AN ANALYSIS OF MULTI-ALLELIC DATA

G. A. WATTERSON

Mathematics Department, Monash University,
Clayton, Victoria, Australia 3168

Manuscript received May 31, 1977
Revised copy received August 25, 1977

ABSTRACT

Some multi-allelic data obtained by COYNE (1976) and by SINGH, LEWONTIN and FELTON (1976) are analyzed for their compatibility with the neutral alleles theory. It is found that strict neutrality appears not to be the case, but that if further alleles were to be distinguished in the samples, neutrality could become a possible explanation.

Section 1

THE papers by COYNE (1976) and SINGH, LEWONTIN and FELTON (1976) indicate that, if sufficiently great effort is expended in performing electrophoresis under various conditions, a considerable number of different alleles can be detected even in small samples of genes. COYNE found 23 distinct alleles in only 60 genes at the xanthine dehydrogenase locus of *Drosophila persimilis*, while SINGH, LEWONTIN and FELTON found 27 alleles in 146 genes from the xanthine dehydrogenase locus of *D. pseudoobscura* using four electrophoresis conditions, and an additional heat-sensitivity test revealed at least 37 alleles in that sample.

The purpose of the present paper is to analyze the allele frequencies reported in those papers to see if they are consistent with strict selective neutrality. We find, using various statistical tests, that neutrality is not a plausible explanation for the data. However, it will also be pointed out that if the more common "alleles" should really be subdivided into further alleles, the strict neutrality hypothesis may possibly be resurrected.

LEWONTIN (personal communication) has analyzed the data using an F approximation for the null hypothesis distribution of the information statistic. We here give exact results, or simulation results, for various test statistics. Our conclusions are qualitatively similar to those of LEWONTIN, and allow us to compare the accuracy of the F approximation.

Section 2. Results for the pooled samples

The 60 genes analyzed by COYNE (1976) may be summarized, as far as their allelic distribution is concerned, by the numbers

$$1^{18},\ 2^{3},\ 4^{1},\ 32^{1} \qquad (1)$$

with the interpretation that 18 alleles were represented once each in the sample, three alleles were represented twice each, one allele had four representatives, and the most common allele was represented by 32 genes. The total number of alleles is $18 + 3 + 1 + 1 = 23$, and the total number of genes is $(18 \times 1) + (3 \times 2) + (1 \times 4) + (1 \times 32) = 60$. More generally, we shall describe a sample by symbols of the type

$$n_1^{\alpha_1},\ n_2^{\alpha_2},\ n_3^{\alpha_3}, \ldots \tag{2}$$

where α_j is the number of alleles each having n_j representatives. The total number of alleles in the sample will be written

$$k = \alpha_1 + \alpha_2 + \alpha_3 + \ldots \tag{3}$$

and the total number of genes in the sample will be written

$$n = \alpha_1 n_1 + \alpha_2 n_2 + \alpha_3 n_3 + \ldots . \tag{4}$$

Assume that the sample (2) was drawn from a large population, in which statistical equilibrium has been reached, in which all genes (of whatever allelic type) have the same mutation rate and all mutations produce new alleles, and in which selection does not operate. Then, EWENS (1972) obtained the following probability for a random sample to be of composition (2), subject to *given* values of k and n as in (3) and (4),

$$\begin{aligned}&\Pr(n_1^{\alpha_1}, n_2^{\alpha_2}, n_3^{\alpha_3}, \ldots \mid k,n)\\ &= \frac{n!}{n_1^{\alpha_1} n_2^{\alpha_2} n_3^{\alpha_3} \ldots \alpha_1!\alpha_2!\alpha_3! \ldots |S_n^{(k)}|}\ ;\end{aligned} \tag{5}$$

where $S_n^{(k)}$ is a Stirling number of the first kind (see ABRAMOWITZ and STEGUN 1965 Table 24.3 for some numerical values). We do not use EWENS' notation, however; in particular, our n corresponds to his $2n$.

The beauty of (5) is that the conditioning on a given value of k has made the sample probability free of any unknown population parameters, especially the size of the population and the mutation rate. The probability (5) has been verified, at least as an approximation, for sampling from various population models subject to selective neutrality. See, for instance, KARLIN and McGREGOR (1972); WATTERSON (1974a,b); KELLY (1977); WATTERSON (1976); KINGMAN (1977).

