J. Indones. Math. Soc.
Vol. 16, No. 2 (2010), pp. 89–104.

ALLELE FREQUENCIES IN MULTIGENE FAMILIES

S. Padmadisastra

Department of Statistics, Universitas Padjadjaran
Jalan Raya Bandung Sumedang Km 21 Jatinangor, Indonesia,
s_padmadisastra@yahoo.com

Abstract. Results on allelle frequencies in three chromosomes, drawn at random from a diploid population, evolving in equilibrium, at a particular generation, are presented in this paper. The genes on each chromosome are subject to unbiased and reciprocal gene conversion and mutation. Using the coalescent approach we find the probability distribution of the allelic configurations in the three chromosomes, and the moments of the allelic numbers that exist in one of the three chromosomes or in a pair of chromosomes. We also consider the identity coefficients of two genes drawn at random, one from each of two chromosomes, and the probability that all genes in the three chromosomes are monomorphic. Numerical examples are also given together with simulation results, and they agree well.

Key words and Phrases: Coalescent approach, allelic configurations, identity coefficients, monomorphic probability.

Abstrak. Dalam makalah ini dibahas mengenai frekuensi alel dalam tiga buah kromosom yang diambil secara acak, pada suatu generasi tertentu, dari sebuah populasi diploid. Dianggap bahwa populasi berada dalam keadaan stabil (equilibrium). Gen dalam tiap kromosom dapat mengalami mutasi dan konversi. Menggunakan pendekatan koalisi (coalescent) diperoleh distribusi peluang dari konfigurasi alel dalam ketiga kromosom, dan momen-momen mengenai banyak alel yang ada dalam sebuah dan sepasang kromosom. Selain itu diperoleh juga koefisien identitas dari dua buah gen yang diambil secara acak, masing-masing satu dari tiap kromosom, dan peluang bahwa semua gen dalam ketiga kromosom monomorfik. Hasil yang diperoleh ternyata sangat sesuai dengan yang diperoleh melalui simulasi.

Kata kunci: Pendekatan koalisi, konfigurasi alel, koefisien identitas, peluang monomorfik.

2000 Mathematics Subject Classification: 62P10.
Received: 23-08-2010, accepted: 08-11-2010.

1. Introduction

Watterson [19] studied the joint allelic configurations of two chromosomes in a population. Here we extend his results by considering one more chromosome; thus the object of study here is the joint allelic configuration in three chromosomes. We employ similar techniques to his; namely the "coalescent approach". In this approach, as we proceed backward in time, two chromosomes coalesce into one class if they are descended from the same ancestor. This approach has been successful in tackling many problems in population genetics; see for instance Griffiths [2], Kingman [5,6], Tavare [14] and Watterson [15, 16] and Kaplan and Hudson [3].

We also use a similar model of evolution; namely that the population follows a Wright-Fisher model, in which each offspring generation is formed through sampling at random with replacement from its parent generation at the end of its life-cycle. Thus generations do not overlap. Here we assume that the population consist of $2N$chromosomes, it evolves in equilibrium and within each chromosome there is a multigene family of n genes. During the formation of a generation, any of the n genes in the chromosome might mutate into a novel allele, never before seen, with rate $\nu/2N$per gene per generation. Thus the number of possible alleles is infinite; this is the infinitely many alleles model of Kimura and Crow [4]. It may also happen that during this time one gene, in a chromosome, converts the type of another gene, on the same chromosome, into its type. We let the probability that the i^{th} gene will convert the j^{th} gene to have its allelic type be $\lambda/2N$ per generation.

The gene conversions are unbiased and reciprocal. We assume that ν, λ, and n are $O(1)$ and that N is large so that there is negligible chance of more than one mutation or gene conversion per chromosome per generation. Further, we assume in this model there is no crossing-over.

In the present study, some analytic formulas for the quantities of interest are obtained and these are illustrated by several numerical examples. We also conduct a simulation study to assess the agreement between theoretical and simulation results. The simulation is done using Watterson's method of simulation [17, 18] and [19]. Simulation is done only for three sampled chromosomes not for the whole population as was done by Ohta [10]. In Section 2 we obtain the main result of this paper, namely the probability generating function of the allelic configurations in the three chromosomes. Some applications of this main result are presented in Section 3. There we study the Trivariate frequency spectrum, Identity coefficients, Number of alleles and Probability of monomorphism. But unfortunately, the analytic formulas of some interesting quantities have not been derived; thus we turn to simulation to study their behaviour. The identity coefficients presented in this paper are the probabilities that two (three) randomly chosen genes from two (three) different chromosomes would be identical. Studies by Ohta [8, 9, 10, 11], Nagylaki and Barton [7] and Kaplan and Hudson [3] present identity coefficients that consider not only different chromosomes but also whether the genes in question are at the same locus or at different loci in the chromosomes.

2. **Main Results**

If we trace back the ancestorship of the three sampled chromosomes, drawn at random from a particular generation of a Wright-Fisher model, it may happen that two of them were descended from a "most recent common ancestor" (MRCA1) in some particular generation, a generation more recent than when the whole three were descended from their most recent common ancestor (MRCA2); see Fig. l. Denote by T_l the time between the present and when the first two coalesce into one class; i.e. the time for their MRCA1, and by T_2 the time when these first two, after coalescing into one group, join together with the remaining chromosome into one class. Here time is measured in units of 2N generations.

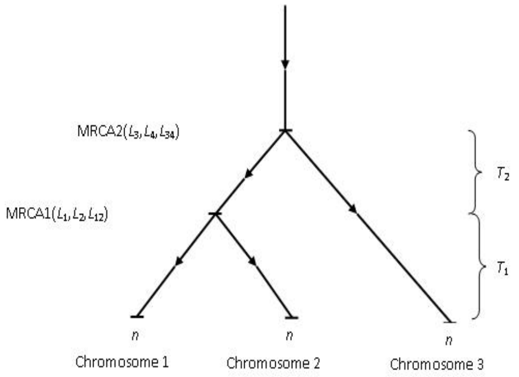

Figure 1. Chromosome configurations

In Fig. l, we have numbered the two chromosomes that join together first as "1" and "2" and the remaining one, "3." The symbols L will be explained shortly. Of course, in practice, we may not be able to determine which pair of chromosomes diverged most recently; however, for our theoretical results we will assume that this is known.

