A new integrative indicator to assess crop genetic diversity? About the publication by Bonneuil et al. (2012), published in Ecological Indicators 23, 280–289

André Gallais, Francois Lefèvre

Letter to the editor

A new integrative indicator to assess crop genetic diversity ?

About the publication by Bonneuil *et al* (2012), *Ecological Indicators* 23, 280-289.

A. Gallais[a,*], F. Lefèvre[b]

[a] *INRAE-UPS-CNRS, UMR Génétique Quantitative et Evolution, Ferme du Moulon, 91190 Gif-sur-Yvette, France*

[b] *INRAE, Ecologie des Forêts Méditerranéennes, URFM, 84914 Avignon, France*

* Corresponding author at INRAE-UPS-CNRS, UMR Génétique Quantitative et Evolution, Ferme du Moulon, 91190 Gif-sur-Yvette, France

E-mail addresses : andre.gallais@inrae.fr (A. Gallais), francois.lefevre.2@inrae.fr (F. Lefèvre)

Abstract

To study the change of genetic diversity in wheat cultivated varieties over the French territory from the end of the nineteenth century to 2006, Bonneuil *et al.* (2012) defined and used an indicator to account for the spatial share of the different varieties. However, we found two errors in the implementation of this indicator. The first error is to combine an estimation of weighted genetic diversity among populations with the unweighted Nei coefficient of differentiation among populations (G_{ST}). Furthermore, the authors considered, what could be justified, that varieties cultivated around the period 1910-1930, i.e. land races, have now lost their within population diversity due to the process of their maintenance. Then, to retrieve the total genetic diversity present at the period when land races were cultivated, they proposed to add an estimate of the within-variety diversity to the current estimate of the between-variety diversity, which they considered as equal to Nei's parameter (D_{ST}) of genetic differentiation between populations. This is the second and main error. Indeed, we show that, when each population is reduced to one single line, the expectation of the between-line genetic diversity is not equal to the D_{ST} at the level of the heterogeneous populations but is near to their total gene diversity. The result of the authors' computation is then a high overestimation of genetic diversity for the period where land races were cultivated. The consequence of the two main errors is that the proposed indicator is not scientifically based and its application leads to erroneous conclusions.

Key words : gene diversity indicator, Nei index, within-population diversity, between-population diversity, weighting gene frequencies, wheat populations.

For studying the change of gene diversity in wheat cultivated varieties over the French territory from the end of the nineteenth century to 2006, Bonneuil *et al*. (2012) have proposed and applied an indicator of gene diversity taking into account the relative acreage of the varieties and the lost of gene diversity during the maintenance process of old varieties (land races). This indicator is based on the Nei index which gives a decomposition of the total gene diversity (H_T) into an average within-subpopulation gene diversity (H_S) and a parameter of subpopulation differentiation (D_{ST}):

$H_T = H_S + D_{ST}$.

To do this, the varieties are considered as subpopulations. According to the authors, one originality of their indicator is the use of allelic frequencies from molecular data weighted by the relative acreage of the varieties, i.e. the relative size of the populations. With such a weighting, following Chakraborty (1974), the total gene diversity can be decomposed in a weighted average gene diversity within subpopulations and a weighted average gene diversity between subpopulations. Therefore, in the definition of the indicator proposed by the authors, $H_T^*=H^*/G_{ST}$, the numerator (H^*, weighted total gene diversity) and the denominator unweighted differentiation parameter, $G_{ST} = D_{ST}/H_T$) are not consistent one with the other. This is a first error. Indeed, referring to the effect of neglecting the unequal population sizes, Chakraborty (1974) mentions that: "*in the case where interpopulation gene diversity is very small as compared to intrapopulation one, the effect of the unequal population sizes is not appreciable*". This is not what is expected in the case where populations are varieties, as in the wheat data considered by the authors, where a significant interpopulation gene diversity is expected in spite of the presence of a relatively large intrapopulation gene diversity within the land races. Note that the authors in their application consider $G_{ST} = 0.40$, which is not a small value. Furthermore, an experimental study given by Chakraborty shows that the estimates of G_{ST} can be quite different with and without weighting of allelic frequencies by the population size.

A second error is at the level of the interpretation of gene diversity from the wheat data. Before the 1910's wheat varieties were mainly land races which were genetically heterogeneous and progressively from 1930 to 1950, land races were progressively replaced by line varieties, genetically homogeneous. It is then considered, although not explicitly, that all the within-variety gene diversity of these 'old' heterogeneous populations has been lost today due to genetic drift and phenotypic selection during the process of their maintenance.

Then, each population variety is considered as represented today by one line. This appears from the authors' considerations: "*Note that as only five individuals derived from self-pollination were used to genotype a variety, we could not assess within-variety diversity but rather the five individuals were considered as homozygous and genetically homogeneous and the rare multi-allelic profiles were replaced by missing data*".

