



## Commentary

## On the sign of the average effect of an allele

Sabin Lessard<sup>a,\*</sup>, Warren J. Ewens<sup>b,1</sup><sup>a</sup> Département de mathématiques et de statistique, Université de Montréal, Montréal, Québec H3C 3J7, Canada<sup>b</sup> Department of Biology and Department of Statistics, The University of Pennsylvania, Philadelphia, PA 19104, USA

## ARTICLE INFO

## Article history:

Received 22 July 2022

Available online 19 October 2022

## Keywords:

Average effect

Average excess

Additive fitness

## ABSTRACT

The concept of the average effect of an allele pervades much of evolutionary population genetics. In this context the average effect of an allele is often considered as the main component of the “fitness” of that allele. It is widely believed that, if this fitness component for an allele is positive, then the frequency of this allele will increase, at least for one generation in discrete-time models. In this note we show that this is not necessarily the case since the average effect of an allele on fitness may be different from its marginal additive fitness even in a one-locus setting in non-random-mating populations.

© 2022 Elsevier Inc. All rights reserved.

## 1 Introduction

The idea that more fit parents leave disproportionately more offspring than less fit parents, that this higher fitness is in part inherited by the offspring of the more fit parents leading to a steady increase in the frequency of the more fit type, is at the core of the Darwinian theory. In terms of Mendelian genetics, this idea became, loosely, that more fit alleles increase in frequency over successive generations, this statement then being the basic idea of the Darwinian theory when cast in Mendelian terms.

The “fitness”, or fitness contribution, of an allele requires a precise definition. In his statement of the Fundamental Theorem of Natural Selection (FTNS), (Fisher 1930, 1958) took the view that the average effect of an allele can be seen as the “fitness” of that allele. The average effect of an allele plays a fundamental role in the analysis of the correlation between relatives, in the concept of the additive genetic variance, and therefore in the concept of narrow heritability defined as the proportion of this variance in the total variance. It is also central to Fisher’s evolutionary ideas and to the modern version of the Fundamental Theorem of Natural Selection see Ewens and Lessard, 2015, and references therein).

In this note we focus on the evolutionary significance of the average effect of an allele. We therefore address the following genetical parallel of the Darwinian paradigm: Does a positive value of the average effect of any allele imply an increase (at least for one generation in discrete-time models) in the frequency of that allele?

## 2 The analysis

It is sufficient for our purposes to limit the discussion below to an infinite diploid population undergoing discrete, nonoverlapping generations and where the fitness of any individual depends on the genotype at one autosomal locus only, denoted locus  $A$ , allowing  $n$  alleles  $A_1, \dots, A_n$ . There are two average effect concepts, and Fisher (1958, p. 35) confusingly uses both without any clarification. Here we focus on the  $\alpha$  definition (Ewens, 2004, p. 62). Suppose that at the time of conception of some parental generation (called time  $t$  below) the frequency of the ordered genotype  $A_i A_j$  is  $P_{ij}$ , so that the frequency  $p_i$  of  $A_i$  at this time is given by  $p_i = \sum_j P_{ij}$ . Suppose also that the fitness of this genotype understood as its mean viability from conception to the time of reproduction is  $w_{ij} = w_{ji} \geq 0$ , so that the mean population fitness  $\bar{w}$  at time  $t$  is  $\sum_{i,j} P_{ij} w_{ij} \geq 0$ . Then, using the least squares method, the “time  $t$ ” values of the average effects  $\alpha_1, \dots, \alpha_n$  of  $A_1, \dots, A_n$ , respectively, are found by minimizing

$$\sum_{i,j} P_{ij} (w_{ij} - \bar{w} - \alpha_i - \alpha_j)^2 \quad (1)$$

with respect to the  $\alpha_i$  values, subject to the constraint  $\sum_i \alpha_i p_i = 0$ . This constraint ensures that, except in the trivial case where all  $\alpha_i$  values are zero, at least one  $\alpha_i$  value is positive and at least one is negative. If for all  $(i, j)$  combinations  $w_{ij}$  can be written exactly in the form  $\bar{w} + c_i + c_j$  for some  $c_i$  and  $c_j$  with  $\sum_i c_i p_i = 0$ , as in the example below, then the  $c_i$  values are precisely equal to the  $\alpha_i$  values.