It can be shown that, in the limit, when $2N \to \infty$, the approximate joint probability density function of (T_1, T_2) is

$$f(t_1, t_2) = 3e^{-(3t_1+t_2)} \quad t_1 \geq 0, t_2 \geq 0 \tag{1}$$

conditional on "1" and "2" joining first. Thus T_l and T_2 are conditionally independent and have exponential densities with mean $\frac{1}{3}$ and 1 respectively.

The joint allelic configurations in the three chromosomes will first be derived conditional on the joint allelic configurations in MRCA1 and MRCA2. We will call a gene from MRCA1 (MRCA2) a "founder" gene, provided it has at least one non-mutant descendant in Chromosome 1 or 2 (in MRCA1 or Chromosome 3). Let us suppose that the numbers of founder genes in MRCA1 for Chromosomes 1 and 2 be L_1 and L_2. Since MRCA1 is also a random sample of one chromosome drawn at time T_1+1, then let us suppose that there are L_3 and L_4 founder genes in MRCA2 for Chromosomes 3 and MRCAl, see Fig. 1.

We can consider the genes in the MRCAl and MRCA2 as consisting of four subsets; namely,

(1) the MRCAl genes which, when copied into the two daughter chromosomes, become founder genes in daughter chromosome "1" but not founder genes in daughter chromosome "2",
(2) those genes which become founder genes in chromosome "2" but not founder genes in chromosome "1",
(3) those genes which become founder genes in both chromosomes "1" and "2",
(4) those genes in the MRCAl which do not become founder genes in either daughter chromosome.

A similar description as above can be given for MRCA2, but to save space we omit it.

We write L_{12} for the number of MRCAl genes in the third subset. Then the four subsets are of sizes $L_1 - L_{12}, L_2 - L_{12}, L_{12}$ and $n - L_1 - L_2 + L_{12}$ respectively. Similarly we write L_{34} for the number of MRCA2 genes which are founders for both Chromosome 3 and MRCAI genes. Then the four subsets of MRCA2 are of sizes $L_3 - L_{34}, L_4 - L_{34}, L_{34}$ and $n - L_3 - L_4 + L_{34}$ respectively.

Of course, in the present situation, the statement that the evolution of a chromosome is mathematically equivalent to the evolution of a population of n genes in Moran's model, proved by Watterson [18] under Shimizu's [12] model for the evolution of multigene families, will also hold. Thus the conditional distribution of the numbers of founder genes, L, which are ancestors for n, non-mutant genes at time T later, is

$$P(L = l|T = t, n) = q_l(t, n), \tag{2}$$

where

$$q_l(t,n) = \sum_{j=l}^{n} e^{-\lambda j + (j+\theta-1)t} a(j,l,n) \quad l = 0, \dots, n, \tag{3}$$

and for $j > 0$,

$$a(j,l,n) = (-1)^{j-l} \frac{(2j+\theta-1)(l+\theta)_{(j-1)} n_{[j]}}{l!(j-l)!(n+\theta)_{(j)}} \tag{4}$$

while for $j = 0$, $a(0,0,n) = 1$. In the above formula, we use the ascending and descending factorial notations;

$$\theta_{(j)} = \theta(\theta+1)\cdots(\theta+j-1), \text{and} \quad \theta_{[j]} = \theta(\theta-1)\cdots(\theta-j+1).$$

Further in the present case the probability distribution of the random variables L_{12} and L_{34}, conditional on given values of $L_1 = l_1$ and $L_2 = l_2$, and of $L_3 = l_3$ and $L_4 = l_4$, say $h(l_1, l_2; l_{12})$ and $h(l_3, 1_4; l_{34})$, will also be hypergeometric as is noted in Watterson [19]. See Eqns. (2.7) - (2.8) in Watterson [19] for their form.

Let $\mathbf{L} = (L_1, L_2, L_{12}, L_3, L_4, L_{34})$, $\mathbf{T} = (T_1, T_2)$ and $P(\mathbf{l}|\mathbf{T})$ denote the probability that $\mathbf{L}$ has value $\mathbf{l}$ given $\mathbf{T}$. Averaging $P(\mathbf{l}|\mathbf{t})$ with respect to (1), we have the joint unconditional distribution of $\mathbf{L}$, $P(\mathbf{l})$,

$$\begin{aligned} P(\mathbf{l}) &= 3h(l_1, l_2, l_{12})\, h(l_3, l_4, l_{34}), \sum_{j_1=l_1}^{n} \sum_{j_2=l_2}^{n} \sum_{j_3=l_3}^{n} \sum_{j_4=l_4}^{l_1+l_2+l_{12}} a(j_1, l_1, n) \\ &\quad \times a(j_2, l_2, n)\, a(j_3, l_3, n)\, a(j_4, l_4, l_1 + l_2 - l_{12}) \qquad (5) \\ &\quad \times \left[3 + \lambda \sum_{i=1}^{3} j_i(j_i + \theta - 1)\right]^{-1} \left[1 + \lambda \sum_{i=3}^{4} j_i(j_i + \theta - 1)\right]^{-1} \end{aligned}$$

Now to describe the joint allelic configurations in the three chromosomes, let Γ_{j_1,j_2,j_3} denote the number of alleles each represented by j_1, j_2 and j_3 genes in our sampled chromosomes, respectively. We obtain the p.g.f. for the trivariate frequency spectrum, $\{\Gamma_{j_1,j_2,j_3}\}$, of the three sampled chromosomes

$$\begin{aligned} E\left(\prod_{j_1,j_2,j_3} V_{j_1,j_2,j_3}^{\Gamma_{j_1,j_2,j_3}}\right) &= \sum_{\mathbf{l}} P(\mathbf{l})\, X(\theta, \mathbf{l}) \qquad (6) \\ &\quad \times \text{coefficient of } \phi_1^n \phi_2^n \phi_3^n \psi_1^{l_4} \psi_2^{l_3} p^{l_1} q^{l_2} \rho^{l_{12}} \eta^{l_{34}} \text{in } G \end{aligned}$$

where

$$\begin{aligned} X(\theta, \mathbf{l}) &= \prod_{i=1}^{3} \frac{(n - l_1)!}{(\theta + l_i)_{(n-l_i)}} \left(\binom{l_1 + l_2 - l_{12}}{l_1 - l_{12}, l_2 - l_{12}}\right)^{-1} \\ &\quad \times \frac{(l_3 - l_{34})!(l_4 - l_{34})! l_{34}!(l_1 + l_2 - l_{12} - l_4)!}{\theta_{(l_3+l_4-l_{34})}(\theta + l_4)_{(l_1+l_2-l_{12}-l_4)}} \end{aligned}$$