In particular, the probability (5), when applied to the COYNE data (1), yields

$$\begin{aligned}\Pr(1^{18}, 2^3, 4^1, 32^1 \mid k = 23, n = 60) &= \frac{60!}{2^3.4.32.18!3!\,|S_{60}^{(23)}|}\\ &= 5.952 \times 10^{-7}\ .\end{aligned}$$

Similarly, the SINGH, LEWONTIN and FELTON data (obtained before the heat-sensitivity test was applied) had $k = 27$ alleles in $n = 146$ genes, the sample composition being

$$1^{10}, 2^3, 3^7, 5^2, 6^2, 8^1, 11^1, 68^1\ , \tag{6}$$

which has probability

$$\Pr(1^{10}, 2^3, 3^7, 5^2, 6^2, 8^1, 11^1, 68^1 \mid k=27,\ n=146)$$
$$= \frac{146!}{2^3.3^7.5^2.6^2.8.11.68.10!3!7!2!2!\ |S_{146}^{(27)}|}$$
$$= 2.326 \times 10^{-9}\ .$$

With a slight change of notation, let $\alpha(j)$ denote the observed number of alleles having j representatives in a sample, let $a(j) = E[\alpha(j)]$ and $\sigma(j)$ = standard deviation of $\alpha(j)$. Then using (4.5) in WATTERSON (1974a), we may calculate $a(j)$ and $\sigma(j)$ corresponding to distribution (5) for the data sets (1) and (6). We have:

COYNE data

j	1	2	3	4	5	6	7	...	32	...
$\alpha(j)$	18	3	0	1	0	0	0	...	1	...
$a(j)$	10.8	4.5	2.5	1.6	1.1	0.7	0.5	...	0.000004	...
$\sigma(j)$	1.9	1.9	1.5	1.2	1.0	0.8	0.7	...	0.002	...

SINGH, LEWONTIN and FELTON data

j	1	2	3	4	5	6	7	8	9	10	11	12	...	68	...
$\alpha(j)$	10	3	7	0	2	2	0	1	0	0	1	0	...	1	...
$a(j)$	8.9	4.2	2.7	1.9	1.4	1.1	0.9	0.8	0.6	0.5	0.5	0.4	...	0.0004	...
$\sigma(j)$	2.2	1.8	1.5	1.3	1.2	1.0	0.9	0.9	0.8	0.7	0.7	0.6	...	0.02	...

Note particularly that the most frequent allele in each case appears to be too frequent and, for the COYNE data, that the number of singleton alleles appears too high for neutrality.

The above calculations show that both sets of data are very unlikely to occur under the neutral alleles distribution (5). But of course each possible sample would have low probability, there being so many possible samples. Therefore, to assess whether either of the data sets (1) and (6) afford evidence against neutrality, we need to consider the probabilities of getting as extreme, *or more extreme*, data sets. Many methods of deciding which samples are "more extreme" have been suggested in the literature. We here concentrate on three such methods, more to illustrate the methods than in the belief that they are really appropriate in the present context.

The homozygosity test: The homozygosity of a sample, or a population, is defined as the sum of the squares of all allele relative frequencies. In the notation (2), it becomes

$$\hat{F} = \alpha_1 \left(\frac{n_1}{n}\right)^2 + \alpha_2 \left(\frac{n_2}{n}\right)^2 + \alpha_3 \left(\frac{n_3}{n}\right)^2 + \dots .$$

For the particular data (1) and (6), the homozygosities are $\hat{F} = 0.2972$ and $\hat{F} = 0.2353$, respectively. It takes a considerable amount of computer time to evaluate the significance level of these statistics by adding the probabilities (5) for all samples having these, or more extreme, $\hat{F}$ values. By computer simulation, however, it is very quick to generate samples having (5) as distribution and to *estimate* the significance levels by the proportion of more extreme $\hat{F}$ values in the simulated samples. We ran 1000 independent samples with $k = 23$, $n = 60$ and found that *no* sample had an $\hat{F}$ value as large or larger than $\hat{F} = 0.2972$, as for Coyne's data. Again, we ran 2000 independent samples with $k = 27$, $n = 146$ and found only 8 had as large an $\hat{F}$ value as did Singh, Lewontin and Felton's data. We may conclude that both data sets depart significantly from the neutral alleles distribution (5) in the direction of excess homozygosity (that is, a deficiency in heterozygosity).

