Genet. Sel. Evol. 31 (1999) 457–464

Original article

On the precision of estimation of genetic distance

Jean-Louis Foulley[a], William G. Hill[b*]

[a] Station de génétique quantitative et appliquée, Institut national de la recherche agronomique, 78352 Jouy-en-Josas cedex, France
[b] Institute of Cell, Animal and Population Biology, The University of Edinburgh, Edinburgh EH9 3JT, UK

(Received 26 April 1999; accepted 15 September 1999)

Abstract – This article gives a formal proof of a formula for the precision of estimated genetic distances proposed by Barker et al. which can be used in designing experimental sampling programmes. The derivation is given in the general multiallelic case using the Sanghvi distance. Two sources of sampling are considered, i.e. i) among individuals (or gametes) within locus and ii) among loci within populations. Distribution assumptions about gene frequencies are discussed, especially the normal used in Barker et al. versus the Dirichlet via simulation. © Inra/Elsevier, Paris

genetic distance / estimation / precision / Dirichlet

Résumé – À propos de la précision de l'estimation des distances génétiques. Cet article présente une démonstration formelle d'une formule de Barker et al. donnant la précision de l'estimation de distances génétiques à des fins de planification expérimentale. Cette démonstration est faite dans le cas général multiallélique sur la base de la distance de Sanghvi. Deux sources d'échantillonnage sont considérées à savoir i) au niveau des individus (ou gamètes) intra-locus et ii) entre loci intra-populations. Les hypothèses sur les lois des fréquences géniques sont discutées via quelques simulations en particulier celle de la loi Normale adoptée par Barker et al. par rapport à la loi de Dirichlet © Inra/Elsevier, Paris

distance génétique / estimation / précision / Dirichlet

1. INTRODUCTION

In a report to the FAO, Barker et al. [2] proposed a formula to express the standard error of an estimate of the genetic distance (d) which was intended

* Correspondence and reprints
E-mail: w.g.hill@ed.ac.uk

to be used in deciding on sample sizes when designing field programmes. They start from the following expression of the estimator:

$$D = [(\hat{p}_1 - \hat{p}_2)^2]/[\hat{\bar{p}}(1 - \hat{\bar{p}})] \tag{1}$$

where $\hat{p}_1$, $\hat{p}_2$ are the observed frequencies of a given allele at one locus in populations 1 and 2, respectively ($\hat{\bar{p}}$ being an estimate of the average frequency) in which $2n = n_1 + n_2$ individuals are sampled assuming $n_1 = n_2$; using equation (1) they infer that the standard deviation of D can be expressed as

$$SE(D) = \sqrt{2/Lk}[d + (1/n)] \tag{2}$$

where L is the number of loci and k is the number of algebraically independent distance estimates per locus, i.e. assuming $k + 1$ alleles.

As no proof of this formula was given in the paper, we thought it might be useful to provide a formal detailed derivation which also helps to clarify the assumptions made throughout and the sources of uncertainty taken into account.

2. THEORY

We will restrict our attention to the multi-allelic case. Let $y_{1j} = 2n\hat{p}_{1j}; y_{2j} = 2n\hat{p}_{2j}$ be the number of A_j alleles observed in the n individuals sampled in populations 1 and 2, respectively, with p_{1j}, p_{2j} designating the corresponding true allele frequencies. Under H_0: $(p_{1j} = p_{2j} = p_j; \forall j)$ the statistic

$$Z^2 = \sum_{j=1}^{J} (y_{1j} - y_{2j})^2 / 4n\hat{p}_j \tag{3}$$

where $\hat{p}_j = (\hat{p}_{1j} + \hat{p}_{2j})/2$, has an asymptotic chi-square distribution with $J - 1$ degrees of freedom [7].

Factorizing n, and the expectation $(J - 1)$ of the chi-square, Z^2 can be written alternatively as:

$$Z^2 = 2n \sum_{j=1}^{J} [(\hat{p}_{1j} - \hat{p}_{2j})^2 / (\hat{p}_{1j} + \hat{p}_{2j})] = n(J - 1)D \tag{4}$$

where D is the so-called Sanghvi's G^2 distance closely related to the θ^2 of Battacharyya [9].

