The estimation of genotypic probabilities in an adult population by the analysis of descendants

ANTONIO BARBADILLA*, HORACIO NAVEIRA†, ALFREDO RUIZ AND MAURO SANTOS

Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain

(*Received 22 July 1991 and in revised form 2 October 1991*)

Summary

There are instances, the most typical being inversion polymorphism in *Drosophila*, where the genotype is not directly accessible in the adult organism, but can be observed in young life-stages. In these cases, if we want to estimate genotypic probabilities in adult populations, we must examine an offspring sample from adults. In this paper we derive the maximum likelihood estimators, and their errors, for genotypic probabilities in an adult population, according to a standard protocol in which collected parents of a random sample are individually crossed with individuals of a laboratory stock with known homozygous genotype, and a fixed number of their offspring is genetically examined in young life-stages. Arnold's probabilistic model for one locus with two alleles is developed for our estimates. An optimum design which generates a minimum variance is proposed, consisting of examining a moderate offspring number (3–4) per parent. Finally, we propose maximum likelihood estimates when several samples with different numbers of parents per sample, and/or examined progeny per parent are obtained.

1. Introduction

One of the descriptive parameters of a genetic polymorphism in a Mendelian population is the relative frequency or probability of a genotype ($p(A_i A_j)$). Although this descriptor is often used with the aim of estimating genic frequencies ($p(A_i)$), it contains *per se* other interesting information too. So, if we compare it with other descriptors we can obtain evidence of (*a*) the structure of the population (whether it is in Hardy–Weinberg equilibrium, HWE, or not, e.g. Selander, 1970), (*b*) the action of natural selection (Dobzhansky & Levene, 1948; Christiansen & Frydenberg, 1973; Stalker, 1976; Barbadilla *et al.* 1992) or (*c*) the association with characters other than fitness (White & Andrew, 1960, 1962; Prevosti, 1966; Krimbas & Loukas, 1980; Ruiz *et al.* 1991).

When genotypes can be determined directly in adults, the maximum likelihood estimators (MLEs) of their different probabilities are inferred from the counts or observed frequencies of each genotype (Weir, 1990). There are cases, though, where genotypes are cryptic in adult organisms but visible in young life-stages, being then necessary to examine an offspring sample to estimate their probabilities. A typical example is inversion polymorphism in *Drosophila*, which is accessible only in third-instar larvae. There are many other examples of missing parental data in other organisms, usually when dealing with allozyme loci (Luykx, 1981; Samollow *et al.* 1983; and references in Arnold & Morrison, 1985). In cases like these, the estimation of genotypic and genic frequencies in adult populations can be made according to two different experimental protocols: (1) A random sample of inseminated females is taken in an adult population and brought to the laboratory, where *n* individuals of each offspring are examined and used to estimate the parental genotypes. (2) Males (females) are collected at random in an adult population and are individually crossed with females (males) of a laboratory stock with a known homozygous genotype. The genotypic and/or genic probabilities in the adult population are then estimated through the examination of *n* descendants of each cross.

Arnold (1981), Arnold & Morrison (1985), and Sobel *et al.* (1986) have investigated the properties of different estimators of genic probabilities for both

[*] Corresponding author.

[†] Departamento de Biología Fundamental, Unidad de Genética, Facultad de Biología, Universidad de Santiago de Compostela, Santiago de Compostela, Spain

https://doi.org/10.1017/S0016672300030342 Published online by Cambridge University Press

protocols. Their estimators are obtained from a model that assumes HWE and Mendelian segregation. Barbadilla & Naveira (1988) have developed MLEs, together with their errors, of the probabilities of mated genotypic pairs ($p(A_i A_j, A_k A_l)$, where $A_i A_j$ is a parental genotype and $A_k A_l$ the other one) for protocol 1. The model in this case was dependent on two assumptions: Mendelian segregation and single female insemination (or a high degree of sperm predominance). It should be pointed out that estimates of genotypic frequencies by each protocol do not correspond exactly to the same adult population. Those from the first protocol are genotypic frequencies in the mated population [$p(A_i A_j \,|\, \text{mated population})$], whereas those from the second one are genotypic frequencies in the adult base population [$p(A_i A_j \,|\, \text{base population})$]. Thus, if we want to estimate the frequencies in the adult (base) population without any additional assumption (such as, for example, that there is neither sexual selection, nor panmixia), or if we are studying sexual selection, then only protocol 2 is suitable.