The significance of the results is confirmed by the calculation of the means and standard deviations of $\hat{F}$ under the distribution (5). Formulas for these were given by Watterson (1977) and some numerical examples in Watterson (1978). In the present cases, with Coyne's data we would expect $\hat{F}$ to be 0.0831 with standard deviation 0.0180; the observed value is nearly 12 standard deviations too big. With the Singh, Lewontin and Felton data, we expect $\hat{F}$ to be 0.0984 with standard deviation 0.0278; the observed value is nearly 5 standard deviations too big.

The information statistic: Defining the information statistic as

$$I = -[\alpha_1 \frac{n_1}{n} \ln \left(\frac{n_1}{n}\right) + \alpha_2 \frac{n_2}{n} \ln \left(\frac{n_2}{n}\right) + \ldots]$$

we find that the Coyne and Singh, Lewontin and Felton data yield respective values of $I = 2.0842$ and $I = 2.2797$. These may be contrasted to their expected values 2.8029 and 2.7325, with standard deviations 0.0935 and 0.1416 respectively, the moments being computed using (4.5) in Watterson (1974a). The observed values are therefore more than 7, and more than 3, standard deviations below their means, respectively. This agrees with the significant results reported above (see also Singh, Lewontin and Felton (1977)).

The most frequent allele: In Coyne's data, the most frequent allele was represented by $32 = n_{\max}$ (say) genes; for the Singh, Lewontin and Felton data, $n_{\max} = 68$. Some significance levels for $n_{\max}$ were given by Ewens (1973), but for samples of 300 or more genes. A formula for the distribution of $n_{\max}$ was given by Watterson and Guess (1977, 4.7) and some mean values were tabulated by them. For the present k,n combinations we compute that the probabilities of getting $n_{\max}$ as large or larger than the observed values are 0.000007 and 0.002758, respectively. Again, the extreme significance of the Coyne data, and the significance of the Singh, Lewontin and Felton data, is noted. For confirmation, in simulations of 1000 samples in each case, we found mean values for $n_{\max}$ of 10.7 and 29.8 respectively, and standard deviations of 3.2 and 9.7 respectively. The observed values of $n_{\max}$ were thus over 6, and nearly 4, standard deviations above their respective means.

All of the above results should be treated with extreme skepticism. Among other assumptions, it has been assumed that the data sets are random samples from their respective species populations, and that all alleles are accurately detected and counted. We shall make some remarks about the latter assumption in the discussion below. Concerning the former assumption, we know that the data were collected from various populations, and that these should be analyzed separately. Fortunately, we can carry out such analyses because the authors reported their results in sufficient detail.

Section 3. Results for individual populations

In the notation (2), (3) and (4), the allele frequencies of COYNE's samples, which were drawn from three populations, may be summarized as

Fish Creek:	$1^7, 2^1, 15^1$	with $k = 9$ alleles in $n = 24$ genes,
Mather:	$1^9, 12^1$	with $k = 10$ alleles in $n = 21$ genes,
Sisters:	$1^{10}, 5^1$	with $k = 11$ alleles in $n = 15$ genes.

The Fish Creek sample could be equally, or more, extreme (with respect to all the statistics $\hat{F}$, I, and $n_{\max}$) if it had consisted of $1^8, 16^1$. Using the formula (5), the probability of getting the observed sample or the more extreme one is only 0.0042. LEWONTIN (personal communication) reports an F-approximation significant at the 0.001 level. Both the Mather and the Sisters samples are as extreme as possible, and have respective probabilities of only 0.0012 and 0.0262. Again, LEWONTIN reports significance levels < 0.001 and 0.004 using the F-approximation.

COYNE (1976 p. 604) remarked "Because of the small sample sizes, however, the degrees of freedom were too small for the values to achieve significance. Much larger sample sizes than those used here will be necessary to properly test the data for correspondence to neutrality". Whatever the validity of COYNE's remarks for the approximate tests he used, our exact tests above show that these samples have very significant departures from neutrality.