Provided that the variance covariance matrices of $\mathbf{y}_1 = \{y_{1j}\}$ and of $\mathbf{y}_2 = \{y_{2j}\}$ are close to each other, Z^2 in equation (4) can be interpreted as a non-central chi-square with $\nu = J - 1$ degrees of freedom with a non-centrality parameter equal to $\lambda = n(J - 1)d$, where $d = (J - 1)^{-1} \sum_{j=1}^{j} [(p_{1j} - p_{2j})^2 / p_j]$

with $p_j = (p_{1j} + p_{2j})/2$ corresponding to the true distance between the two populations.

Therefore, $E(Z^2) = (J-1) + \lambda$ and $\text{var}(Z^2) = 2[(J-1) + 2\lambda]$ [1]; hence

$$E(D) = d + 1/n \tag{5}$$

and

$$\text{var}_c(D) = 2(J-1)^{-1}[(1/n^2) + (2d/n)] \tag{6}$$

Normalizing D by dividing $2\sum_{j=1}^{J}(\hat{p}_{1j} - \hat{p}_{2j})^2/(\hat{p}_{1j} + \hat{p}_{2j})$ by $(J-1)$ allows the metric to be adjusted for the number of alleles.

For a locus (k) chosen at random in the genome, the value of the distance d_k becomes a random variable, and we will consider the expectation and variance of d_k (later on designated as d for simplicity) with respect to sampling the true frequencies of alleles in populations 1 and 2 from a larger population; this results basically from sampling loci in the two populations from a pool of 'exchangeable' loci [3, 12].

Let the distribution of the vector $\mathbf{p}_{i(J\times 1)}$ of gene frequencies in a given line (say i) over 'exchangeable loci' have mean $\boldsymbol{\pi}$ and variance covariance $\rho_i\mathbf{C}$, i.e.

$$\mathbf{p}_{i(J\times 1)} \sim (\boldsymbol{\pi}, \rho_i\mathbf{C}) \tag{7}$$

where

$$\mathbf{C} = \mathbf{C}(\boldsymbol{\pi}) = \text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T \tag{8}$$

$\rho_i\mathbf{C}$ measures the 'between loci' within line component of variance in gene frequencies, which, under pure genetic drift and random mating, is also a 'between lines' within locus component of variance. Thus, in these conditions, ρ_i can be interpreted as the inbreeding coefficient F_i in line i, the value of which depends only on the effective population size (N) and the number (t) of generations of drift $F = 1 - (1 - 1/2N)^t$ [15].

The true distance $d = (J-1)^{-1}\sum_{j=1}^{J}[(p_{1j} - p_{2j})^2/\overline{p}_j]$ can be expressed as a quadratic form $d = \boldsymbol{\delta}^T\mathbf{Q}\boldsymbol{\delta}$ with $\boldsymbol{\delta}_{J\times 1} = \{\delta_j = p_{1j} - p_{2j}\}$ and the $(J \times J)$ matrix $\mathbf{Q}$ of the quadratic form being $(J-1)^{-1}\,\text{diag}(\overline{p}_j^{-1})$. Assuming $\overline{\mathbf{p}} \approx \boldsymbol{\pi}$, and taking the expectation of d with respect to the distributions of $\mathbf{p}_1$ and $\mathbf{p}_2$ requires the evaluation of:

$$E_{\mathbf{p}_1,\mathbf{p}_2}(d) = [E_{\mathbf{p}_1,\mathbf{p}_2}(\boldsymbol{\delta})]^T\mathbf{Q}E_{\mathbf{p}_1,\mathbf{p}_2}(\boldsymbol{\delta}) + \text{tr}[\mathbf{Q}\text{var}_{\mathbf{p}_1,\mathbf{p}_2}(\boldsymbol{\delta})]$$

As populations 1 and 2 are derived from the same founding population with allele frequency $\boldsymbol{\pi}$, $E(\boldsymbol{\delta}) = \mathbf{0}$. The second term is the trace of $\mathbf{Q}[\text{var}_{\mathbf{p}_1}(\mathbf{p}_1) + \text{var}_{\mathbf{p}_2}(\mathbf{p}_2)]$. As $\mathbf{C}(\overline{\mathbf{p}})$ is close to $\mathbf{C}$ $(\boldsymbol{\pi})$ if $\overline{\mathbf{p}} \approx \boldsymbol{\pi}$, this reduces to

$$\text{tr}\{\mathbf{Q}[\text{var}_{\mathbf{p}_1}(\mathbf{p}_1) + \text{var}_{\mathbf{p}_2}(\mathbf{p}_2)]\} = \rho_1 + \rho_2 \text{ since } \text{tr}(\mathbf{QC}) = (J-1)^{-1}\sum_{j=1}^{J}(1-\pi_j) = 1$$