$P(\mathbf{l})$ is given in (5) and

$$\begin{aligned} G = &\\ &exp\left\{\theta \sum_{\mathbf{i}} \sum_{\mathbf{j}} V_{\mathbf{j}} \phi_1^{j_1} \phi_2^{j_2} \phi_3^{j_3} \psi_1^{i_4} \psi_2^{i_3} p^{i_1} q^{i_2} \binom{j_1 - 1}{i_1 - 1} \binom{j_2 - 1}{i_2 - 1} \binom{j_3 - 1}{i_3 - 1}\right. \\ &\times \sum_{j_{12}} \sum_{j_{34}} \rho^{j_{12}} \eta^{j_{34}} \binom{i_1 + i_2 - j_{12} - 1}{i_4 - 1} \\ &\times \frac{(i_3 + i_4 - j_{34} - 1)!}{(i_3 - j_{34})!(i_4 - j_{34})! j_{34}!} \frac{(i_1 + i_2 - j_{12})!}{(i_1 - j_{12})!(i_2 - j_{12})! j_{12}!} \qquad (7) \\ &+ \theta \sum_{i_1,i_2} \sum_{j_1 \geq i_1} \sum_{j_2 \geq i_2} V_{j_1,j_2,0} \phi_1^{j_1} \phi_2^{j_2} p^{i_1} q^{i_2} \binom{j_1 - 1}{i_1 - 1} \binom{j_2 - 1}{i_2 - 1} \\ &\times \sum_{j_{12}} \rho^{j_{12}} \frac{(i_1 + i_2 - j_{12} - 1)!}{(i_1 - j_{12})!(i_2 - j_{12})! j_{12}!} \\ &\left. + \theta \sum_{j_1=1} V_{j_1,0,0} \phi_1^{j_1}/j_1 + \theta \sum_{j_2=1} V_{0,j_2,0} \phi_2^{j_2}/j_2 + \theta \sum_{j_3=1} V_{0,0,j_3} \phi_3^{j_3}/j_3 \right\}, \end{aligned}$$

where the sum over $\mathbf{i}$ means the sum over i_1, i_2, i_3, and i_4, with $i_1 + i_2 + i_3 \geq 1$ and the sum over $\mathbf{j}$ means the sum over $j_1 \geq i_1$, $j_2 \geq i_2$, and $j_3 \geq i_3$. It can be shown that the p.g.f. is correctly normalized.

3. Applications

3.1. **Trivariate Frequency Spectrum.** Differentiating (6) with respect to V_{j_1,j_2,j_3} and then putting all V's equal to 1 yields $E(\Gamma_{j_1,j_2,j_3})$. After some algebra, we found, for $j_1 + j_2 + j_3 \geq 1$

$$
\begin{aligned}
E(\Gamma_{j_1,j_2,j_3}) = & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (8) \\
& \theta \prod_{i=1}^{3} \frac{j_i!}{(\theta+n-1)_{[j_i]}} \sum_{\mathbf{l}} p(\mathbf{l}) \left\{ \sum_{\mathbf{i}} \sum_{j_{12}} \sum_{j_{34}} \frac{i_1 i_2 i_3}{j_1 j_2 j_3} \, \frac{i_4}{(i_1+i_2-j_{12})(i_3+i_4-j_{34})} \right. \\
& \times \binom{\theta+l_1-1}{i_1} \binom{\theta+l_2-1}{i_2} \binom{\theta+l_3-1}{i_3} \binom{\theta+l_4-1}{i_4} \\
& \times \binom{n-l_1}{j_1-i_1} \binom{n-l_2}{j_2-i_2} \binom{n-l_3}{j_3-i_3} \binom{l_1-l_{12}}{i_1-j_{12}} \binom{l_2-l_{12}}{i_2-j_{12}} \\
& \times \binom{l_{12}}{j_{12}} \binom{l_3-l_{34}}{i_3-j_{34}} \binom{l_4-l_{34}}{i_4-j_{34}} \binom{l_{34}}{j_{34}} \binom{l_1+l_2-l_{12}-l_4}{i_1+i_2-j_{12}-i_4} \\
& \times \left[\binom{l_1+l_2-l_{12}}{i_i+i_2-j_{12}} \binom{\theta+l_1+l_2-l_{12}-1}{i_1+i_2-j_{12}} \binom{\theta+l_3+l_4-l_{34}-1}{i_3+i_4-j_{34}} \right]^{-1} \\
& +\delta_{0,0,j_3} \sum_{i_1,i_2} \sum_{j_{12}} \frac{i_1 i_2}{j_1 j_2} \frac{1}{(i_1+i_2-j_{12})} \binom{\theta+l_1-1}{i_1} \binom{\theta+l_2-1}{i_2} \\
& \times \binom{n-l_1}{j_1-i_1} \binom{n-l_2}{j_2-i_2} \binom{l_1-l_{12}}{i_1-j_{12}} \binom{l_2-l_{12}}{i_2-j_{12}} \binom{l_{12}}{j_{12}} \\
& \times \binom{l_1+l_2-l_{12}-l_4}{i_1+i_2-j_{12}} \left[\binom{\theta+l_1+l_2-l_{12}-1}{i_1+i_2-j_{12}} \binom{l_1+l_2-l_{12}}{i_1+i_2-j_{12}} \right]^{-1} \\
& \left. +\delta_{0,j_2,j_3} \binom{n-l_1}{j_1} \Big/ j_1 + \delta_{j_1,0,j_3} \binom{n-l_2}{j_2} \Big/ j_2 + \delta_{j_1,j_2,0} \binom{n-l_3}{j_3} \Big/ j_3 \right\}.
\end{aligned}
$$