The SINGH, LEWONTIN, and FELTON data were drawn from twelve distinct populations. Two provided very small samples:

Population AU:	$1^2, 2^2$	with $k = 4$, $n = 6$,
Population GU:	2^2	with $k = 2$, $n = 4$,

and in each case there was only one other sample configuration possible having the same k,n values, namely $(1^3,3^1)$ and $(1^1,3^1)$, respectively. Whichever of these sample configurations had arisen, they could not have provided significant evidence against neutrality. For most of the other populations, only the most extreme possible sample would have been significant at the 5% level; this actually occurred with the HR population, whose composition $1^5,7^1$ with $k = 6$, $n = 12$ has probability 0.0427. Otherwise, although the samples were fairly extreme, they were too small to be significantly so. The exact probabilities for getting at least as extreme a sample (as large or larger homozygosity) for all twelve populations are compared below with the F-approximation probability obtained by LEWONTIN (personal communication) for the information statistic.

Population	SZ	SS	CN	SC	WR	CH	CE	MV	HR	AU	GU	BO
Probability	0.25	0.13	0.22	0.28	0.09	0.18	0.13	0.70	0.04	1	1	0.38
F approximation	0.14	0.08	0.08	0.18	0.05	0.16	0.08	0.59	0.01	0.83	1.00	0.36

DISCUSSION

In the above analyses, I have quoted one-sided significance levels, namely the probabilities for "neutral" samples to have as high, or higher, homozygosities than those observed. These may be doubled if two-sided tests and significance levels are required, at least in those majority of cases when the observed homozygosity is higher than its median value. In any case, there is strong evidence of non-neutrality in the individual populations of *D. persimilis*, and less strong evidence in the *D. pseudoobscura* data. It is interesting to note that the F-approximation underestimates the error probability. As we have remarked before, the conclusion relies on the data correctly indicating the allele frequencies in the samples. But SINGH, LEWONTIN, and FELTON found that by an additional heat sensitivity test, they could distinguish 37 (rather than 27) alleles in the 146 genomes examined. Combining their Tables 1 and 4, we find that the pooled sample of *D. pseudoobscura* had a composition, in order as listed by SINGH, LEWONTIN, and FELTON:

"Allele" frequency	1	1	5	3	1	1	8	1	2	5	2	1	3	3	6	2	6
Number of aliased alleles	1	1	1	1	1	1	1	1	1	2	2	1	1	1	2	1	2

											Total
"Allele" frequency	11	68	3	1	1	1	3	1	3	3	146
Number of aliased alleles	2	5	2	1	1	1	1	1	1	1	37

In particular, we see that the most frequent "allele", whose high frequency of 68 was the main cause of the significance of the results for these data in section 2, is really the combined total of (at least) five distinct alleles. It could well be that if the 68 genes were appropriately redistributed among 5 alleles (and the other aliased alleles' counts be similarly redistributed), the significance would evaporate. In fact, however, SINGH, LEWONTIN, and FELTON report that the most common allele among the five has frequency 45. If this were observed in a sample from a single population, with $k = 37$ and $n = 146$, it would still be significantly large. Unfortunately, due to some ambiguous lines with respect to heat sensitivity, the detailed compositions of the twelve populations' samples were not found.

It is an interesting consequence of neutral allele theory that we can calculate a probability for the above pooled data even though some "allele" frequencies are in fact totals for aliased alleles. Let $N_1, N_2, \ldots, N_k$ denote the numbers of each of k alleles in a sample size $n = N_1 + N_2 \ldots + N_k$. We assume that both k and n are given, and that the alleles are here listed in some order (for instance, in electrophoretic order). (So far, we have ignored the ordering of the alleles.)

The joint probability of getting a particular sample, under neutrality, is (see WATTERSON 1974a, 2.26)

$$\Pr(N_1 = n_1, N_2 = n_2, \ldots, N_k = n_k | k, n) = \frac{\prod_{j=1}^{k} f_1(n_j)}{f_k(n)} \qquad (7)$$

where $n_1, n_2, \ldots, n_k > 0$ and $n_1 + n_2 + \ldots + n_k = n$. Here $f_k(n)$ is a "first type Stirling distribution"

$$f_k(n) = \frac{k!}{n!} |S_n^{(k)}| \, \phi^n / [-\log(1-\phi)]^k, \; n = k, \; k+1, \ldots$$