Then,

$$E_{\mathbf{p}_1,\mathbf{p}_2}(d) = 2\overline{\rho} \tag{9}$$

where

$$\overline{\rho} = (\rho_1 + \rho_2)/2$$

So far, no assumption about a specific gene frequency distribution was needed since the expectation of a quadratic form depends only on the first two moments. Several assumptions can be made at that stage. For the sake of simplicity, a normal approximation for the distributions of true gene frequencies can be considered as in Barker et al. [2] and Lewontin and Krakauer [7]. One may also rely on the Dirichlet distribution which is the natural conjugate of the multinomial. The first alternative results in

$$d = (J-1)^{-1} \sum_{j=1}^{J} [(p_{1j} - p_{2j})^2 / \overline{p}_j] \sim 2\overline{\rho}(J-1)^{-1} \chi^2_{J-1}$$

Hence, as in equation (9) and as expected $E_{\mathbf{p}_1,\mathbf{p}_2}(d) = 2\overline{\rho}$, and

$$\mathrm{var}_{\mathbf{p}_1,\mathbf{p}_2}(d) = 8\overline{\rho}^2 (J-1)^{-1} \tag{10}$$

Remember that the total variance can be decomposed into $\mathrm{var}(D) = E_{\mathbf{p}_1,\mathbf{p}_2}[\mathrm{var}(D|\mathbf{p}_1,\mathbf{p}_2)] + \mathrm{var}_{\mathbf{p}_1,\mathbf{p}_2}[E(D|\mathbf{p}_1,\mathbf{p}_2)]$. The expressions for $E(D|\mathbf{p}_1,\mathbf{p}_2)$ and $\mathrm{var}(D|\mathbf{p}_1,\mathbf{p}_2)$ were given in equations (5) and (6) and correspond to effects on the first two moments of multinomial sampling of individuals or alleles within the two populations 1 and 2. Now

$$E_{\mathbf{p}_1,\mathbf{p}_2}[\mathrm{var}_c(D)] = 2(J-1)^{-1}[(1/n^2) + (4\overline{\rho}/n)] \tag{11}$$

$$\mathrm{var}_{\mathbf{p}_1,\mathbf{p}_2}[E_c(D)] = 8\overline{\rho}^2 (J-1)^{-1} \tag{12}$$

Combining these two formulae results in the expression for the unconditional sampling variance of the estimation of the genetic distance:

$$\mathrm{var}(D) = 2(J-1)^{-1}[2\overline{\rho} + (1/n)]^2 \tag{13}$$

the expectation being equal to $E(D) = 2\overline{\rho} + (1/n)$.

3. DISCUSSION

Formula (13) is identical to that given by Barker et al. [2] for $L = 1$ locus and $k = J - 1$ algebraically independent estimates of the genetic distance.

Incidentally, formula (9) for the expectation of d is identical to the one given by Weir [16], Laval [5] and Laval et al. [6] although these last authors considered a different distance measure, namely Reynolds'. This clearly shows the interest in normalizing the squared differences $(p_{1j} - p_{2j})^2$ by the degree of heterozygosity as in Sanghvi's and Reynolds' distances but not in Rogers', $d = \frac{1}{2}\sum_{j=1}^{J}(p_{1j} - p_{2j})^2$. Takezaki and Nei [15] consider alternative estimators of

genetic distance, and show that while the simple estimator D used here is not the best, it is only marginally less so.

To derive the expectation of d (9) it was assumed that $\overline{\mathbf{p}} \approx \boldsymbol{\pi}$. This implies computing $\overline{\mathbf{p}}$ in D (formula 4) from the whole collection of the I populations involved in the distance study either as an unweighted $\overline{\mathbf{p}} = (\sum_{i=1}^{I} \mathbf{p}_i)/I$, or as a weighted mean; to that respect we suggest for unbalanced designs with n_i individuals sampled in population i, $\overline{\mathbf{p}} = (\sum_{i=1}^{I} \alpha_i \mathbf{p}_i)/\sum_{i=1}^{I} \alpha_i$ with weights α_i inversely proportional to $\rho_i + [(1-\rho_i)/n_i]$.