In the above we interpret i_1/j_1 as 1 if $i_1 = j_1 = 0$ and similarly for i_2/j_2, i_3/j_3. The $\mathbf{i}$ sum is over i_1, i_2, i_3, i_4 with $i_1 + i_2 + i_3 \geq 1$ and $i_4/(i_l + i_2 - j_{12}) = 1$ if $i_1 = i_2 = j_{12} = i_4 = 0$. The i_1, i_2 sum is over i_1, i_2 with $i_1 + i_2 \geq 1$. Here, of course, the following relations hold,

$$\sum_{j_1,j_2,j_3} j_1 E(\Gamma_{j_1,j_2,j_3}) \quad = \quad \sum_{j_1,j_2,j_3} j_2 E(\Gamma_{j_1,j_2,j_3}) \quad = \quad \sum_{j_1,j_2,j_3} j_3 E(\Gamma_{j_1,j_2,j_3}) = n$$

3.2. **Identity Coefficients.** The results above enable us to calculate the probability of the event that two (three) randomly choosen genes from two (three) different chromosomes would be identical. Let $\hat{I}_2$ ($\hat{I}_3$) denote the probability of the event in the two (three) chromosomes, then

$$\hat{I}_2 = \tag{9}$$
$$\frac{1}{3} \times \left(\sum_{j_1,j_2} \frac{j_1 j_2}{n^2} E(\Gamma_{j_1,j_2,0}) + \sum_{j_1,j_3} \frac{j_1 j_3}{n^2} E(\Gamma_{j_1,0,j_3}) + \sum_{j_2,j_3} \frac{j_2 j_3}{n^2} E(\Gamma_{0,j_2,j_3}) \right)$$

and

$$\hat{I}_3 \;=\; \sum_{j_1,j_2,j_3} \frac{j_1 j_2 j_3}{n^3} E(\Gamma_{j_1,j_2,j_3}) \tag{10}$$

Equations (9) and (10) are obtained using similar reasoning as in the two chromosomes case of Watterson [19], except that in the three chromosomes case, the probability that two genes drawn at random from two randomly chosen chromosomes, $\hat{I}_2$ or $\hat{I}$ in Watterson's notation, now is the average of the probabilities calculated for different pairs of chromosomes "1" and "2", or "1"and "3".

3.3. **Numbers of Alleles.** The number of alleles that exist in one, in a pair, or in all three sampled chromosomes can be expressed in terms of frequency spectrum Γ'$_{j_1,j_2,j_3}$. We write

$$K_1 = \sum_{j_1=1}^{n} \sum_{j_2=0}^{n} \sum_{j_3=0}^{n} \Gamma_{j_1,j_2,j_3}, \quad K_2 = \sum_{j_1=0}^{n} \sum_{j_2=1}^{n} \sum_{j_3=0}^{n} \Gamma_{j_1,j_2,j_3}$$

,

$$K_3 = \sum_{j_1=0}^{n} \sum_{j_2=0}^{n} \sum_{j_3=1}^{n} \Gamma_{j_1,j_2,j_3}$$

for the numbers of alleles in the first, second and third sampled chromosome. Let the number of alleles that exist on both chromosome 1 and 2 be K_{12} on 1 and 3 be K_{13}, and on 2 and 3 be K_{23}. Then

$$K_{12} = \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} \sum_{j_3=0}^{n} \Gamma_{j_1,j_2,j_3}, \quad K_{13} = \sum_{j_1=1}^{n} \sum_{j_2=0}^{n} \sum_{j_3=1}^{n} \Gamma_{j_1,j_2,j_3}$$

$$K_{23} = \sum_{j_1=0}^{n} \sum_{j_2=1}^{n} \sum_{j_3=1}^{n} \Gamma_{j_1,j_2,j_3}$$

Clearly the number of alleles that exist in all three chromosomes, K_{123} say, is

$$K_{123} = \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} \sum_{j_3=1}^{n} \Gamma_{j_1,j_2,j_3}$$

To find the joint p.g.f. of K_1, K_2, K_3, K_{12}, K_{13}, K_{23} and K_{123} we assign certain values to V_{j_1,j_2,j_3}, the dummy variables in the p.g.f. of Γ_{j_1,j_2,j_3} given in (7). The assignment is done according to the definition of K's given above. Thus into (6) we substitute

$$V_{j_1,j_2,j_3} = \begin{cases} S_1S_2S_3S_{12}S_{13}S_{23}S_{123} & \text{if } j_1 \geq 1,\ j_2 \geq 1,\ j_3 \geq 1, \\ S_1S_2S_{12} & \text{if } j_1 \geq 1,\ j_2 \geq 1,\ j_3 = 0, \\ S_1S_3S_{13} & \text{if } j_1 \geq 1,\ j_2 = 0,\ j_3 \geq 1, \\ S_2S_3S_{23} & \text{if } j_1 = 0,\ j_2 \geq 1,\ j_3 \geq 1, \\ S_1 & \text{if } j_1 \geq 1,\ j_2 = 0,\ j_3 = 0, \\ S_2 & \text{if } j_1 = 0,\ j_2 \geq 1,\ j_3 = 0, \\ S_3 & \text{if } j_1 = 0,\ j_2 = 0,\ j_3 \geq 1, \end{cases}$$

Then we obtain a formula for the joint p.g.f of K_1, K_2, K_3, K_{12}, K_{13}, K_{23}, and K_{123}, that is for $E(S_1^{K_1}S_2^{K_2}S_3^{K_3}S_{12}^{K_{12}}S_{13}^{K_{13}}S_{23}^{K_{23}}S_{123}^{K_{123}})$.

To find the p.g.f of K, the total number of alleles present in at least one of the three chromosomes, put $S_1 = S_2 = S_3 = S_{123} = S$ and $S_{12} = S_{13} = S_{23} = S^{-1}$, thus

$$\begin{aligned} E\left(S_1^{K_1}S_2^{K_2}S_3^{K_3}S_{12}^{K_{12}}S_{13}^{K_{13}}S_{23}^{K_{23}}S_{123}^{K_{123}}\right) &= E\left(S^{K_1+K_2+K_3-K_{12}-K_{13}-K_{23}+K_{123}}\right) \\ &= E\left(S^K\right). \end{aligned}$$

In turns out that,

$$\begin{aligned} &E(S^K) = \qquad (11) \\ &\sum_{\mathbf{l}} P(\mathbf{l}) \left(\prod_{j=l_1+1}^{n} \frac{j+\theta S-1}{j+\theta-1}\right) \left(\prod_{j=l_2+1}^{n} \frac{j+\theta S-1}{j+\theta-1}\right) \\ &\times \left(\prod_{j=l_3+1}^{n} \frac{j+\theta S-1}{j+\theta-1}\right) \left(\prod_{j=l_4+1}^{l_3+l_4-l_{34}} \frac{j+\theta S-1}{j+\theta-1}\right) \left(\prod_{j=1}^{l_1+l_2-l_{12}} \frac{j+\theta S-1}{j+\theta-1}\right) \end{aligned}$$

