-1)$
      Do 10  $i = 1, k$
10    $(r+\theta-1)V(i, 1, r) = (r-1)V(i, 1, r-1) + \theta V(i-1, 1, r)$
      Do 11  $m = 2, n$
      Do 12  $r = 1, l$
      $(r+m-1)(r+m+\theta-2)W(0, 1, r) = (r+2m-2)(r-1)W(0, 1, r-1)$
                                        $+ m(m-1)V(0, 1, r)$
      Do 13  $i = 1, k$
13    $(r+m-1)(r+m+\theta-2)W(i, 1, r) = (r+2m-2)(r-1)W(i, 1, r-1)$
                                        $+ m(m-1)V(i, 1, r)$
                                        $+ (r-1)\theta W(i-1, 1, r)$
      Do 12  $j = 2, \min(l, m)$
      Do 12  $i = j-1, k$
```

```
12    $(r+m-1)(r+m+\theta-2)W(i,j,r)=(r+2m-2)(r-1)W(i,j,r-1)$
                                   $+m(m-1)V(i,j,r)$
                                   $+(r-1)\theta W(i-1,1,r)$
                                   $+m\theta V(i-1,j-1,r+1)$
      Do 14   $r=1,l$
      Do 14   $j=1,\min(l,m)$
      Do 14   $i=j-1,k$
14    $V(i,j,r)=W(i,j,r)$
      $(m+\theta-1)A(0)=(m-1)B(0)$
      Do 15   $i=1,k$
15    $(m+\theta-1)A(i)=(m-1)B(i)+\theta A(i-1)$
      $(m+\theta-1)C(1)=(m-1)D(1)$
      Do 16   $j=2,l$
16    $(m+\theta-1)C(j)=(m-1)D(j)+\theta D(j-1)$
      Do 17   $i=0,k$
17    $B(i)=A(i)$
      Do 18   $j=1,l$
18    $D(j)=C(j)$
```

Print out the probability distributions

$$P(S=i, A=j; m)=W(i,j;1), \quad P(S=i)=A(i), \quad P(A=j)=C(j)$$

for $i=0,1,\cdots,k; j=1,2,\cdots,\min(m,l)$ as an option for intermediate sample sizes $m=2,\cdots,n-1$ and print if $m=n$.

```
11    Continue
```

The joint probability distribution of (A, S) was calculated for different values of θ and $n=100$ as an illustration and is given in Table 1. Only probabilities which are greater than 0·0000 are shown. A general impression is that the most probable values are concentrated around $(A, S=A+1)$, $(A, S=A)$, $(A, S=A-1)$. Of course $A \leqq S+1$ always. For small allele values from one to around the mean number of alleles the most probable pairs are $(A, S=A+1)$, with the probabilities decreasing for a given A down the columns. For larger allele values the most probable pairs are when S is larger than $A+1$. The table shows that there is a very strong structure in a sequence of completely linked sites and segregation is far from random. Provided θ and the sample size are not too small it seems that when a site is segregating it very often delineates an allele.

As θ increases the segregating sites are closer to being randomly distributed on the genes and the probabilities are not so concentrated on the diagonal.