Actually this condition turns out to be mandatory as demonstrated by a simulation study based on the Dirichlet distribution. This distribution and its particular case of the beta for two categories have been used by population geneticists, mostly in a Bayesian context, to specify prior information about allele frequencies [16]. Under recurrent mutation, migration and drift but without selection, Wright [17] also obtained gene frequencies at a biallelic locus which are beta distributed. Thus, that assumption makes sense as long as selection is absent or weak.

Results based on the Dirichlet distribution in the case of $J = 5$ alleles show a non-negligible downwards bias increasing with F and disequilibrium among allele frequencies when using the standard formula (*figure 1*).

One can guess at its direction by considering populations taken towards fixation: either they are fixed for the same allele or fixed for different alleles. In the biallelic case, the line is either AA or aa. If it is AA (probability π) the average distance between this line and another line is $(0 \times \pi) + \left[(1-\pi) \times \frac{(1-0)^2}{1/4}\right]$, i.e. $4(1-\pi)$. The same reasoning applies given the line is aa leading to 4π so that the expectation of the distance is $[\pi \times 4(1-\pi)] + [(1-\pi)] \times 4\pi$, i.e. $8\pi(1-\pi)$ which is lower than $2F$, here equal to 2 for the limit case. The higher the deviation of π from $1/2$, the higher the bias as observed in the simulation.

Regarding the variance of the true distance d, simulation indicates that the normal approximation overestimates it in the case of an equal frequency distribution over alleles and underestimates it under large heterogeneity (*figure 2*). The approximation works reasonably well as long as the effective number of alleles does not fall below about 70 % of its nominal number and provided the averaging of gene frequencies in the denominator is made over all populations (a value of 15 was taken in the simulation).

This makes this formula worthwhile on account of its simplicity relative to its main objective, i.e. of providing a rough estimate of the precision of estimated genetic distances, particularly when designing programmes of data collection for distance estimation, as discussed by Barker et al. [2] for breeds of livestock. For instance, using this formula with the aim of having a standard deviation of 0.03 or less for distance values of 0.1, they recommended basing breed characterization on 25 animals per breed assayed and 25 micro-satellite loci, each with an effective allele number of at least 2.