Next, the reasoning in the publication is the following: since all the within population diversity has been lost in the samples available today, to reconstruct the total gene diversity present when population varieties were cultivated, on the basis of Nei's equation, it is necessary to add to the between population differentiation (D_{ST}) an estimate of the within population diversity (H_S). In practice, considering that they could apply Nei's decomposition to unbalanced weighting, the authors used the relationship $H_T = D_{ST}/G_{ST}$, G_{ST} being the relative differentiation parameter introduced by Nei (1973), $G_{ST} = D_{ST}/H_T$. To do this, the authors considered that D_{ST} estimated at the level of populations currently reduced to one line was equal to the D_{ST} at the level of the original heterogeneous populations, and they used *a priori* fixed values of G_{ST}, $G_{ST} = 0.40$ for population varieties (taken from the literature) and obviously $G_{ST} = 1$ for line varieties.

The problem is to have a correct estimate of D_{ST} at the level of the original land races. It is considered in the publication that the total diversity estimated among single lines derived from population varieties, which is equal to the D_{ST} among lines (because for one line, $H_S = 0$), is equivalent to the differentiation parameter D_{ST} between the original heterogeneous populations. This is wrong, as we show in appendix for the situation of balanced weighting. Indeed, precisely, the expectation of D_{ST} among lines (hereafter D_{STL}), with one line per population, is equal to the total gene diversity at the level of heterogeneous populations minus $1/s\ H_S$. When s is sufficiently high, as in the wheat data, it is near to the total gene diversity H_T. This can be easily understood by the fact that the allele drawn from each population with

one line per population depends in additive manner on the gene diversity among populations (D_{ST}) and also on the gene diversity within a population (H_S). Thus, at the level of all populations, with a large number of populations, in expectation, the gene diversity among lines, with one line per population, is equal to the total gene diversity $H_T = H_S + D_{ST}$.

Consequently, the proposed estimator of the total gene diversity $H_t = D_{ST}/G_{ST}$ divides D_{STL}, which is near to the total gene diversity for a sufficiently large number s of populations (for example $s > 10$), by the parameter G_{ST} which is less than 1. With the value $G_{ST} = 0.40$, as in the wheat study, this results in a high overestimation of the total gene diversity. For a number s of populations around 50, as in the wheat study presented by the authors, the overestimation is approximately 1.5 times the total gene diversity at the population variety level, i.e., the true total gene diversity has been multiplied by 2.5; the true total gene diversity is doubled with only three populations ($s = 3$). If $G_{ST} = 1$ (case of line varieties) there is obviously no overestimation. Therefore, the apparent high decrease in total gene diversity from the years 1910-1920 to the years 1940-1950 expresses mainly a decrease of the overestimation related to the proportion of line varieties but not a decrease due to the loss of total gene diversity. Indeed, without the error in the reasoning, the conclusion is quite different as it appears only a tendency for a decrease in genetic diversity (compare figures 2 and 3 of the publication, pages 286-287).

Note also that the use by the authors of a G_{ST} value from the literature, thus independent of the data, and the use of an average G_{ST} for a given period of time in order to take into account the proportion of the different types of varieties (land races, lines) raises another issue. Indeed, G_{ST} parameter is a measure of differentiation between subpopulations conditionally to the level of within-subpopulation diversity (Hedrick, 2005 ; Jost, 2008 ; Gerlach *et al*., 2010). A consequence is that G_{ST} values from different sets of subpopulations

that show different levels of within-subpopulation diversity are not comparable and cannot be compiled together, as proposed by the authors in their new index.

In conclusion, the issue of the effective gene diversity in use, raised in this publication, is relevant and any integrative indicator accounting for both unequal use of the varieties and unequal within-population diversities would be useful. Unfortunately, due to the two main errors, i.e. application of Nei's index to weighted allelic frequencies and estimation of D_{ST} among heterogeneous populations by the D_{ST} among derived lines, the proposed indicator is not scientifically based and therefore its use leads to erroneous conclusions.

References

Bonneuil, C., Goffaux, R., Bonnin, I., Montalent, P., Hamond, C., Balfourier, F., Goldringer, I., 2012. A new integrative indicator to assess crop genetic diversity. *Ecological Indicators* 23, 280–289.

Chakraborty, R., 1974. A Note on Nei's Measure of Gene Diversity in a Substructured Population. *Humangenetik* 21, 85-88.

Gerlach, G., Jueterbock, A., Kraemer, P., Deppermann, J., Harmand, P., 2010. Calculations of population differentiation based on G_{ST} and D: forget G_{ST} but not all of statistics! *Molecular Ecology* 19, 3845-3852.