The means, variances, covariances and correlation of the number of alleles and number of segregating sites in a sample are given in Table 2 for representative values of θ. A striking result is the very high correlation between the

TABLE 1

Joint probability distribution of the number of alleles and segregating sites in a random sample of 100 genes

(a) $\theta = 0{\cdot}1$

segregating	alleles						
sites	1	2	3	4	5	6	>6
0	0·6005						
1		0·3018					
2		0·0084	0·0716				
3		0·0006	0·0036	0·0108			
4		0·0001	0·0003	0·0007	0·0012		
5				0·0001	0·0001	0·0001	
>5							

(b) $\theta = 0{\cdot}5$

segregating	alleles										
sites	1	2	3	4	5	6	7	8	9	10	>10
0	0·0887										
1		0·2027									
2		0·0194	0·2224								
3		0·0052	0·0396	0·1565							
4		0·0017	0·0115	0·0391	0·0796						
5		0·0005	0·0038	0·0122	0·0250	0·0312					
6		0·0002	0·0013	0·0041	0·0083	0·0116	0·0099				
7		0·0001	0·0004	0·0014	0·0028	0·0041	0·0042	0·0026			
8			0·0001	0·0005	0·0010	0·0014	0·0015	0·0012	0·0006		
9				0·0002	0·0003	0·0005	0·0005	0·0005	0·0003	0·0001	
10				0·0001	0·0001	0·0002	0·0002	0·0002	0·0001	0·0001	
11						0·0001	0·0001	0·0001			
>11											

(c) $\theta = 1{\cdot}0$

segregating	alleles													
sites	1	2	3	4	5	6	7	8	9	10	11	12	13	>13
0	0·0100													
1		0·0419												
2		0·0057	0·0851											
3		0·0022	0·0222	0·1120										
4		0·0010	0·0093	0·0419	0·1072									
5		0·0005	0·0046	0·0193	0·0512	0·0797								
6		0·0003	0·0023	0·0097	0·0255	0·0457	0·0479							
7		0·0001	0·0012	0·0050	0·0132	0·0244	0·0317	0·0240						
8		0·0001	0·0006	0·0025	0·0069	0·0130	0·0179	0·0178	0·0102					
9			0·0003	0·0013	0·0035	0·0068	0·0098	0·0106	0·0084	0·0038				
10			0·0001	0·0006	0·0018	0·0035	0·0052	0·0060	0·0052	0·0033	0·0012			
11			0·0001	0·0003	0·0009	0·0018	0·0027	0·0032	0·0030	0·0022	0·0012	0·0003		
12				0·0002	0·0005	0·0009	0·0014	0·0017	0·0016	0·0013	0·0008	0·0004	0·0001	
13				0·0001	0·0002	0·0005	0·0007	0·0009	0·0009	0·0007	0·0005	0·0002	0·0001	
14					0·0001	0·0002	0·0004	0·0005	0·0005	0·0004	0·0003	0·0002	0·0001	
15					0·0001	0·0001	0·0002	0·0002	0·0002	0·0002	0·0001	0·0001		
>15					0·0001	0·0001	0·0002	0·0002	0·0002	0·0002	0·0002	0·0001	0·0001	

TABLE 1(*cont'd*)

(d) $\theta = 1{\cdot}5$

segretating	alleles															
sites	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	15
0	0·0013															
1		0·0078														
2		0·0012	0·0223													
3		0·0005	0·0069	0·0415												
4		0·0003	0·0034	0·0187	0·0566											
5		0·0002	0·0020	0·0102	0·0330	0·0602										
6		0·0001	0·0012	0·0061	0·0197	0·0425	0·0520									
7		0·0001	0·0007	0·0038	0·0123	0·0274	0·0427	0·0375								
8			0·0004	0·0023	0·0077	0·0177	0·0295	0·0349	0·0231							
9			0·0003	0·0014	0·0048	0·0113	0·0197	0·0256	0·0238	0·0123						
10			0·0002	0·0009	0·0030	0·0071	0·0128	0·0176	0·0185	0·0139	0·0058					
11			0·0001	0·0005	0·0018	0·0044	0·0082	0·0117	0·0131	0·0113	0·0070	0·0024				