where ϕ is an arbitrary parameter, $0 < \phi < 1$, which cancels out from (7). In particular, $f_1(.)$ is the logarithmic distribution, parameter ϕ. If we now add some of the allele numbers together, *e.g.*, let

$$S_1 = \sum_{j=1}^{k_1} N_j, \; S_2 = \sum_{j=k_1+1}^{k_1+k_2} N_j, \; S_3 = \sum_{j=k_1+k_2+1}^{k_1+k_2+k_3,} N_j, \ldots$$

be the totals of $k_1, k_2, k_3, \ldots$ allele counts, then (7) is replaced by

$$\Pr(S_1 = s_1, S_2 = s_2, \ldots | \sum_j S_j = n \, , \quad \sum_j k_j = k)$$
$$= \frac{\prod_j f_{k_j} \; (s_j)}{f_k(n)} \, .$$

In particular, the probability of the aliased-alleles data above, in their listed order, is

$$[f_1(1)]^{10} \, [f_1(2)]^2 f_2(2) \; [f_1(3)]^6 f_2(3)$$
$$f_1(5) \; f_2(5) \; [f_2(6)]^2 f_1(8) \; f_2(11) \; f_5(68)/f_{37} \; (146)$$
$$= (w_{1,1})^{10} \, (w_{1,2})^2 \, w_{2,2} \, (w_{1,3})^6 \, w_{2,3} \, w_{1,5} \, w_{2,5}$$
$$(w_{2,6})^2 \, w_{1,8} \, w_{2,11} \, w_{5,68} \, / \, w_{37,146}$$
$$= 3.547 \times 10^{-23} \, ,$$

where, in EWEN's (1972) notation,

$$w_{k,n} = k! \, |S_n^{(k)}| / n! \; .$$

However, it is not clear how neutrality should be tested, because it is an open question as to which possible samples would be "more extreme" than the observed one. The difficulty may be overcome eventually by the complete sequencing of all genes so that their alleles will be fully distinguished.

Averages of sample homozygosities have often been used as statistics for estimating $\Theta = 4N_e u$, where N_e is the effective population size and u is the mutation rate per gene per generation. For instance, SINGH, LEWONTIN and FELTON (1976, p. 625) estimated $\Theta/4$ as 0.68 on the basis of a maximum heterozygosity of 0.73 (homozygosity 0.27). Assuming neutrality, the *population* homozygosity, F, has an expected value $1/(1+\Theta)$ so that an intuitively reasonable estimate of Θ would be F^{-1}—1. Using the *sample* homozygosity, $\hat{F}$, suggests an estimate $\tilde{\Theta} = \hat{F}^{-1}$—1, but this has been shown, by KIRBY (1975), to have a much greater mean-square error as an estimate of Θ than does the maximum likelihood estimate, $\hat{\Theta}$, based on k and n alone (see EWENS 1972). Moreover, under neutrality, $\hat{F}$ does not have mean $1/(1+\Theta)$, but rather $n^{-1} + (1 - n^{-1})/(1+\Theta)$, see NEI and ROYCHOUD-

HURY (1974) and WATTERSON (1977, 4.3.1), and this suggests the use of the estimators $\tilde{\tilde{\Theta}} = n(1-\hat{F})/(n\hat{F}-1)$, which is greater than $\tilde{\Theta}$ but will have a similarly large variance.

For the twelve samples analyzed by SINGH, LEWONTIN and FELTON (using their criteria 1–4, but not heat sensitivity) we find the following estimates:

Population	SZ	SS	CN	SC	WR	CH	CE	MV	HR	AU	GU	BO
$\hat{\Theta}$	11.65	1.32	1.14	4.90	3.43	6.22	1.32	7.88	4.11	4.06	0.88	2.68
$\tilde{\Theta}$	5.26	0.65	0.59	3.36	1.72	3.59	0.65	5.26	1.67	2.60	1.00	2.13
$\tilde{\tilde{\Theta}}$	10.14	0.72	0.71	4.87	2.14	5.18	0.72	10.14	2.14	6.50	2.00	2.88

In the neutral case, all three estimatorss are biased, $\hat{\Theta}$ and $\tilde{\tilde{\Theta}}$ being slightly too large on the average and $\tilde{\Theta}$ too small. Their biases are also subject to any selective influences. As pointed out to me by EWENS, the homozygosity tests of neutrality used above are, in effect, tests for the compatibility of $\hat{\Theta}$ and $\tilde{\tilde{\Theta}}$. Preliminary calculations suggest that $E(\tilde{\tilde{\Theta}})$ is more reduced by deleterious alleles than is $E(\hat{\Theta})$ when Θ and/or selection increase. Further work is needed to see whether the substantially smaller mean-square error of $\hat{\Theta}$ compared with $\tilde{\tilde{\Theta}}$ (or $\tilde{\Theta}$) is maintained from the neutral case into non-neutral cases.