The minimization deriving from the expression given in (1) leads to a set of simultaneous equations defining the  $\alpha_i$  values, typified by

$$p_i \alpha_i + \sum_j P_{ij} \alpha_j = \bar{w} - p_i \quad (2)$$

\* Corresponding author.

E-mail address: [sabin.lessard@umontreal.ca](mailto:sabin.lessard@umontreal.ca) (S. Lessard).<sup>1</sup> W.J. Ewens and S. Lessard contributed equally to this work.

In this equation  $p_i$  is the inter-generational change of the frequency of the allele  $A_i$  between the times of conception of the parental and offspring generations, respectively. An alternative expression for Equation 2) given in Lessard 1997) is

$$2 \sum_j C_{ij} \alpha_j = \bar{w} (p_i), \tag{3}$$

where  $C_{ij} = E(f_i f_j)$  with  $f_i$  and  $f_j$  being defined as the frequencies of  $A_i$  and  $A_j$ , respectively, with possible values 0, 1/2 or 1 in the diploid genotype of an individual chosen at random and  $E$  denoting the expectation. Note that, owing to the constraint on the  $\alpha_i$  values, we may take  $C_{ij} = \text{Cov}(f_i, f_j) = E(f_i f_j) - p_i p_j$  in the above expression with  $\text{Cov}$  denoting the covariance. Note also that  $\bar{w} (p_i) = p_i a_i$ , where  $a_i$  is the deviation of the marginal fitness of  $A_i$  from the mean fitness in the population, called its average excess see, e.g., Lessard, 1997, and references therein).

Our focus is on the set of  $\alpha_i$  values and their relation to the set of  $p_i$  values. The collection of equations in 2) provides an implicit relation between these two sets of values, and in principle leads to explicit expressions for the  $\alpha_i$  values in terms of the various  $p_i$  values. Our aim is to explore the relation between these two sets of values to answer the question posed earlier: **Does  $\alpha_i > 0$  imply  $p_i > 0$ , and conversely?**

If mating is at random in the generation preceding the parental generation, so that Hardy–Weinberg proportions hold for genotype frequencies in the parental generation, then  $P_{ij} = p_i p_j$  for all  $(i, j)$  combinations and the second term on the left-hand side of Eq. 2) is

$$\sum_j P_{ij} \alpha_j = p_i \sum_j p_j \alpha_j = 0. \tag{4}$$

Therefore, we have  $p_i \alpha_i = \bar{w} (p_i)$ , from which  $\alpha_i$  and  $p_i$  always have the same sign if  $p_i > 0$  and  $\bar{w} > 0$ .

If mating is not necessarily at random in the generation preceding the parental generation but there are only two possible alleles at locus  $A$ , so that  $p_2 = 1 - p_1$ ,  $\alpha_2 p_2 = -\alpha_1 p_1$  and  $f_2 = 1 - f_1$  with  $0 < p_1 < 1$ , then Eq. 3) for  $i = 1$  becomes

$$\bar{w} (p_1) = 2 C_{11} \alpha_1 + C_{12} \alpha_2 = \frac{2\alpha_1}{1 - p_1} E(f_1^2) - p_1^2 = \frac{2\alpha_1}{1 - p_1} \text{Var}(f_1) \tag{5}$$

with  $\text{Var}$  denoting the variance. First, we note that  $\bar{w} (p_1) = p_1 \alpha_1$  if and only if  $\text{Var}(f_1) = p_1 (1 - p_1)/2$ , which means Hardy–Weinberg proportions for genotype frequencies. More importantly, we have in the general case of two alleles at a single locus that, if  $\bar{w} > 0$ , then  $\alpha_1$  is of the same sign as  $p_1$ , and similarly by symmetry  $\alpha_2$  and  $p_2$  have the same sign. This result is in agreement with Equations 2.62) and 2.66) in Ewens 2004). So far, therefore, the view that  $\alpha_i$  can be taken as the “fitness” contribution of  $A_i$ , in the sense that a positive value of  $\alpha_i$  leads to an increase in the frequency of  $A_i$ , can be sustained.