12			0·0001	0·0003	0·0011	0·0027	0·0051	0·0075	0·0088	0·0082	0·0060	0·0031	0·0009			
13				0·0002	0·0007	0·0017	0·0032	0·0048	0·0058	0·0056	0·0045	0·0028	0·0012	0·0003		
14				0·0001	0·0004	0·0010	0·0020	0·0030	0·0037	0·0037	0·0031	0·0021	0·0011	0·0004	0·0001	
15				0·0001	0·0002	0·0006	0·0012	0·0018	0·0023	0·0024	0·0021	0·0015	0·0009	0·0004	0·0001	
>15				0·0001	0·0004	0·0009	0·0018	0·0029	0·0037	0·0039	0·0036	0·0028	0·0018	0·0001	0·0005	0·0003

TABLE 2

sample size	$\mathscr{E}(A)$	$\mathscr{E}(S)$	Var (A)	Var (S)	Cov (A, S)	Corr (A, S)
			$\theta = 0{\cdot}01$			
2	1·0099	0·0100	0·0098	0·0101	0·0099	0·9950
10	1·0281	0·0283	0·0280	0·0284	0·0281	0·9973
50	1·0446	0·0448	0·0445	0·0450	0·0446	0·9982
100	1·0516	0·0518	0·0515	0·0519	0·0516	0·9984
200	1·0586	0·0587	0·0584	0·0589	0·0586	0·9986
500	1·0677	0·0679	0·0676	0·0681	0·0677	0·9988
			$\theta = 0{\cdot}5$			
2	1·3333	0·5000	0·2222	0·7500	0·3333	0·8165
10	2·1333	1·4145	0·9245	1·7994	1·1212	0·8692
50	2·9378	2·2396	1·7091	2·6458	1·9242	0·9049
100	3·2843	2·5887	2·0531	2·9974	2·2707	0·9153
200	3·6309	2·9365	2·3985	3·3465	2·6173	0·9238
500	4·0891	3·3954	2·8559	3·8061	3·0754	0·9328
			$\theta = 1{\cdot}0$			
2	1·5000	1·0000	0·2500	2·0000	0·5000	0·7071
10	2·9290	2·8290	1·3792	4·3687	1·8701	0·7619
50	4·4992	4·4792	2·8741	6·1039	3·4313	0·8192
100	5·1874	5·1774	3·5524	6·8123	4·1190	0·8373
200	5·8780	5·8730	4·2381	7·5130	4·8095	0·8523
500	6·7928	6·7908	5·1499	8·4338	5·7242	0·8686
			$\theta = 1{\cdot}5$			
2	1·6000	1·5000	0·2400	3·7500	0·6000	0·6325
10	3·5426	4·2435	1·6437	7·7094	2·4107	0·6773
50	5·8430	6·7188	3·7838	10·3745	4·6863	0·7480
100	6·8680	7·7661	4·7869	11·4446	5·7098	0·7714
200	7·9002	8·8095	5·8081	12·4994	6·7416	0·7912
500	9·2702	10·1862	7·1714	13·8828	8·1114	0·8129
			$\theta = 2{\cdot}0$			
2	1·6667	2·0000	0·2222	6·0000	0·6667	0·5774
10	4·0398	5·6579	1·8076	11·8170	2·8191	0·6100
50	7·0376	8·9584	4·5356	15·4573	5·7681	0·6889
100	8·3946	10·3548	5·8542	16·8943	7·1219	0·7161
200	9·7660	11·7461	7·2061	18·3057	8·4924	0·7394
500	11·5896	13·5816	9·0179	20·1534	10·3157	0·7652

number of alleles and segregating sites in the sample. Indeed as the sample size increases the correlation tends to 1 as shown in Section 2. The conditional expectations were calculated for representative values of θ and are given in Tables 3 and 4. Table 3 shows that for larger sample sizes $S < \mathscr{E}(A \mid S) \leqq S+1$, and that $\mathscr{E}(A \mid S)$ increases as the sample size increases, probably to $S+1$. As θ increases $\mathscr{E}(A \mid S)$ increases. Table 4 is of particular interest, since in practice

TABLE 3
Expected number of alleles, given S segregating sites

	number of segregating sites, S			
sample size	2	3	4	5
		$\theta = 0{\cdot}01$		
10	2·6789	3·1005	3·3457	3·4823
50	2·8514	3·5472	4·0912	4·4974
100	2·8861	3·6477	4·2805	4·7866
200	2·9101	3·7197	4·4210	5·0104