The above result shows that K may be written as a sum of independent Bernoulli random variables, given $\mathbf{L} = \mathbf{l}$; i.e.,

$$\begin{aligned} K = {} & (X_{1_n} + X_{1_{n-1}} + \cdots + X_{1_{l_1+1}}) + (X_{2_n} + X_{2_{n-1}} + \cdots + X_{2_{l_2+1}}) \qquad (12) \\ & + (X_{3_n} + X_{3_{n-1}} + \cdots + X_{3_{l_3+1}}) + (X_{4_{l_1+l_2-l_{12}}} + X_{4_{l_1+l_2-l_{12}-1}} + \cdots + X_{4_1}) \\ & + (X_{5_{l_3+l_4-l_{34}}} + X_{5_{l_3+l_4-l_{34}-1}} + \cdots + X_{5_{l_4+1}}) \end{aligned}$$

where

$$X_{i_j} = \begin{cases} 0, & \text{with probability } (j-1)/(j+\theta-1) \\ 1, & \text{with probability } \theta/(j+\theta-1) \end{cases}$$

The interpretation of these variables is that X_{1_j}, and X_{2_j}, denote the numbers (0 or 1) of mutations which occur during T_1 in chromosomes 1 and 2, while X_{3_j}, mutations occur during $T_1 + T_2$ in chromosome 3, and because of the model, these mutations will produce new alleles which exist only in one chromosome but not in the other two, respectively. Those that founded both chromosomes 1 and 2 but arose as mutants during T_2 are represented by X_{4_j}, and similiarly X_{5_j}, represents the number of mutations that gave rise to the founder genes that founded both chromosomes 3 and MRCA1.

To find moments of K we can use either (11) or (12). Either way we find

$$E(K) = \sum_{\mathbf{l}} P(\mathbf{l})\theta\, B_1 eq13) \qquad (13)$$

and

$$E(K(K-1)) \;=\; \sum_{\mathbf{l}} P(\mathbf{l})\theta^2\,(B_1^2 - B_2) \qquad (14)$$

where

$$\begin{aligned} B_1 \;=\; & \sum_{j=0}^{l_1+l_2-l_{12}-1} (\theta+j)^{-1} + \sum_{j=0}^{l_3-l_{34}-1} (\theta+j+l_4)^{-1} \\ & + \sum_{j=0}^{n-l_1-1} (\theta+j+l_1)^{-1} + \sum_{j=0}^{n-l_2-1} (\theta+j+l_2)^{-1} + \sum_{j=0}^{n-l_3-1} (\theta+j+l_3)^{-1} \\ B_2 \;=\; & \sum_{j=0}^{l_1+l_2-l_{12}-1} (\theta+j)^{-2} + \sum_{j=0}^{l_3-l_{34}-1} (\theta+j+l_4)^{-2} \\ & + \sum_{j=0}^{n-l_1-1} (\theta+j+l_1)^{-2} + \sum_{j=0}^{n-l_2-1} (\theta+j+l_2)^{-2} + \sum_{j=0}^{n-l_3-1} (\theta+j+l_3)^{-2} \end{aligned}$$

We can find, at least theoretically, the probability that $K = k$, for any $k = 1, 2, \ldots$, but the most interesting case is perhaps when $k = 1$. This will give the probability that all the genes in the three chromosomes have only one type. i.e. are monomorphic. Thus from (11)

$$\begin{aligned} P_{\text{mono}} \;=\; & P(K=1) = P(K_1 = K_2 = K_3 = K_{12} = K_{13} = K_{23} = K_{123} = 1) \\ =\; & \theta\left(\frac{(n-1)!}{\theta_n}\right)^3 \sum_{\mathbf{l}} P(\mathbf{l}) \frac{\theta_{l_1}\theta_{l_2}\theta_{l_3}\theta_{l_4}}{(l_1-1)!(l_2-1)!(l_3-1)!(l_4-1)!} \\ & \times \frac{(l_1+l_2-l_{12})!(l_3+l_4-l_{34})!}{\theta_{(l_1+l_2-l_{12})}\theta_{(l_3+l_4-l_{34})}} \qquad (15) \\ =\; & \text{coefficient of } S^1 \text{ in } E(S^K) \end{aligned}$$

We can interpret the meaning of the terms in the above expression as follows. The first factor θ is saying that the one allelic type arose by mutation at some time. The second factor, $\frac{(n-1)!}{\theta_n}$, is related to Ewens' sampling formula; i.e. $\frac{\theta(n-1)!}{\theta_n}$ is the probability that a sample of n genes contains only 1 allelic type. The applicability of the Ewens' distribution to one chromosome in this model was first shown by Shimizu [13] and discussed further by Watterson [19]. Other terms in the above expression can also be described in terms of subsamples having to be monomorphic. Finally, since we work conditionally on the number of founders, then we have to average over the founders distribution to get the final result for P_{mono}.

To study the the marginal behavior of (K_1, K_2, K_3), put $S_3 = S_{13} = S_{23} = S_{123} = 1$ and leave the other S's unchanged. After some algebra, we found

$$E(S_1^{K_1} S_2^{K_2} S_{12}^{K_{12}}) = \sum_{\mathbf{l}} P(\mathbf{l}) \sum_m H_m \tag{16}$$

where

$$\begin{aligned} H_m &= \frac{(l_1 - l_{12})!(l_2 - l_{12})!}{(\theta + l_1)_{(n-l_1)}(\theta + l_2)_{(n-l_2)}(\theta)_{(l_1+l_2-l_{12})}} \\ &\quad \times \frac{(\theta S_1 S_2 S_{12})_{(m+l_{12})}\,(\theta S_1 + l_{12} + m)_{(n-l_{12}-m)}(\theta S_2 + l_{12} + m)_{(n-l_{12}-m)}}{m!\,(l_1 - l_{12} - m)!\,(l_2 - l_{12} - m)!} \end{aligned}$$

In the calculation of the moments, other than $E(K_1)$ and $\mathrm{Var}(K_1)$, we have to use $P(\mathbf{l})$ as in (eq5) because the marginal distribution of (L_1, L_2, L_{12}) obtain by summing out L_3, L_4, and L_{34} is not the same as distribution of (L_1, L_2, L_3) in the two chromosomes case of Watterson [19]. Perhaps the mere knowledge that there is a Chromosome "3" older than "1" and "2", means that "1" and "2" are not randomly chosen.