One clear advantage $\hat{\Theta}$ has over $\tilde{\Theta}$ and $\tilde{\tilde{\Theta}}$ is that $\hat{\Theta}$ requires knowledge of only the number of alleles in the sample together with the sample size, whereas $\tilde{\Theta}$ and $\tilde{\tilde{\Theta}}$ rely on knowing also the allele frequencies. In the pooled sample studied by SINGH, LEWONTIN and FELTON, there were 37 alleles distinguished in the 146 genes using the five criteria; had the sample been from a single population, we could calculate $\hat{\Theta} = 15.65$, whereas, not knowing $\hat{F}$ exactly, we could not calculate $\tilde{\Theta}$ or $\tilde{\tilde{\Theta}}$.

I cannot see much point in averaging the sets of twelve estimates $\hat{\Theta}$ (or $\tilde{\Theta}$ or $\tilde{\tilde{\Theta}}$) or estimating a parameter Θ by first averaging the twelve homozygosities. There seems to be no reason why the effective sizes of twelve populations and hence their Θ values should be equal. Had the samples been taken from *one* population at one time, it would be better to pool them first and treat them as a single large sample.

I thank PROFESSOR W. J. EWENS for much advice in the analysis of these data, and MRS. M. WU for help with the computing. I especially thank PROFESSOR R. LEWONTIN for sending me his analysis of the data. Thanks are also due to the referee for some helpful comments.

LITERATURE CITED

ABRAMOWITZ, M. and I. A. STEGUN, 1965 *Handbook of Mathematical Functions.* Dover, New York.

COYNE, J. A., 1976 Lack of genic similarity between two sibling species of Drosophila as revealed by varied techniques. Genetics **84:** 593–607. ——, 1977 Corrigenda, Genetics **85,** No. 3.

EWENS, W. J., 1972 The sampling theory of selectively neutral alleles. Theor. Pop. Biol. **3**: 87–112. ——, 1973 Testing for increased mutation rate for neutral alleles. Theor. Pop. Biol. **4**: 251–258.

KARLIN, S. and J. MCGREGOR, 1972 Addendum to a paper of W. EWENS. Theor. Pop. Biol. **3**: 113–116.

KELLY, F., 1977 Exact results for the Moran neutral allele model. Adv. Appl. Prob. (In press).

KINGMAN, J. F. C., 1977 The population structure associated with Ewens' sampling formula. Theor. Pop. Biol. **11**: 274–283.

KIRBY, K., 1975 A discussion of simulation results for various aspects of the neutral alleles model. Theor. Pop. Biol. **7**: 277–287.

NEI, M. and A. K. ROYCHOUDHURY, 1974 Sampling variances of heterozygosity and genetic distance. Genetics **76**: 379–390.

SINGH, R. S., R. C. LEWONTIN and A. A. FELTON, 1976 Genetic heterogeneity within electrophoretic "alleles" of xanthine dehydrogenase in *Drosophila pseudoobscura*. Genetics **84**: 609–629. ——, 1977 Corrigenda, Genetics **85**, No. 3.

WATTERSON, G. A., 1974a The sampling theory of selectively neutral alleles. Adv. Appl. Prob. **6**: 463–488. ——, 1974b Models for the logarithmic species abundance distributions. Theor. Pop. Biol. **6**: 217–250. ——, 1976 The stationary distribution of the infinitely-many neutral alleles diffusion model. J. Appl. Prob. **13**: 639–651. ——, 1977 Heterosis on neutrality? Genetics **85**: 789–814. ——, 1978 The homozygosity test of neutrality. Genetics (In press).

WATTERSON, G. A. and H. A. GUESS, 1977 Is the most frequent allele the oldest? Theor. Pop. Biol. **11**: 141–160.

Corresponding editor: D. L. HARTL