Hedrick, P.W., 2005. A standardized genetic differentiation measure. *Evolution* 59, 1633-1638.

Jost, L., 2008. G_{ST} and its relatives do not measure differentiation. *Molecular Ecology* 17, 4015–4026.

Nei, M., 1973. Analysis of Gene Diversity in Subdivided Populations. *Proc. Nat. Acad. Sci.* USA 70, 12, 3321-3323.

Appendix : D_{ST} between lines derived from heterogeneous populations is not equal to D_{ST} between these populations but is near to the total genetic diversity at the level of the populations.

H_T defined by Nei (1973) as the total gene diversity, is the probability of drawing two different alleles at the level of the whole population either from the same population or from two distinct populations. Then, we can write

$$H_T = 1/s\ H_W + (1-1/s)\ H_B \qquad (1)$$

where s is the number of populations, H_W (= H_S in Nei's equation) and H_B are respectively the average probability to draw two different genes within the same population and the average probability to draw one gene in a population and the other in another population. Then at the level of populations as $H_T = H_S + D_{ST}$, using (1), it results

$$D_{ST} = (s-1)/s\ (H_B - H_S). \qquad (2)$$

Adding subscript P for populations we can write

$$D_{STP} = (s-1)/s\ (H_{BP} - H_{SP}). \qquad (3)$$

Then, following the author approach, we consider the case where one line is extracted at random from each population (as wheat is a autogamous species, populations can be considered as a mixture of homozygous lines). With the added subscript L for the lines derived from the populations, with one line per population, we can write, from Nei's decomposition $H_{TL} = D_{STL}$ because $H_{SL} = 0$, and from (2) with $H_{SL} = 0$, it results

$$D_{STL} = (s-1)/s\ H_{BL} = H_{TL}. \qquad (4)$$

But the expectation of H_{BL}, the probability to draw two different alleles from two different lines (one per population), in the absence of selection, is equal in average to the probability H_{BP} to draw two different alleles from two different populations because the whole allele sampling is the same. Thus the expectation of D_{STL} is

$$E(D_{STL}) = (s-1)/s\ H_{BP}\ , \qquad (5)$$

and from (3) and (5) it results

$D_{STP} = E(D_{STL}) - (s\text{-}1)/s\ H_{SP}$ or $E(D_{STL}) = D_{STP} + H_{SP} - 1/s\ H_{SP}$,

So, $E(D_{STL})$ is not equal to the D_{STP} as assumed in the publication, and

$$E(D_{STL}) = H_{TP} - 1/s\ H_{SP}, \quad \text{or} \quad H_{TP} = E(D_{STL}) + 1/s\ H_{SP}. \tag{6}$$

This result can be shown more directly in the case where the number of populations s is sufficiently large (for example $s > 10$). In this situation we directly have from (1) $H_{TP} \approx H_{BP}$ and $E(H_{TL}) \approx E(H_{BL}) \approx E(D_{STL}) \approx H_{TP}$ (because $E(H_{BL}) \approx H_{BP}$, as noted before). Thus $E(D_{STL})$ is not equal to D_{STP} but is approximately equal to the total gene diversity H_{TP}.

Consequently, according to (6), to have an estimate of the total gene diversity at the level of the populations it is necessary to add only $1/s\ H_{SP}$ to the estimated diversity H_{TL} which is near to the total gene diversity of the populations H_{TP} when s is sufficiently large, whereas the author approach by using Ht[1] $= E(D_{STL})/G_{STP} = (H_{TP} - 1/s\ H_{SP})/G_{STP}$, leads to add H_{SP} to the estimated diversity H_{TL}. The result is an overestimation of the total diversity:

$Ht - H_{TP} = (H_{TP} - 1/s\ H_{SP})/G_{STP} - H_{TP} = (s\text{-}1)/s\ H_{TP}\ (1\text{-}G_{STP})/G_{STP}$ or in relative value

$Ht/H_{TP} = 1 + (s\text{-}1)/s\ (1\text{-}G_{STP})/G_{STP} = [1 - (1\text{-}G_{STP})/s]/G_{STP}$, for $s > 1$, or more simply when s is sufficiently large $Ht/H_{TP} \approx 1/G_{STP}$. Example with $G_{STP} = 0.40$, as considered by the authors and $s = 10$, $Ht/H_{TP} = 2.35$ and with the approximation that s is sufficiently large $Ht/H_{TP} \approx 2.50$. So it is not necessary to have s very large in order to have an approximation of the overestimation by $Ht/H_{TP} \approx 1/G_{STP}$.

[1] we note author's indicator Ht and not H_T as in the publication, in order to differentiate it from the true total gene diversity H_T, as noted by Nei (1973).