The aim of this work is to obtain estimators with optimum statistical properties for the different genotypic probabilities in an adult natural population, when protocol 2 is followed. The usual procedure in earlier estimates consisted of determining the parental genotype through the analysis of 6 or more descendants per parent (Dobzhansky, 1944; Dobzhansky & Levene, 1948; Stalker, 1976; Salceda & Anderson, 1988; Ruiz *et al.* 1991). Here, we propose MLEs together with their variances, for any fixed number of examined offspring ($n \geqslant 2$). In addition, we have studied the experimental design that generates a minimum variance (or maximum information) for a given experimental effort. It is shown that the best strategy does not consist of determining the parental genotype but of estimating it through a moderate number of descendants (3–4). Lastly, we also propose ML estimates when several samples with different numbers of parents per sample and/or examined progeny per parent are available.

Our probabilistic model is an extension to k alleles of a one locus–two alleles Arnold's model (Arnold, 1981). HWE is not assumed, the only assumption being Mendelian segregation, i.e. the absence of any process that may produce a deviation from the 1:1 ratio in the analysed offspring of a heterozygous parent. A test for this assumption will be shown later.

2. Probabilistic model

(i) *One locus with two alleles*

Consider one locus with two alleles (A and a). A collected individual carries y copies of the allele a ($y = 0, 1, 2$). The analysis of a fixed number n of its offspring ($n \geqslant 2$) yields a profile $\underline{n} = (x, n-x)$ where x and $n-x$ are the respective numbers of a and A alleles inherited from the parent. Family profiles are defined in a sample space $\Lambda = \{\underline{n}: 0 \leqslant x \leqslant n\}$. Three mutually exclusive events can be obtained, which are shown in the first column of Table 1. These events constitute a partition of the different possible family profiles n in the sample space Λ. This partition identifies family profiles by the presence or absence of an allele in the family. The probabilities K_0, K_1, K_2 of each of these three events can be computed from the model specification

$$\text{prob}(\underline{n} \,|\, \Theta_0 \Theta_1, \Theta_2) = \sum_{y=0}^{2} \text{prob}(\underline{n} \,|\, y) \quad \text{prob}(y \,|\, \Theta_0, \Theta_1, \Theta_2),$$

and are shown in the second column of Table 1. $\Theta_0, \Theta_1, \Theta_2$ are the probabilities of the different genotypes in the base population. α is the probability, according to Mendelian segregation, that a heterozygous parent is assigned correctly after the analysis of n descendants.

Our problem is to infer $\Theta_0, \Theta_1, \Theta_2$, the probabilities of the different genotypes. In a collection of N parents, the numbers N_0, N_1, N_2 are the counts of the different types of family profiles, based on the presence or absence of one or other allele. The probability density of $\underline{N} = (N_0, N_1, N_2)$ is multinomial, and is given by

$$\text{prob}(\underline{N} \,|\, \Theta_0, \Theta_1, \Theta_2) = \binom{N}{\underline{N}} K_0^{N_0} K_1^{N_1} K_2^{N_2},$$

where

$$\binom{N}{\underline{N}} = N!/N_0!N_1!N_2!$$

For a random sample from the base population, the list of counts $N = (N_0, N_1, N_2)$ is a sufficient statistic for the estimation of the probabilities of the events K_0, K_1, K_2, that is to say, $\underline{N}$ contains all the available information of the sample for the estimation of the probabilities. Since the number of independent parameters of the model is equal to the number of degrees of freedom, then we can find the MLEs for the parameters of the model equating the observations with their expected values from the model specification of Table 1 (Bailey, 1951), as follows:

$$\left.\begin{aligned} \hat{\Theta}_0 &= N_0/N - (\tfrac{1}{2})^n \hat{\Theta}_1 \\ \hat{\Theta}_1 &= N_1/N\alpha \\ \hat{\Theta}_2 &= 1 - (\hat{\Theta}_0 + \hat{\Theta}_1). \end{aligned}\right\} \qquad (1)$$

These estimators (denoted by a circumflex) are consistent, unbiased and of minimum variance. The asymptotic variances and covariances are given in the Appendix.

An optimum experimental design is that which produces maximum information per unit of observation cost. If we fix a given experimental effort T ($T = Nn$ = total number of examined offspring; $N =$

https://doi.org/10.1017/S0016672300030342 Published online by Cambridge University Press

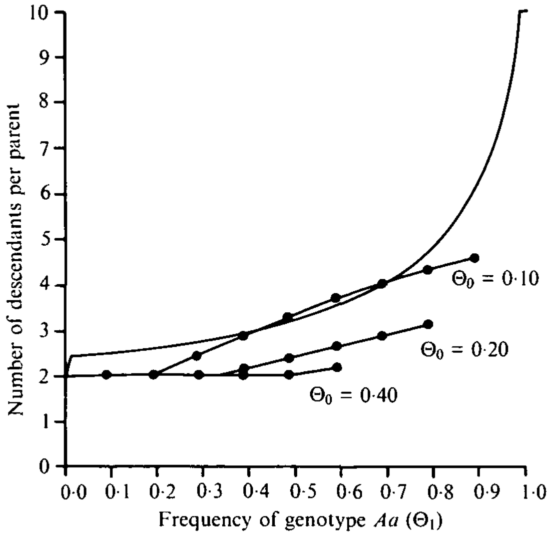

Fig. 1. Optimum distribution for a given experimental effort ($T = Nn$) between N (number of collected parents) and n (number of examined descendants per parent) which minimizes the variance of the estimators $\hat{\Theta}_1$ (——) and $\hat{\Theta}_0$ (—●—). The variance of $\hat{\Theta}_1$ is a function of Θ_1. The variance of $\hat{\Theta}_0$ depends on both Θ_0 and Θ_1.

