



HAL
open science

**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

► **To cite this version:**

André Gallais, Francois Lefèvre. A new integrative indicator to assess crop genetic diversity? About the publication by Bonneuil et al. (2012), published in Ecological Indicators 23, 280–289. Ecological Indicators, 2020, 116, pp.106390. <10.1016/j.ecolind.2020.106390>. <hal-03321450>

HAL Id: hal-03321450

<https://hal.science/hal-03321450v1>

Submitted on 6 Jan 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY 4.0 - Attribution - International License

1 Letter to the editor

2 **A new integrative indicator to assess crop genetic diversity ?**

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

4 A. Gallais^{a,*}, F. Lefèvre^b

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

7 ^b *INRAE, Ecologie des Forêts Méditerranéennes, URFM, 84914 Avignon, France*

8

9

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

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

14

15 **Abstract**

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

34

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

37

38

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

46 $H_T = H_S + D_{ST}$.

47 To do this, the varieties are considered as subpopulations. According to the authors,
48 one originality of their indicator is the use of allelic frequencies from molecular data weighted
49 by the relative acreage of the varieties, i.e. the relative size of the populations. With such a
50 weighting, following Chakraborty (1974), the total gene diversity can be decomposed in a
51 weighted average gene diversity within subpopulations and a weighted average gene diversity
52 between subpopulations. Therefore, in the definition of the indicator proposed by the authors,
53 $H_T^* = H^*/G_{ST}$, the numerator (H^* , weighted total gene diversity) and the denominator
54 unweighted differentiation parameter, $G_{ST} = D_{ST}/H_T$) are not consistent one with the other.
55 This is a first error. Indeed, referring to the effect of neglecting the unequal population sizes,
56 Chakraborty (1974) mentions that: “*in the case where interpopulation gene diversity is very*
57 *small as compared to intrapopulation one, the effect of the unequal population sizes is not*
58 *appreciable*”. This is not what is expected in the case where populations are varieties, as in
59 the wheat data considered by the authors, where a significant interpopulation gene diversity is
60 expected in spite of the presence of a relatively large intrapopulation gene diversity within the
61 land races. Note that the authors in their application consider $G_{ST} = 0.40$, which is not a small
62 value. Furthermore, an experimental study given by Chakraborty shows that the estimates of
63 G_{ST} can be quite different with and without weighting of allelic frequencies by the population
64 size.

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

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

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

87 The problem is to have a correct estimate of D_{ST} at the level of the original land races.
88 It is considered in the publication that the total diversity estimated among single lines derived
89 from population varieties, which is equal to the D_{ST} among lines (because for one line, $H_S =$
90 0), is equivalent to the differentiation parameter D_{ST} between the original heterogeneous
91 populations. This is wrong, as we show in appendix for the situation of balanced weighting.
92 Indeed, precisely, the expectation of D_{ST} among lines (hereafter D_{STL}), with one line per
93 population, is equal to the total gene diversity at the level of heterogeneous populations minus
94 $1/s H_S$. When s is sufficiently high, as in the wheat data, it is near to the total gene diversity
95 H_T . This can be easily understood by the fact that the allele drawn from each population with

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

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

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

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

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

128

129 **References**

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

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

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

138 Hedrick, P.W., 2005. A standardized genetic differentiation measure. *Evolution* 59, 1633-
139 1638.

140 Jost, L., 2008. G_{ST} and its relatives do not measure differentiation. *Molecular Ecology* 17,
141 4015–4026.

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

144

145

146

147

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

151
152 H_T defined by Nei (1973) as the total gene diversity, is the probability of drawing two
153 different alleles at the level of the whole population either from the same population or from
154 two distinct populations. Then, we can write

$$155 \quad H_T = 1/s H_W + (1-1/s) H_B \quad (1)$$

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

$$160 \quad D_{ST} = (s-1)/s (H_B - H_S). \quad (2)$$

161 Adding subscript P for populations we can write

$$162 \quad D_{STP} = (s-1)/s (H_{BP} - H_{SP}). \quad (3)$$

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

$$168 \quad D_{STL} = (s-1)/s H_{BL} = H_{TL}. \quad (4)$$

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

$$173 \quad E(D_{STL}) = (s-1)/s H_{BP}, \quad (5)$$

174 and from (3) and (5) it results

175 $D_{STP} = E(D_{STL}) - (s-1)/s H_{SP}$ or $E(D_{STL}) = D_{STP} + H_{SP} - 1/s H_{SP}$,

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

177 $E(D_{STL}) = H_{TP} - 1/s H_{SP}$, or $H_{TP} = E(D_{STL}) + 1/s H_{SP}$. (6)

178 This result can be shown more directly in the case where the number of populations s is
 179 sufficiently large (for example $s > 10$). In this situation we directly have from (1) $H_{TP} \approx H_{BP}$

180 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

181 $E(D_{STL})$ is not equal to D_{STP} but is approximately equal to the total gene diversity H_{TP} .

182 Consequently, according to (6), to have an estimate of the total gene diversity at the
 183 level of the populations it is necessary to add only $1/s H_{SP}$ to the estimated diversity H_{TL}

184 which is near to the total gene diversity of the populations H_{TP} when s is sufficiently large,

185 whereas the author approach by using $Ht^1 = E(D_{STL})/G_{STP} = (H_{TP} - 1/s H_{SP})/G_{STP}$, leads to add

186 H_{SP} to the estimated diversity H_{TL} . The result is an overestimation of the total diversity:

187 $Ht - H_{TP} = (H_{TP} - 1/s H_{SP})/G_{STP} - H_{TP} = (s-1)/s H_{TP} (1-G_{STP})/G_{STP}$ or in relative value

188 $Ht/H_{TP} = 1 + (s-1)/s (1-G_{STP})/G_{STP} = [1 - (1-G_{STP})/s]/G_{STP}$, for $s > 1$, or more simply

189 when s is sufficiently large $Ht/H_{TP} \approx 1/G_{STP}$. Example with $G_{STP} = 0.40$, as considered by the

190 authors and $s = 10$, $Ht/H_{TP} = 2.35$ and with the approximation that s is sufficiently large

191 $Ht/H_{TP} \approx 2.50$. So it is not necessary to have s very large in order to have an approximation

192 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).