When there are three or more alleles at locus  $A$ , the relation between  $\alpha_i$  and  $p_i$  is not immediately clear, and we proceed by considering a simple example. Suppose that there are three alleles possible,  $A_1, A_2$  and  $A_3$ , and that the fitness values of the possible genotypes are given by the following symmetric array, with the paternally-derived gene indicated by the columns and the maternally-derived gene indicated by the rows:

	$A_1$	$A_2$	$A_3$	
$A_1$	1.018	1.018	0.937	
$A_2$	1.018	1.018	0.937	
$A_3$	0.937	0.937	0.856	6)

Moreover, the ordered genotype frequencies assuming non-random mating in the previous generation are supposed to be as follows:

	$A_1$	$A_2$	$A_3$	
$A_1$	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{1}{9}$	
$A_2$	$\frac{2}{9}$	$\frac{2}{9}$	0	
$A_3$	$\frac{1}{9}$	0	0	7)

The allele frequencies are  $p_1 = 4/9$  for  $A_1$ ,  $p_2 = 4/9$  for  $A_2$ , and  $p_3 = 1/9$  for  $A_3$ . Combining 6) and 7), the mean fitness  $\bar{w}$  is 1.

Given that  $\bar{w} = 1$ , the fitnesses given in 6) are exactly in the form  $\bar{w} + \alpha_i + \alpha_j$  for  $i, j = 1, 2, 3$ , with  $\alpha_1 = \alpha_2 = 0.009$ ,  $\alpha_3 = 0.072$ , and  $p_1 \alpha_1 + p_2 \alpha_2 + p_3 \alpha_3 = 0$ . This shows the fact that these  $\alpha_i$  values are the average effects of  $A_1, A_2$  and  $A_3$ , as the notation anticipated.

Changing his example to fitness instead of stature, and paraphrasing, the form  $\bar{w} + \alpha_i + \alpha_j$  is precisely what Fisher 1930, page 32) described as not necessarily representing the real fitnesses of  $A_i A_j$ , but as being more intimately involved with evolutionary changes than the fitnesses  $w_{ij}$  themselves. It is in effect the interpretation of this view that we are addressing.

Since  $\bar{w} = 1$  there is no normalization needed to find the genotype frequencies at the time of reproduction of the parental generation. Allowing for the various fitness values given in 6), these are :

	$A_1$	$A_2$	$A_3$	
$A_1$	$\frac{1}{9} \times 1.018$	$\frac{2}{9} \times 1.018$	$\frac{1}{9} \times 0.937$	
$A_2$	$\frac{2}{9} \times 1.018$	$\frac{2}{9} \times 1.018$	0	
$A_3$	$\frac{1}{9} \times 0.937$	0	0	8)

Now, whatever the mating scheme is and as long as there is no difference in fertilities, the frequency of  $A_1$  in the offspring generation is

$$\frac{3}{9} \times 1.018 + \frac{1}{9} \times 0.937 = \frac{1}{9} \times 3.991 \tag{9}$$

compared to 4/9 in the parental generation at conception. The change  $p_1$  in the frequency of  $A_1$  is  $3.991/9 - 4/9 = 0.001 < 0$ , despite the fact that  $\alpha_1$  is positive. Note, however, that the average effect and average excess for the other two alleles are of the same sign, since  $p_2 = 4 \times 1.018/9 > 0$  and  $p_3 = 0.937/9 < 0$  with  $\alpha_2 > 0$  and  $\alpha_3 < 0$ .

### 3 Discussion

The assumption that fitnesses depend on the alleles at one locus only is of course quite unrealistic. We have chosen the simple example above in this setting to show that it is not necessarily true that a positive value of the average effect  $\alpha_i$  of an allele  $A_i$  implies an increase in the frequency of  $A_i$ . Owing to Eq. 2) that implicitly defines the average effects, an increase in the frequency of  $A_i$  occurs under the conditions  $p_i > 0$  and  $\bar{w} > 0$  if and only if  $\bar{\alpha}_i > 0$ , where

$$\bar{\alpha}_i = \alpha_i + \sum_j \frac{P_{ij}}{p_i} \alpha_j \tag{10}$$

is the marginal additive fitness of  $A_i$ . This is the expected additive fitness of an individual given that a gene chosen at random at locus  $A$  in this individual is  $A_i$ . Let us recall that the additive fitness of an individual, or its genetic value in fitness, is defined as the sum of the average effects of all the alleles in this individual on fitness. We note that the sign of  $\bar{\alpha}_i$  may differ from the sign of  $\alpha_i$  in non-random-mating populations as a result of correlation

between allele frequencies within an individual so that  $P_{ij} = p_i p_j$  does not necessarily hold for all  $j$  indices. Curiously, more than two alleles at the same locus are necessary for this correlation to overcome the effect of the average effect on the sign of the change in the allele frequency. It is worth noting that Fisher used a two-allele setting in the original edition of *The Genetical Theory of Natural Selection* (Fisher, 1930) and made only superficial changes to address the more general case of multiple alleles at one locus in the revised edition (Fisher, 1958). As for the multi-locus setting, it was not analyzed in Fisher's book.