number of collected parents; n = number of examined offspring per parent), it is possible to find the optimum distribution of effort between N and n, that is, the value of n and $N = T/n$ that produces a minimum (maximum) variance (information) of the estimator of a genotypic probability. In Fig. 1 the values of n that minimize the variance of each one of the estimators of the independent parameters ($\hat{\Theta}_0$ and $\hat{\Theta}_1$) are graphed. The variance of $\hat{\Theta}_1$ (single line) is a function of Θ_1, while that of $\hat{\Theta}_0$ (lines with solid symbols) depends on both Θ_1 and Θ_0. As the graph shows, unless Θ_1 is high ($> 0{\cdot}7$), the best strategy to obtain the maximum information from the sample consists of analysing 3–4 descendants per collected parent. This result means that if we want to estimate genotypic probabilities in the base population, then it is better to examine few descendants from each parent, but a large number of parents, than to examine a large offspring number from a reduced sample of parents. The increase in efficiency may be substantial. For example, if $\Theta_1 = 0{\cdot}5$ and we apply an effort $T = Nn$ of 1001 examined descendants, distributed to $N = 143$ parents, and $n = 7$ descendants per parent, then the variance of the estimate of Θ_1 would be the same as in a second design with $T = 712$ descendants, distributed to $N = 178$ parents and $n = 4$ descendants per parent. This latter design implies a saving of 289 descendants, that is, 29% of T for the first design!

(ii) *One locus with k alleles*

The extension of the estimators of genotypic probabilities to more than two alleles is straightforward. Let $\hat{\Theta}_{ii}$ be the probability of homozygote $A_i A_i$ and $\hat{\Theta}_{ij}$ ($i \neq j$) that of heterozygote $A_i A_j$. Following the same reasoning as before, the optimum estimators of the probability of the different genotypes are:

$$\hat{\Theta}_{ii} = N_{ii}/N - (\tfrac{1}{2})^n \hat{\Theta}_{i.}$$

$$\hat{\Theta}_{ij} = N_{ij}/N\alpha,$$

where N_{ii} and N_{ij} are the counts (observations) of the homozygous $A_i A_i$ and heterozygous $A_i A_j$ genotypes, respectively, and $\Theta_{i.} = \sum_{j=1, i \neq j}^{k} \hat{\Theta}_{ij}^{i}$ is the estimate of the frequency of all heterozygotes carrying the allele i.

Variances of the estimators are given in the Appendix. As for the case of two alleles, unless $\Theta_{i.}$ is high, the best strategy to minimize variances is to analyse 3–4 descendants per collected parent.

(iii) *Several samples with different N and n*

Frequently we have several independent samples from different collections, or only one sample with different subsamples differing in the number of parents (N) and/or in the number of examined offspring by parent (n). The probability density of $\underline{N} = (N_{ii,l}, N_{ij,l})$, where l refers to the sample and ii and ij to homozygous and heterozygous genotypes respectively, is multinomial. MLEs must be found from the likelihood function. The log likelihood of density function is, for k alleles,

$$\ln L = \sum_{l=1}^{m} \sum_{i=1}^{k} N_{ii,l} \ln (\hat{\Theta}_{ii,l} + (\tfrac{1}{2})^{n_l} \hat{\Theta}_{i..l})$$
$$+ \sum_{l=1}^{m} \sum_{i=2}^{k} \sum_{j=1}^{i-1} N_{ij,l} \ln (\alpha_l \hat{\Theta}_{ij,l}), \qquad (2)$$

where m is the number of samples and n_l and α_l the number of examined descendants and the probability of a correct assignation of a heterozygous individual in the sample l, respectively. No simple algebraic expressions to find the MLEs and their errors exist, but numerical maximisation is straightforward (see Section 3).