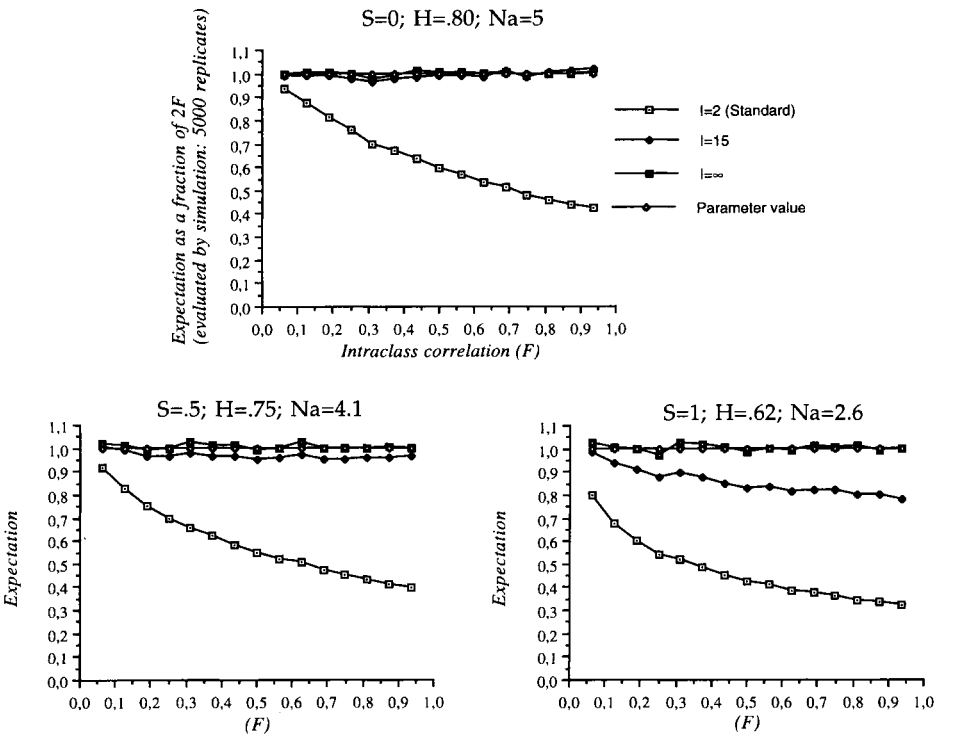

Figure 1. Expectation of standard and modified Sanghvi's distances. S: standard deviation shift of the mean in the underlying scale; H: heterozygosity; Na: effective number of alleles. $S = 0$: $p1 = p2 = p3 = p4 = p5 = 0.20$; $S = 0.5$: $p1 = 0.366$, $p2 = 0.231$, $p3 = 0.177$, $p4 = 0.136$, $p5 = 0.090$; $S = 1$: $p1 = 0.563$, $p2 = 0.209$, $p3 = 0.123$, $p4 = 0.072$, $p5 = 0.033$. The mean true frequency $\overline{p}_j$ of allele A_j in the denominator of d was obtained from a set of $I = 15$ populations

Moreover, improving it analytically might be a tedious task even for approximations. For instance, using the so-called delta method based on Taylor expansions, one should go beyond the second order expansion to obtain different results and assume specific forms for the third and higher moments of gene frequency distributions. Anyway, for those interested in further adjustments, one may recommend basing them on the following general formula (derived from equations (11) and (12)):

$$\text{var}(D) = 2(J-1)^{-1}[E(d) + (1/n)]^2 + E^2(d)[CV_d^2 - 2(J-1)^{-1}]$$

where $E(d)$ and CV_d are the expectation and coefficient of variation of the true distance, respectively.

Formule (13) also provides a means for combining inter loci information in the expression of the distance. Now, for K independent loci, a 'natural' estimator of the distance is obtained from $D = \sum_{k=1}^{K}(w_k D_k)/w_+$ where the weight w_k is proportional to the reciprocal of the variance of the distance D_k

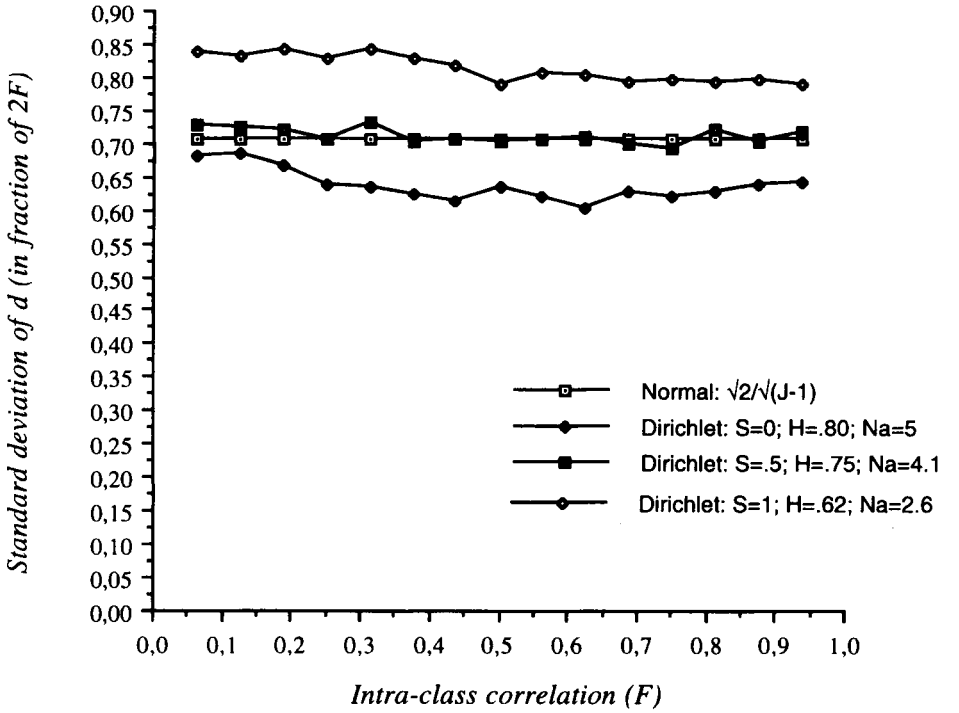

Figure 2. Normal approximation versus Dirichlet simulation: effect on the standard deviation of true distance (d) as a function of F and of the effective number of alleles ($J = 5$ allelles, $I = 15$ populations, $N = 5\,000$ replicates).