Further, by putting all S's equal to 1 except $S_3 = S$, it can be shown what the p.g.f of K_3 equals

$$E(S^{K_3}) = \frac{(\theta S)_{(n)}}{\theta_{(n)}}$$

Thus the number of distinct alleles in Chromosome 3 also follows Ewens' sampling formula.

It is rather unfortunate that we can not exploit (6) any further; in particular we are not able to simplify any other p.g.f. that involves Chromosome 3, such as the marginal p.g.f. $E(S_1^{K_1} S_3^{K_3} S_{13}^{K_{13}})$ etc. But Chromosome 3's behaviour still can be studied through simulation.

4. **Discussion and Examples**

In this section we present some numerical example of the results obtained earlier together with some simulation results. The simulation is done using Watterson's method of simulation, Watterson [16,17,18,19]. In each replicate of simulation we produce three families of gene each of size n and all the simulation results presented here are obtained from 10000 replicates. Throughout the tables, the simulation result are given in parentheses. In general, in all of the tables, the simulation results agree very closely with the results calculated from the theoretical formulas, except in some entries where the latter are not available.

Table I shows some results of the number of alleles existing in Chromosome 1; i.e. K_1. (Because of symmetry between chromosomes 1 and 2, and as it is shown in section 3, Chromosome 3 has similar behavior with chromosome 1, we do not present any results of Chromosome 2 and 3.) the entries under $E(K_1)$ and $\mathrm{Var}(K_1)$ are obtained from Eqns. (3.10) and (3.11) in Watterson [19], the reason for this are explained in Section 3.3. as they should be, the values of $E(K_1)$ and $\mathrm{Var}(K_1)$ are exactly the same with those in Table I of Watterson [19]. Other features of the table are also similar to that found in the two chromosomes case; namely keeping the mutation parameter fixed at $\nu = 0.2$ and increasing the values of θ produces larger number of alleles in each chromosomes. While keeping θ constant at $\theta = 0.45$ and varying the values of ν does not produce any effect; the marginal moments are constant for different values of ν. The numerical results on the joint behavior of two chromosomes can be seen in Table II. In this table the moments of the number of common alleles between chromosomes 1 and 2, K_{12}, and between Chromosomes 1 and 3, K_{13} are given. $E(K_{13})$ is found from $K_{13} = \sum_{j_1=1}\sum_{j_2=0}\sum_{j_3=1}\Gamma_{j_1j_2j_3}$ and (8) and the $\mathrm{Var}(K_{13})$ values are from simulation. Although the marginal behaviours of K_1, K_2, and K_3 are the same, it turns out that behaviours of K_{12} and K_{13} are not. Comparing the values of $E(K_{12})$ with those of $E(K_{13})$, we can say that $E(K_{12})$ has higher values than $E(K_{13})$ in any combination of parameter values, as we expect. One reason for this is because Chromosomes 1 and 2, at time of sampling, will have lower diversity than Chromosomes 1 and 3 since the

Table I. Alleles in One Chromosomes
Moments of K_1, K_2 and K_3

n	θ	$E(K_1)$	$\mathrm{Var}(K_1)$
		$\nu = 0.2$, throughout, $\theta = \nu/\lambda$	
2	0.045	1.0431 (1.0466)	0.0412 (0.0444)
	0.09	1.0826 (1.0824)	0.0758 (0.0756)
	0.225	1.1837 (1.1826)	0.1499 (0.1493)
	0.45	1.3103 (1.3080)	0.2140 (0.2132)
5	0.045	1.0910 (1.0960)	0.0883 (0.0904)
	0.09	1.1768 (1.1754)	0.1668 (0.1652)
	0.225	1.4078 (1.4072)	0.3562 (0.3556)
	0.45	1.7256 (1.7269)	0.5683 (0.5692)
10	0.045	1.1243 (1.1245)	0.1214 (0.1192)
	0.09	1.2429 (1.2442)	0.2320 (0.2381)
	0.225	1.5699 (1.5762)	0.5128 (0.5239)
	0.45	2.0392 (2.0409)	0.8615 (0.8539)
		$\theta = 0.45$ throughout	
10	0.02	2.3092 (2.0506)	0.8615 (0.8679)
	0.2	2.3092 (2.0409)	0.8615 (0.8463)
	2.0	2.3092 (2.0453)	0.8615 (0.8651)
		$\lambda = 0.44$ throughout	
10	0.02	1.1255 (1.1236)	0.1225 (0.1201)
	0.2	2.0479 (2.0512)	0.8674 (0.8493)
	2.0	5.6473 (5.6391)	2.0359 (2.0365)

Table II Alleles in Two Chromosomes
Moments of K_{12} and K_{13}

	$E(K_{12})$	$\mathrm{Var}(K_{12})$	$E(K_{13})$	$\mathrm{Var}(K_{13})$
θ	$\nu = 0.2, n = 2$ throughout, $\theta = \nu/\lambda$			
0.045	0.9047 (0.9049)	0.0982 (0.1007)	0.6443 (0.6457)	(0.2295)
0.09	0.9297 (0.9311)	0.1044 (0.1020)	0.6583 (0.6659)	(0.2256)
0.225	1.0052 (1.0052)	0.1445 (0.1392)	0.7003 (0.7011)	(0.2365)
0.45	1.1125 (1.1161)	0.2086 (0.2062)	0.7651 (0.7621)	(0.2747)
ν		$\theta = 0.45, n = 2$ throughout		
0.02	1.2847 (1.2843)	0.2057 (0.2059)	1.2159 (1.2129)	(0.2004)
0.2	1.1125 (1.1161)	0.2086 (0.2062)	0.7651 (0.7621)	(0.2747)
2.0	0.5205 (0.5126)	0.3143 (0.3117)	0.1020 (0.0989)	(0.0910)
ν		$\lambda = 0.44,\ n = 2$ throughout		
0.02	1.0219 (1.0425)	0.0315 (0.0315)	0.9746 (0.9770)	(0.0439)
0.2	1.1144 (1.1161)	0.2097 (0.2118)	0.7663 (0.7589)	(0.2742)
2.0	0.7556 (0.7443)	0.5701 (0.5720)	0.1480 (0.1459)	(0.1683)

former were descended from a more recent common ancestor than Chromosomes 1 and 3 did. Another noteworthy feature is that the values of $E(K_{12})$ are generally higher than those of $E(K_{12})$ in the two chromosomes case (Watterson [19]) but the opposite happens to $E(K_{13})$; i.e. its values are smaller. The effect of the parameter values on their joint behaviour is similar to that in the two chromosomes case.