The values of $\hat{\Theta}_{ii,l}$ and $\hat{\Theta}_{ij,l}$ for each sample may not be estimating the same parameters, Θ_{ii} and Θ_{ij}. We can test for this, computing the value of the likelihood ratio test statistic or G statistic (Sokal & Rohlf, 1981) given by the expression

$$G = 2\left[\sum_{l=1}^{m} \sum_{i=1}^{k} N_{ii,l} \ln \frac{N_{ii,l}}{N_l(\hat{\Theta}_{ii} + (\tfrac{1}{2})^{n_l} \hat{\Theta}_{i.})}\right.$$
$$\left. + \sum_{l=1}^{m} \sum_{i=2}^{k} \sum_{j=1}^{i-1} N_{ij,l} \ln \frac{N_{ij,l}}{N_l \alpha_l \hat{\Theta}_{ij}}\right]. \qquad (3)$$

This value can be compared with a χ^2-distribution with $(m-1)(\tfrac{1}{2}k(k+1)-1)$ degrees of freedom. If the G value is significant, then this indicates that the differences among samples for $\Theta_{ii,l}$ and $\Theta_{ij,l}$ are too large to be attributed to sampling errors.

(iv) *Estimation of gene frequencies*

If we define the gene frequency of allele A_i as a function of genotype frequencies, $\Theta_i = \Theta_{ii} + (\tfrac{1}{2})\Theta_{i.}$,

https://doi.org/10.1017/S0016672300030342 Published online by Cambridge University Press

Table 1. *Specification of the probabilistic model*

Event	Probability	Count (N)	Score	Example
$\Lambda_0 = \{\underline{n}: x = 0\}$	$K_0 = \Theta_0 + (\frac{1}{2})^n \Theta_1$	N_0	0	*AA*, *Aa*
$\Lambda_1 = \{\underline{n}: 0 < x < n\}$	$K_1 = \alpha\Theta_1$	N_1	1	*Aa*
$\Lambda_2 = \{\underline{n}: x = n\}$	$K_2 = \Theta_2 + (\frac{1}{2})^n \Theta_1$	N_2	2	*aa*, *Aa*

Θ_0, Θ_1, Θ_2: probabilities of the different genotypes. For a locus with two alleles, 0 = only one allele, 1 = both alleles, 2 = only the other allele. $\alpha = [1-(\frac{1}{2})^{n-1}]$.

then we can estimate Θ_i from the estimators of Θ_{ii} and Θ_{ij}. The MLE is

$$\hat{\Theta}_i = N_{ii}/N + N_{i.}/2N, \tag{4a}$$

where $N_{i.} = \sum_{j=1, i \neq j}^{k} N_{ij}$. This equation is 'Dobzhansky's estimator' (Dobzhansky & Epling, 1944; Arnold, 1981) or 'gene counting' estimator. Its variance is

$$\mathrm{Var}(\hat{\Theta}_i) = \mathrm{Var}(\hat{\Theta}_{ii}) + (\tfrac{1}{4})\mathrm{Var}(\hat{\Theta}_{i.}) + \mathrm{Cov}(\hat{\Theta}_{ii}, \hat{\Theta}_{i.}). \tag{4b}$$

Arnold (1981) has investigated the properties of this estimator under the assumption of HWE. In this case, the variance simplifies to:

$$\mathrm{Var}(\hat{\Theta}_i) = \frac{\Theta_i(1-\Theta_i)}{2N}(1+(\tfrac{1}{2})^{n-1}).$$

When HWE is assumed, the Dobzhansky's estimator is not the MLE of the gene frequency ($\hat{\Theta}_{ML}$). The likelihood equation of $\hat{\Theta}_{ML}$ is given by Arnold (1981) in his result 5. Arnold shows, however, that the Dobzhansky's estimator is highly efficient and easier to compute than $\hat{\Theta}_{ML}$, hence it can alternatively be used as valid approximation.

(v) *Testing for HWE*

The likelihood ratio test can also be used to test HWE. The standard procedure consists of comparing the observed number of the different genotypes with the expected values according to HWE. But in our case, the observed number, N_y, $y = 0, 1, 2$ for two alleles, is not, in general, the actual number of each genotype in the sample, due to the error of misdiagnosis. Thus, the expected number to be compared with the observed number should include both HWE and probability of misdiagnosis. The likelihood ratio test statistic is

$$G = \sum_{y=0}^{2} Ny \ln \frac{N_y}{N\hat{K}_y}, \tag{5a}$$

where $\hat{K}_y$ are computed from equations of second column in Table 1. In these equations, the Θ_y values are estimated from the MLE of Θ ($\hat{\Theta}_{ML}$), the gene frequency of allele A, according to HWE. But, as has been pointed out, Θ can be approximated from (4). If genotypic frequencies are in HWE, the G statistic will be χ^2-distributed with 1 D.F.