In cases where the fitness of any individual depends on the allelic genes that the individual has at multiple loci, the concept of average effect can be extended (see, e.g., Castilloux and Lessard, 1995; Lessard, 1997, or Ewens, 2004, Section 7.4). With  $n_k$  alleles  $A_{k1}, \dots, A_{kn_k}$  at locus  $k$ , the average effects are implicitly defined by the set of linear equations in the form

$$p_{ki}\alpha_{ki} + \sum_j P_{ki,kj}\alpha_{kj} + \sum_{l \neq k} \sum_j Q_{ki,lj}\alpha_{lj} = \bar{w} \quad (11)$$

Here,  $\alpha_{ki}$  is the average effect of allele  $A_{ki}$  whose frequency at locus  $k$  is  $p_{ki}$ , while  $P_{ki,kj}$  is the frequency of the ordered genotype  $A_{ki}A_{kj}$  at locus  $k$  and  $Q_{ki,lj}$  twice the probability that, in an individual chosen at random, a gene chosen at random from locus  $k$  is  $A_{ki}$  and a gene chosen independently at locus  $l$  is  $A_{lj}$ . The left-hand side in Eq. 11) is equal to  $p_{ki}\bar{\alpha}_{ki}$  where  $\bar{\alpha}_{ki}$  represents the marginal additive fitness of  $A_{ki}$  at locus  $k$  with the same interpretation as in a one-locus setting. In a more general multi-locus setting, however, we may expect the occurrence of more discrepancies between the signs of the average effect and the corresponding change in allele frequency with correlation between allele frequencies across loci coming into play. This is perhaps especially the case where the population mean fitness decreases from one generation to the next due to recombination and epistatic effects, which can happen even for the case of random mating (see, e.g., Ewens, 2004, section 6.2). Let us recall that the equality  $Q_{ki,lj} = 2p_{ki}p_{lj}$  for  $l \neq k$ , which corresponds to

linkage equilibrium, does not generally hold whatever the mating system is. It would be an interesting challenge to delimit the cases where a positive value of  $\alpha_{ki}$  implies an increase in frequency of  $A_{ki}$ : it is perhaps only rarely that this does not happen. However, the calculation of average effects as the numbers of loci and alleles at each locus increase might be out of reach given the amount of data on the population state to collect and the dimension of the linear system to solve. For now we can at least say that the identification of the sign (positive or negative) of the average effect of an allele determines the sign of the change in its frequency is not necessarily correct. In other words, the sign of the average effect may differ from the sign of the average excess and, reminding ourselves that an average effect is the result of linear regression of fitness on allele frequency, this can occur more generally in multiple linear regression for any variables.

#### Data availability

No data was used for the research described in the article.

#### Acknowledgments

S. Lessard is supported by the Natural Sciences and Engineering Research Council of Canada, grant no. 8833.

#### References

- Castilloux, A.-M., Lessard, S., 1995. The fundamental theorem of natural selection in Ewens' sense (Case of many loci). *Theor. Popul. Biol.* 48, 306–315.
- Ewens, W.J., 2004. *Mathematical Population Genetics*. Springer, New York, USA.
- Ewens, W.J., Lessard, S., 2015. On the interpretation and relevance of the fundamental theorem of natural selection. *Theor. Popul. Biol.* 104, 59–67.
- Fisher, R.A., 1930. *The Genetical Theory of Natural Selection*. Clarendon, Oxford, UK.
- Fisher, R.A., 1958. *The Genetical Theory of Natural Selection*. Dover, New York, USA.
- Lessard, S., 1997. Fisher's fundamental theorem of natural selection revisited. *Theor. Popul. Biol.* 52, 59–67.