We also note a relationship between $E(K_{12})$, $E(K_{13})$, $E(K_{23})$ in our results with $E(K_{12})$ in Watterson [19]. Their relationship can be formulated as

$$\begin{aligned} E(\text{Watterson's}\, K_{12}) &= \frac{1}{3}(E(K_{12} + E(K_{13} + E(K_{23}) \\ &= \frac{1}{3}E(K_{12}) + \frac{2}{3}E(K_{13}) \end{aligned}$$

For example, when $\nu = 0.2$, $n = 2$ and $\theta = 0.045$, from Table III of Watterson [19], we see $\mathrm{E}(K_{12}) = 0.731$. This exactly the average of $E(K_{12})$, $E(K_{13})$, and $E(K_{23})$ given in our Table II for same parameter values.

The behaviour of the number, K_{123}, of the alleles common to the three chromosomes, the number, K, of alleles altogether, and the probability that all alleles in all the three chromosomes are of one allelic type, are given in Table III. The expected number of alleles in common to all three chromosomes, $E(K_{123})$, is obtained from the relation $K_{123} = \sum_{j_1=1} \sum_{j_2=1} \sum_{j_3=1} \Gamma_{j_1 j_2 j_3}$ and (eq8), and the $\mathrm{Var}(K_{123})$ values are from simulation. Whereas $E(K)$ is from (13), $\mathrm{Var}(K)$ is from (13)-(14) and P_{mono} is from (15). As is expected, when the mutation parameter ν is held fixed, the number of alleles in common to all three chromosomes and the total number of alleles tend to increase as θ increases, while the probability of there being only one allele present decreases. On the other hand when θ is fixed and ν and λ

Tabel III. Alleles in Three Chromosomes
Moments of K_{123}, K, and $P_{\text{mono}} = P(K = 1)$

θ	$E(K_{123})$	$\text{Var}(K_{123})$	$E(K)$	$\text{Var}(K)$	P_{mono}
	$\nu = 0.2$, $n = 2$ throughout, $\theta = \nu/\lambda$				
0.045	0.6123	–	1.5483	0.4458	0.5435
	(0.6116)	(0.2378)	(1.5514)	(0.4470)	(0.5420)
0.09	0.6266	–	1.6280	0.5274	0.5032
	(0.6334)	(0.2338)	(1.6234)	(0.5268)	(0.5074)
0.225	0.6653	–	1.8105	0.6946	0.4175
	(0.6633)	(0.2444)	(1.7955)	(0.6817)	(0.4250)
0.45	0.7237	–	2.0121	0.8395	0.3304
	(0.7163)	(0.2800)	(2.0191)	(0.8420)	(0.3270)
ν	$\theta = 0.45, n = 2$ throughout				
0.02	1.2051	–	1.4196	0.3230	0.6183
	(1.2011)	(0.1949)	(1.4242)	(0.3235)	(0.6135)
0.2	0.7273	–	2.0121	0.8395	0.3304
	(0.0772)	(0.2800)	(2.0191)	(0.8420)	(0.3270)
2.0	0.0815	–	3.2881	1.0805	0.0321
	(0.0772)	(0.0722)	(3.2957)	(1.0856)	(0.0307)
ν	$\lambda = 0.44, n = 2$ throughout				
0.02	0.9682	–	1.1275	0.1248	0.8791
	(0.9691)	(0.0477)	(1.1243)	(0.1209)	(0.8817)
0.2	0.7248	–	2.0153	0.8415)	0.3290
	(0.7120)	(0.2781)	(2.0248)	(0.8385)	(0.3245)
2.0	0.1158	–	4.5232	1.2572	0.0049
	(0.1108)	(0.1277)	(4.5388)	(1.2892)	(0.0046)

are varied, the number of alleles expected on each chromosome remains constant while the number of alleles in common to two and three chromosomes decreases and the total number of alleles increases, as we expect. Holding λ fixed and increasing u has the same effect as above except that the number of alleles in each chromosome is not constant. Another obvious feature is that the number of alleles in common to all three chromosomes is always smaller than those in common to two chromosomes.

All the entries in Table IV are obtained from simulation, except those of covariances between the number of alleles in Chromosomes 1 and 2, $\text{Cov}(K_1, K_2)$. To save space, not all of the results found for Table IV are presented here. But the discussion given here are based on all the results found. Since the H_m term in (16) is exactly the same as H_m in (3.9) of Watterson [19], then Cov (K_1, K_2) will also be the same as $\text{Cov}(K_1, K_{12})$, see Watterson [19] for proof. Therefore we do not include both the theoretical and simulation values of $\text{Cov}(K_1, K_{12})$ into this table. The values of $\text{Cov}(K_1, K_2)$ are obtained from (3.15) of Watterson [19] but with

Table IV. Covariances

	$\mathrm{Cov}(K_1, K_2)$		$\mathrm{Cov}(K_1, K_3)$	$\mathrm{Cov}(K_1, K_{13})$
θ		$\nu = 0.2,\ n = 2$ throughout		
0.045	0.0057	(0.0076)	(0.0004)	(0.0003)
0.09	0.0179	(0.0169)	(0.0027)	(0.0022)
0.225	0.0612	(0.0607)	(0.0100)	(0.0126)
0.45	0.1151	(0.1184)	(0.0307)	(0.0296)
ν		$\theta = 0.45,\ n = 2$ throughout		
0.02	0.1971	(0.1963)	(0.1560)	0.1576)
0.2	0.1151	(0.1176)	(0.0307)	(0.0336)
2.0	0.0223	(0.0221)	(0)	(0.0019)
ν		$\lambda = 0.44,\ n = 2$ throughout		
0.02	0.0258	(0.0261)	(0.0106)	(0.0110)
0.2	0.1159	(0.1184)	(0.0307)	(0.0296)
2.0	0.0348	(0.0348)	(0.0041)	(0.0018)

different probability terms. Perhaps the most interesting feature of this table is the change from negative to positive covariances, in particular between the number of alleles in common to two chromosomes and the total number of alleles. The total number of alleles among the three chromosomes includes the alleles common to both Chromosomes 1 and 2. Thus one would expect a positive correlation between K and K_{12}. But when θ is low we expect little diversity, and perhaps the number of common alleles tends to inhibit the number of non-common alleles sufficiently to produce a negative correlation.