When we have several samples with different number of examined offspring per parent, the statistic is

$$G = \sum_{l=1}^{m} \sum_{y=0}^{2} N_{ly} \ln \frac{N_{ly}}{N_l \hat{K}_{ly}}, \tag{5b}$$

where l refers to the sample and y to the genotype. G is χ^2-distributed with $2m-1$ D.F. when HWE holds. For k alleles, the statistic would be the same, but changing the limits of values of y, ranging from 1 to $k(k+1)/2$ (the number of karyotypes). The degrees of freedom of the χ^2-distribution would be $m[\frac{1}{2}(k+1)-1]-(k-1)$.

3. An illustration

Our research team is carrying out an exhaustive study of the inversion polymorphism of the species *Drosophila buzzatii* in a natural population from Carboneras (Almería, Spain). Ruiz *et al.* (1986) give a detailed description of the population. In this population *D. buzzatii* is polymorphic for 2 out of 4 (excluding the dot chromosome) autosomes (this polymorphism has been described by Ruiz *et al.* 1984). In July 1990 a random sample of males and females was collected. A total of 314 males was individually crossed with two females of a homokaryotypic control stock and 2 larvae in each offspring family were examined. Among the collected females, 173 of them were crossed with two males of the same stock, and 8 larvae were examined in each offspring to determine the maternal karyotype. The control stock has an inversion (*5I*) that is not present in natural populations. It appeared in a genetically unstable line produced by introgressive hybridization (*In*(5)*F2b*; *F2e*, Naveira & Fontdevila, 1985). This inversion was used as a chromosome marker, to be sure that the progeny of wild females were fathered by the laboratory stock males. Previously, we showed the existence of a high degree of sperm predominance of the last mated male in other laboratory stocks of the same species (Barbadilla *et al.* 1991).

Columns 1 and 3 in Table 2 show the counts of each type of family profile for males and females, respectively, for the polymorphism of chromosome 4. This polymorphism is due to two inversions, 4*st* (*standard*) and 4*s*. From these data we want to obtain

https://doi.org/10.1017/S0016672300030342 Published online by Cambridge University Press

Table 2. *List of counts* (N_y, $y = 0, 1, 2$) *and estimates of genotypic probabilities* (Θ_y) *from the collected sample of adults in Carboneras* (*Almería, Spain*) *in July, 1990.* $\Theta_0 = p(4st/st)$, $\Theta_1 = p(4st/s)$, $\Theta_2 = p(4s/s)$

Males		Females		Total
$N_0 = 207$	$\hat{\Theta}_0 = 0{\cdot}560 \pm 0{\cdot}069$	$N_0 = 111$	$\hat{\Theta}_0 = 0{\cdot}640 \pm 0{\cdot}072$	$\hat{\Theta}_{0.} = 0{\cdot}598 \pm 0{\cdot}049$
$N_1 = 62$	$\hat{\Theta}_1 = 0{\cdot}395 \pm 0{\cdot}088$	$N_1 = 53$	$\hat{\Theta}_1 = 0{\cdot}309 \pm 0{\cdot}069$	$\hat{\Theta}_{1.} = 0{\cdot}349 \pm 0{\cdot}052$
$N_2 = 45$	$\hat{\Theta}_2 = 0{\cdot}045 \pm 0{\cdot}04$	$N_2 = 9$	$\hat{\Theta}_2 = 0{\cdot}051 \pm 0{\cdot}033$	$\hat{\Theta}_{2.} = 0{\cdot}053 \pm 0{\cdot}027$
$N = 314$		$N = 173$		

Data from Barbadilla, Ruiz, Santos & Fontdevila (unpublished).

the following information: (*a*) an estimation of karyotypic frequencies in males and females in the base population, (*b*) a test of the differences between males and females in these frequencies, (*c*) if differences between sexes are not significant, an estimation of karyotypic frequencies in the whole population, irrespectively of sex, (*d*) an estimation of inversion frequencies, and a test of whether the population is in HWE, (*e*) finally, we would like to know whether the assumption of Mendelian segregation in the offspring, is actually fulfilled in the progeny of heterozygous females.

Columns 2 and 4 in Table 2 show the estimates ± their 95 % confidence semi-interval ($\pm 1{\cdot}96\sqrt{\text{variance}}$) of the probabilities of different karyotypes. We have applied formulae (1) and (A 1). The value of the G statistic computed from (3) to compare both samples was not significant, $G = 2{\cdot}594$, 2 D.F., $P = 0{\cdot}228$. Hence we may accept the null hypothesis that there are no differences in frequencies between sexes.

The MLEs for genotypic frequencies in the whole population were computed with the help of LE program from the BMDP statistical software (Dixon, 1990). This program obtains ML estimates that maximize the likelihood function, using the iterative Newton–Raphson algorithm. The program also estimates the asymptotic standard errors and the correlation matrix of the estimated parameters. Given the multinomial distribution of our data, we have followed the example LE.4 (pp. 730–2). The program needs to use an initial estimate (initial guess) for $\hat{\Theta}_0$ and $\hat{\Theta}_1$ (the independent parameters). The simple average from both samples, $\hat{\Theta}_0 = 0{\cdot}600$ and $\hat{\Theta}_1 = 0{\cdot}352$, were chosen. In column 5, the ML estimates for the whole population together with their 95 % confidence semi-interval are given.