500	2·9319	3·7864	4·5550	5·2310
		$\theta = 0{\cdot}5$		
10	2·7357	3·2447	3·5816	3·7990
50	2·8916	3·6679	4·3271	4·8726
100	2·9199	3·7520	4·4910	5·1346
200	2·9386	3·8090	4·6052	5·3228
500	2·9549	3·8593	4·7078	5·4957
		$\theta = 1{\cdot}0$		
10	2·7678	3·3297	3·7273	4·0033
50	2·9126	3·7319	4·4553	5·0834
100	2·9370	3·8054	4·6008	5·3202
200	2·9528	3·8536	4·6980	5·4824
500	2·9661	3·8947	4·7821	5·6251
		$\theta = 1.5$		
10	2·7880	3·3847	3·8240	4·1425
50	2·9252	3·7708	4·5340	5·2145
100	2·9472	3·8372	4·6664	5·4321
200	2·9611	3·8796	4·7523	5·5763
500	2·9725	3·9150	4·8247	5·6993
		$\theta = 2{\cdot}0$		
10	2·8019	3·4233	3·8930	4·2433
50	2·9336	3·7968	4·5871	5·3037
100	2·9539	3·8582	4·7099	5·5067
200	2·9665	3·8967	4·7879	5·6377
500	2·9767	3·9281	4·8522	5·7470

the number of alleles in a sample can be observed more easily than the number of segregating sites. For the parameter values in the table $A-1<\mathscr{E}(S\mid A)\leqq A+1$, though as $\theta\rightarrow\infty$, $\mathscr{E}(S\mid A)\rightarrow\infty$ for any value of A. As the sample size increases, $\mathscr{E}(S\mid A)$ decreases, probably to $A-1$.

Simple values of $\mathscr{E}(S\mid A)$ and $\mathscr{E}(A\mid S)$ are not included in Tables 2 or 3. $\mathscr{E}(A\mid S=0)=1$, $\mathscr{E}(A\mid S=1)=2$, $\mathscr{E}(S\mid A=1)=0$, whatever θ or the sample size.

Computing algorithms for Tables 2, 3 and 4 are available from the author.

TABLE 4
Expected number of segregating sites, given A alleles

	number of alleles, A			
sample size	2	3	4	5
$\theta = 0{\cdot}01$				
10	1·0054	2·0092	3·0120	4·0143
50	1·0036	2·0064	3·0086	4·0104
100	1·0032	2·0056	3·0076	4·0093
200	1·0028	2·0050	3·0068	4·0084
500	1·0024	2·0044	3·0060	4·0074
$\theta = 0{\cdot}5$				
10	1·2872	2·4818	3·6266	4·7410
50	1·1918	2·3354	3·4486	4·5412
100	1·1670	2·2955	3·3986	4·4842
200	1·1476	2·2639	3·3588	4·4383
500	1·1279	2·2312	3·3170	4·3899
$\theta = 1{\cdot}0$				
10	1·6002	3·0004	4·2951	5·5265
50	1·4021	2·7002	3·9335	5·1235
100	1·3500	2·6172	3·8304	5·0064
200	1·3094	2·5515	3·7479	4·9118
500	1·2681	2·4833	3·6613	4·8120
$\theta = 1{\cdot}5$				
10	1·9340	3·5482	4·9965	6·3470
50	1·6279	3·0895	4·4485	5·7396
100	1·5466	2·9610	4·2899	5·5600
200	1·4832	2·8590	4·1624	5·4146
500	1·4186	2·7531	4·0285	5·2608
$\theta = 2{\cdot}0$				
10	2·2850	4·1197	5·7246	7·1954
50	1·8672	3·4999	4·9893	6·3844
100	1·7549	3·3237	4·7729	6·1405
200	1·6674	3·1836	4·5986	5·9424
500	1·5781	3·0380	4·4153	5·7325

References

ABRAMOWITZ, M. AND STEGUN, I. A. (1970) *Handbook of Mathematical Functions.* Dover, New York.

EWENS, W. J. (1972) The sampling theory of selectively neutral alleles. *Theoret. Popn Biol.* **3,** 87–112.

EWENS, W. J. (1979) *Mathematical Population Genetics.* Biomathematics **9,** Springer-Verlag, Berlin.

WATTERSON, G. A. (1975) On the number of segregating sites in genetical models without recombination. *Theoret. Popn Biol.* **7,** 256–276.