In Table V we give two illustrations of the trivariate frequency spectrum (8) and their simulation values. Similar to Table IV, to save space, this table is in reduced form. At the bottom of the table are the values of $\hat{I}_2$ and $\hat{I}_3$ as in (9) and (10) and P_{mono} values equalling $E(\Gamma_{2,2,2})$, as $n = 2$ here. The entries in the table show that, when $\theta = 0.045$, the dominant figures are of $E(\Gamma_{0,2,2})$, $E(\Gamma_{2,2,0})$, and $E(\Gamma_{2,2,2})$. Therefore there is a reasonably high chance for all the genes in the three families to have the same allelic type. If they are not all the same, then we can expect that all the genes in Chromosome 3 to have the same allelic type and the type is different with those in the other two. And it is more likely that all the genes in Chromosomes 1 and 2 will have the same allelic type than any other possibilities that can occur to Chromosomes 1 and 2. When $\theta = 0.45$, however, the $E(\Gamma_{2,2,2})$ is not dominant. Now there is a fair expectation for the alleles to be different on different chromosomes. In Chromosome 3, the chance that all the genes will be different is higher than the chance that they all will be the same. As to Chromosome 1 and 2, the chance that they will share only one aIlelic type which is represented by two genes in each chromosome is smaller than the chance they will share one allelic type with one representative gene in each. In Chromosome 1

Tabel V Trivariate Frequency Spectrum Γ_{j_1,j_2,j_3}
ν= 0.2, n = 2 trhoughout

j_1	j_2	j_3	$\theta = 0.045$	$\theta = 0.45$
0	0	1	0.0566 (0.0530)	0.3306 (0.3388)
0	0	2	0.3103 (0.3073)	0.1733 (0.1695)
0	1	0	0.0385 (0.0431)	0.1320 (0.1278)
	⋮		⋮	⋮
1	0	0	0.0385 (0.0404)	0.1320 (0.1278)
1	0	1	0.0002 (0.0003)	0.0155 (0.0176)
1	0	2	0.0027 (0.0036)	0.0128 (0.0015)
	⋮		⋮	⋮
2	2	0	0.2612 (0.2585)	0.1365 (0.1370)
2	2	1	0.0240 (0.0206)	0.0980 (0.0989)
2	2	2	0.5435 (0.5420)	0.3304 (0.3270)
$\hat{I}_2$			0.3343	0.2578
$\hat{I}_3$			0.5435	0.3304
P_{mono}			0.5435	
P_{mono}				0.3304

itself, two alleles with one representative gene each is more likely than one allele with two representatives.

Acknowledgement. I thank jurusan Statistika, Fmipa – Unpad, for their generous support.

References

[1] Ewens, W. J., "The Sampling Theory of Selectively Neutral Alleles", *Theor. Pop. Biol.* **3** (1972), 87-112.

[2] Griffiths, R. C., "Lines of Descent in the Diffusion Approximation of Neutral Wright-Fisher Models", *Theor. Pop. Biol.* **17** (1980), 37-50.

[3] Kaplan, N. L. and Hudson, R. R., "On the Divergence of Genes. I. Multigene Families ", *Theor. Pop. Biol.* **31** (1987), 178-194.

[4] Kimura, M. and Crow, J. F., "The Number of Alleles that can be Maintained in a Finite Population", *Genetics* **49** (1964), 725-738.

[5] Kingman, J. F. C., "On Genealogy of Large Populations", *J. Appl. Prob. A* **19** (1982a), 27-43.

[6] Kingman, J. F. C., "The Coalescent", *Stoch. Proc. Appl.* **13** (1982b), 235-248.

[7] Nagylaki, T. and Barton, N., "Intrachromosomal Gene Conversion, Linkage, and the Evolution of Multigene Families", ", *Theor. Pop. Biol.* **29** (1986), 407-437.

[8] Ohta, T., "On the Evolution of Multigene Families", *Theor. Pop. Biol.* **23** (1983), 216-240.

[9] Ohta, T., "Some Models of Gene Conversion for Treating the Evolution of Multigene Families", *Genetics* **106** (1984), 517-528.

[10] Ohta, T., "Actual Number of Alleles Contained in a Multigene Family", *Genet. Res.* **48** (1986), 119-123.

[11] Ohta T., "Some Models of Gene Conversion for Treating the Evolution of Multigene Families and Other Repetitive DNA Sequenes", in Stochastic models in Biology (M. Kimura, G. Kallianpur, T. Hida, Eds.) *Theor. Pop. Biol.* **23** (1987), 216-240.

[12] Shimizu, A., *Stationary Distribution of a Diffusion Process Taking Values in Probability Distributions on the Partitions*, in Stochastic models in Biology (M. Kimura, G. Kallianpur, T. Hida, Eds.) Springer-Verlag Berlin, 1987.

[13] Tavare. S., "Lines of Descent and Genealogical Processes, and Their Applications in Population Genetics Models", *Theor. Pop. Biol.* **26** (1984), 119-164.

[14] Watterson, G. A., "Lines of Descent and the Coalescent", *Theor. Pop. Biol.* **26** (1984a), 77-92.

[15] Watterson, G. A., "Allele Frequencies after a Bottleneck", *Theor. Pop. Biol.* **26** (1984b), 387-407.

[16] Watterson, G. A., "Genetic Divergence of Two Populations", *Theor. Pop. Biol.* **27** (1985a), 298-317.

[17] Watterson, G. A., *Estimating Species Divergence Times using Multilocus Data*, in Population Genetics and Molecular Evolution (T. Ohta and K. Aoki, Eds.) Springer-Verlag Berlin,1985b.

[18] Watterson, G. A., "Allele Frequencies in Multigene Families . I. Diffusion Equations Approach", *Theor. Pop. Biol.* **35** (1989a), 142-160.

[19] Watterson, G. A., "Allele Frequencies in Multigene Families . II. Coalescent Approach", *Theor. Pop. Biol.* **35** (1989b), 161-180.