The estimate of *st* inversion frequency ($\pm 1{\cdot}96\sqrt{\text{variance}}$) from (4) was $\hat{\Theta} = 0{\cdot}773 \pm 0{\cdot}030$. The G statistic value [from equation (5*b*)] to test HWE was $G = 2{\cdot}607$, 3 D.F., $P = 0{\cdot}456$. So, we can infer that karyotype frequencies are in HWE.

Finally, we test the assumption of Mendelian segregation. If the number of examined progeny per parent is $\geqslant 3$, then it is always possible to test whether the alleles of a heterozygous individual appear in the same proportion in the progeny (1:1). For 424 inversions coming from 53 heterozygous females ($53 \times 8 = 424$), 211 inversion are *st* and 213 *s*. Therefore, the null hypothesis of Mendelian segregations is accepted ($\chi^2 = 0{\cdot}094$, 1 D.F., $P = 0{\cdot}759$).

4. Discussion

The probabilistic model presented in this paper has led to MLEs of genotypic probabilities, together with their variances. Estimates for a heterogeneous set of samples have also been proposed. All these estimates do not assume HWE. Therefore, any deviation from expected frequencies can be tested, and evidence of selection or information on the structure of the population be obtained.

We have also seen that the examination of 3–4 descendants per parent is the design that produces most information on the parameters to be estimated, for a given experimental effort. In our initial model n was fixed. A uniformly most efficient design would be to stop examining the progeny of an individual whenever a definitive combination is obtained (a combination that allows an unequivocal identification of the parental karyotype), instead of examining all the n-prefixed descendants (Sobel *et al.* 1986). In this case, the average saving in the number of larvae per individual ($E(SL)$) is:

$$E(SL) = E(n) - E^*(n),$$

where $E(n) = n$ is the expected number of examined larvae per individual when the prefixed number is actually examined, and $E^*(n)$ is the expectation when we stop as soon as we get a definitive combination. This last expectation is given by the equation:

$$E^*(n) = [\Theta_0 + \Theta_2 + (\tfrac{1}{2})^{n-1}\Theta_1]\,n + \Theta_1 \sum_{i=2}^{n} i(\tfrac{1}{2})^{i-1}.$$

With $\Theta_1 = 0{\cdot}25$, $0{\cdot}50$, or $0{\cdot}75$, $E^*(n)$s and $E(SL)$s (within parentheses) for $n = 4$, and $n = 8$ are as follows

	$n = 4$ $E^*(n)\,(E(SL))$	$n = 8$ $E^*(n)\,(E(SL))$
$\Theta_1 = 0{\cdot}25$	3·687 (0·312)	6·746 (1·254)
$\Theta_1 = 0{\cdot}50$	3·375 (0·625)	5·492 (2·508)
$\Theta_1 = 0{\cdot}75$	3·062 (0·937)	4·238 (3·762)

https://doi.org/10.1017/S0016672300030342 Published online by Cambridge University Press

Although the strategy of stopping the analysis as soon as we find a definitive combination is more efficient, analysing all the prefixed progeny could be interesting if we want to test the hypothesis of Mendelian segregation. If no evidence on this aspect is available, it would be more advisable to examine all the prefixed progeny, at least for the first time.

One of the processes that may produce a deviation of the ratio 1:1 in the analysed offspring is viability differences among the carriers of different inversions under laboratory conditions (Stalker, 1976). If there were a deviation of Mendelian segregation and its magnitude were known, then it could easily be introduced into the model, substituting in equation (1) the $1/2$ ratio for both alleles, A and a, by p and $1-p$ respectively, where p is the proportion of allele A in the progeny of a heterozygous parent. The estimator of p and its error may be derived from the estimators for incomplete binomial distributions given by Mantel (1951) and Li (1970).

It is a pleasure to acknowledge E. Betrán, J. E. Quezada-Díaz, F. Peris and an anonymous referee for their helpful comments on this MS. A. B. is grateful to the Comissió Interdepartamental de Recerca i Innovació Tecnològica (CIRIT), Generalitat de Catalunya, Catalonia, Spain, for providing funds for research equipment. This work was supported by grant no. PB85-0071 to A.F. from the Comisión Interministerial de la Ciencia y la Tecnología (CICYT), Spain.

References

Arnold, J. (1981). Statistics of natural populations. I. estimating an allele probability in cryptic fathers with a fixed number of offspring. *Biometrics* **37**, 495–504.