pertaining to locus k, and with $w_+ = \sum_{k=1}^{K} w_k$. From equation (13), $w_k \propto J_k - 1$ which is equivalent to weighting each locus by its number of alleles minus 1 so that the formula for the pooled distance reduces to

$$D = (J_+ - K)^{-1} \left(\sum_{k=1}^{K} \sum_{j=1}^{J} \frac{(\widehat{p}_{1jk} - \widehat{p}_{2jk})^2}{\widehat{\overline{p}}_{jk}} \right) \tag{14}$$

and its estimated variance to

$$\text{Est}[\text{var}D] = 2(J_+ - K)^{-1}[D + (1/n)]^2 \tag{15}$$

Finally, issues tackled here with respect to sampling of loci and of lines at a given locus are closely related to theories developed for testing selective neutrality: [7, 9, 11, 13, 14]. In particular, assumptions made in the distribution of gene frequencies in equation (7) rely on the type (a) structure shown in Robertson ([14], Figure 1), i.e. a set of equivalent populations deriving independently from a common base population. For more complex relationships involving some kind of splitting or fusion, one will have to adjust the mean and variance of the gene frequencies accordingly: see, for example, techniques proposed by Felsenstein [4].

ACKNOWLEDGEMENTS

The authors are grateful to Stuart Barker (University of NSW, Armidale, AU), Jean-Jacques Colleau (Inra, Jouy-en-Josas) and Christine Dillmann (INA-PG, Paris) for their comments and criticisms which helped to clarify the subject and to improve the manuscript. Thanks are also expressed to Joe Felsenstein (University of Washington, Seattle) for having provided additional references on the subject.

REFERENCES

[1] Abramowitz M., Stegun I.A., Handbook of Mathematical Functions, 9th ed., Dover publications, New York, 1972.

[2] Barker J.S.F., Bradley D.G., Fries R., Hill W.G., Nei M., Wayne R.K., An integrated global programme to establish the genetic relationships among the breeds of each domestic animal species, FAO report, Rome, Italy, 1993.

[3] Felsenstein J., Confidence limits on phylogenies: an approach using the bootstrap, Evolution 39 (1985) 783–791.

[4] Felsenstein J., Phylogenies from gene frequencies: a statistical problem, Syst. Zool. 34 (1985) 300–311.

[5] Laval G., Modélisation et mesure de la différenciation génétique des races animales à l'aide de marqueurs microsatellites, thesis, Université de Tours, 1997.

[6] Laval G., San Cristobal M., Chevalet C., Distances génétiques intra-spécifiques, 6èmes rencontres de la société francophone de classification, Montpellier, 21–23 September, 1998, pp. 135–138.

[7] Lewontin R. C., Krakauer J., Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms, Genetics 74 (1973) 174–195.

[8] McCullagh P., Nelder J., Generalized Linear Models, 2nd ed., Chapman and Hall, 1989.

[9] Nei M., Maruyama T., Lewontin-Krakauer test for neutral genes, Genetics 80 (1975) 395.

[10] Nei M., Molecular Evolutionary Genetics, Columbia University Press, 1987.

[11] Raufaste N., Bonhomme F., Properties of bias and variance of two multiallelic estimators of Fst, Theor. Popul. Biol. (1999) in press.

[12] Robert C., L'analyse statistique bayésienne, Economica, Paris, 1992.

[13] Robertson A., Remarks on the Lewontin-Krakauer test, Genetics 80 (1975) 386.

[14] Robertson A., Gene frequency distributions as a test of selective neutrality, Genetics 81 (1975) 775–785.

[15] Takezaki N., Nei M., Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA, Genetics 144 (1996) 389–399.

[16] Weir B.S., Genetic Data Analysis II, Sinauer Associates, Sunderland, MA, 1996.

[17] Wright S., Evolution in Mendelian populations, Genetics 16 (1931) 97–159.