Arnold, J. & Morrison, M. L. (1985). Statistics of natural populations. II. Estimating an allele probability in families descended from cryptic mothers. *Genetics* **109**, 785–798.

Bailey, N. T. J. (1951). Testing the solubility of maximum likelihood equations in the routine application of scoring methods. *Biometrics* **7**, 268–274.

Barbadilla, A. & Naveira, H. (1988). The estimation of parental genotypes by the analysis of a fixed number of their offspring. *Genetics* **119**, 465–472.

Barbadilla, A., Quezada-Díaz, J. E., Ruiz, A., Santos, M. & Fontdevila, A. (1991). The natural history of *Drosophila buzzatii*. XVII. Double mating and sperm predominance. *Genetics, Selection, Evolution* **23**, 133–140.

Barbadilla, A., Ruiz, A., Santos, M. & Fontdevila, A. (1992). Analysis of adult selection components associated with inversion polymorphism in a natural population of *Drosophila buzzatii* (submitted).

Christiansen, F. B. & Frydenberg, O. (1973). Selection components analysis of natural polymorphisms using population samples including mother-child combinations. *Theoretical Population Biology* **4**, 425–445.

Dixon, W. J. (ed.). *BMDP Statistical Software*, pp. 721–738. Manual University of California Press, California.

Dobzhansky, Th. (1944). Chromosomes races in *Drosophila pseudoobscura* and *Drosophila persimilis*, pp. 47–114. Publication No. 554. Washington, D.C.: Carnegie Institution.

Dobzhansky, Th. & Epling, C. (1944). Contributions to the Genetics, Taxonomy, and Ecology of *Drosophila pseudoobscura* and its Relatives. Publication No. 554. Washington, D.C.: Carnegie Institution.

Dobzhansky, Th. & Levene, H. (1948). The genetics of natural populations. XVII. Proof of operation of natural selection in wild populations of *Drosophila pseudoobscura*. *Genetics* **33**, 537–547.

Krimbas, C. B. & Loukas, M. (1980). The inversion polymorphism of *Drosophila subobscura*. *Evolutionary Biology* **12**, 402–416.

Li, C. C. (1970). The incomplete binomial distribution. *Mathematical Topics in Population Genetics* (ed. K. Kojima), pp. 337–366. Berlin: Springer-Verlag.

Luykx, P. (1981). A sex-linked esterase locus and translocation heterozygosity in a termite. *Heredity* **46**, 315–320.

Manly, B. F. J. (1985). *The Statistics of Natural Selection*. London: Chapman and Hall.

Mantel, N. (1951). Evaluation of a class of diagnosis tests. *Biometrics* **7**, 240–246.

Naveira, H. & Fontdevila, A. (1985). The evolutionary history of *Drosophila buzzatii*. IX. High frequencies of new chromosome rearrangements induced by introgressive hybridization. *Chromosoma* **91**, 87–94.

Prevosti, A. (1966). Inversion heterozygosity and size in a natural population of *Drosophila subobscura*. *Mutation in Population, Proceedings of the Symposium on the Mutational Process*, pp. 49–53. Praha: Publishing House of the Czechoslovak Academy of Sciences.

Ruiz, A., Naveira, H. & Fontdevila, A. (1984). La historia evolutiva de *Drosophila buzzatii*. IV. Aspectos citogenéticos de su polimorfismo cromosómico. *Genética Ibérica* **36**, 13–35.

Ruiz, A., Fontdevila, A., Santos, M., Seoane, M. & Torroja, E. (1986). The evolutionary history of *D. buzzatii*. VIII. Evidence for endocyclic selection acting on the inversion polymorphism in a natural population. *Evolution* **40**, 740–755.

Ruiz, A., Santos, M., Barbadilla, A., Quezada-Díaz, J. E., Hasson, E. & Fontdevila, A. (1991). Genetic variance for body size in a natural population of *Drosophila buzzatii*. *Genetics* **128**, 739–750.

Salceda, V. M. & Anderson, W. W. (1988). Rare male mating advantage in a natural population of *Drosophila pseudoobscura*. *Proceedings of the National Academy of Sciences, U.S.A.* **85**, 9870–9874.

Samollow, P. B., Dawson, P. S. & Riddle, R. A. (1983). X-linked and autosomal inheritance patterns of homologous genes in two species of *Tribolium*. *Biochemical Genetics* **21**, 167–176.

Selander, R. K. (1970). Behavior and genetic variation in natural population (*Mus musculus*). *American Zoologist* **10**, 53–66.

Sobel, M. J., Arnold, J. & Sobel, M. (1986). Statistics of natural populations. III. Sequential sampling plans for the estimation of gene frequencies. *Biometrics* **42**, 45–65.

Sokal, R. R. & Rohlf, F. J. (1981). *Biometry*, 2nd edn. San Francisco: Freeman.

Stalker, H. D. (1976). Chromosome studies in wild population of *Drosophila melanogaster*. II. Relationship of inversion frequencies to latitude, season, wing-loading and flight activity. *Genetics* **95**, 211–223.

Weir, B. S. (1990). *Genetic Data Analysis*. Sunderland: Sinauer Associates.

White, M. J. D. & Andrew, L. E. (1960). Cytogenetics of the grasshopper *Moraba scurra*. V. Biometric effect of chromosomal inversions. *Evolution* **14**, 284–292.

White, M. J. D. & Andrew, L. E. (1962). Effects of chromosomal inversions on size and relative viability in the grasshopper *Moraba scurra*. *The Evolution of Living Organisms* (ed. G. W. Leeper), pp. 94–101. Melbourne: Melbourne University Press.

https://doi.org/10.1017/S0016672300030342 Published online by Cambridge University Press

Appendix

Variances and covariances for a locus with k alleles

For a locus with k alleles the asymptotic variances and covariances associated with the estimators of different genotypic probabilities are the following:

$$\left.\begin{aligned}
\mathrm{Var}\,(\hat{\Theta}_{ii}) &= \frac{1}{N}\{\Theta_{ii}+(\tfrac{1}{2})^n\,\Theta_{i.}][1-(\Theta_{ii}+(\tfrac{1}{2})^n\,\Theta_{i.})] \\
&\quad +(2^{2n}\,\alpha)^{-1}\,\Theta_{i.}(1-\alpha\Theta_{i.})+\Theta_{i.}(\tfrac{1}{2})^{n-1} \\
&\quad \times[\Theta_{ii}+(\tfrac{1}{2})^n\,\Theta_{i.}]\} \\
\mathrm{Var}\,(\hat{\Theta}_{ij}) &= \frac{\Theta_{ij}(1-\alpha\Theta_{ij})}{Na} \\
\mathrm{Cov}\,(\hat{\Theta}_{ii},\hat{\Theta}_{jj}) &= \frac{1}{N}\{(\tfrac{1}{2})^n[\Theta_{j.}(\Theta_{ii}+(\tfrac{1}{2})^n\,\Theta_{i.}) \\
&\quad +\Theta_{i.}(\Theta_{jj}+(\tfrac{1}{2})^n\,\Theta_{j.})] \\
&\quad +(2^{2n}\,\alpha)^{-1}[\Theta_{ij}-\alpha\Theta_{i.}\Theta_{j.}] \\
&\quad -(\Theta_{ii}+(\tfrac{1}{2})^n\,\Theta_{i.})\,(\Theta_{jj}+(\tfrac{1}{2})^n\,\Theta_{j.})\} \\
\mathrm{Cov}\,(\hat{\Theta}_{ii},\hat{\Theta}_{ij}) &= \frac{1}{N}[(\tfrac{1}{2})^n\,\alpha^{-1}(\alpha\Theta_{ij}\,\Theta_{i.}-\Theta_{ij}) \\
&\qquad\qquad -\Theta_{ij}(\Theta_{ii}+(\tfrac{1}{2})^n\,\Theta_{i.})].
\end{aligned}\right\}\quad (\mathrm{A}\ 1)$$

Two kinds of errors are incorporated in all these formulae: that due to the examined offspring per individual (error of diagnosis of the parental genotype) and that due to the size of the sample of parents in the base population (error of sampling from the population). Both kinds of errors affect the estimation of genotypic frequencies in the population.

As an example of the derivation of these formulae we examine the variance of a heterozygote (ij). Since

$$\hat{\Theta}_{ij} = \frac{N_{ij}}{N\alpha},$$

we have that

$$\begin{aligned}
\mathrm{Var}\,(\hat{\Theta}_{ij}) = \mathrm{Var}\,[N_{ij}/(N\alpha)] &= \mathrm{Var}\,(\hat{K}_{ij}/\alpha) \\
&= [\mathrm{Var}\,(\hat{K}_{ij})]/\alpha^2.
\end{aligned}$$

The variance of $\hat{K}_{ij}$ is that of a binomial distribution, so

$$\mathrm{Var}\,(\hat{K}_{ij}) = K_{ij}(1-K_{ij})/N$$

and

$$\mathrm{Var}\,(\hat{\Theta}_{ij}) = K_{ij}(1-K_{ij})/N\alpha^2.$$

Putting K_{ij} as a function of Θ_{ij} (Table 1) we obtain

$$\mathrm{Var}\,(\hat{\Theta}_{ij}) = \frac{\Theta_{ij}(1-\alpha\Theta_{ij})}{N\alpha}.$$

https://doi.org/10.1017/S0016672300030342 Published online by Cambridge University Press