Measuring genetic distances between breeds: use of some distances in various short term evolution models

Guillaume Laval, Magali San Cristobal, Claude Chevalet

Genet. Sel. Evol. 34 (2002) 481–507

DOI: 10.1051/gse:2002019

Original article

Measuring genetic distances between breeds: use of some distances in various short term evolution models

Guillaume LAVAL*, Magali SANCRISTOBAL**, Claude CHEVALET

Laboratoire de génétique cellulaire, Institut national de la recherche agronomique, BP 27, Castanet-Tolosan cedex, France

(Received 9 May 2001; accepted 21 December 2001)

Abstract – Many works demonstrate the benefits of using highly polymorphic markers such as microsatellites in order to measure the genetic diversity between closely related breeds. But it is sometimes difficult to decide which genetic distance should be used. In this paper we review the behaviour of the main distances encountered in the literature in various divergence models. In the first part, we consider that breeds are populations in which the assumption of equilibrium between drift and mutation is verified. In this case some interesting distances can be expressed as a function of divergence time, t, and therefore can be used to construct phylogenies. Distances based on allele size distribution (such as $(\delta\mu)^2$ and derived distances), taking a mutation model of microsatellites, the Stepwise Mutation Model, specifically into account, exhibit large variance and therefore should not be used to accurately infer phylogeny of closely related breeds. In the last section, we will consider that breeds are small populations and that the divergence times between them are too small to consider that the observed diversity is due to mutations: divergence is mainly due to genetic drift. Expectation and variance of distances were calculated as a function of the *Wright-Malécot* inbreeding coefficient, F. Computer simulations performed under this divergence model show that the Reynolds distance [57] is the best method for very closely related breeds.

microsatellites / breeds / divergence / mutation / genetic drift

1. INTRODUCTION

Assuming a species-like evolution pattern (evolution scheme as a dichotomy), the time scale that separates breeds is rather low with regards to the hundreds of thousands of years separating species. In order to measure the

* Present address: Computational and Molecular Population Genetics Laboratory, Zoologisches Institut, Baltzerstrasse 6, 3012 Bern, Switzerland

** Correspondence and reprints

E-mail: msc@toulouse.inra.fr

genetic distances between closely related populations like breeds, it is desirable to use highly polymorphic markers such as microsatellites [3,4,9,15,18,24,37, 40,53,59,60,70].

The high number of microsatellites distributed over whole genomes coupled with their very rapid evolution rates make them particularly useful for working out relationships among very closely related populations [14,21,22,62,64,66]. Microsatellite markers are a class of tandem repeat loci exhibiting a high mutation rate. Therefore, a high level of polymorphism can be maintained within relatively small samples. The within breed average heterozygosity is generally higher than 0.5 [37,40,54] with extreme values above 0.8 observed for several loci [33]. For a large proportion of microsatellites, the number of alleles observed across mammalian populations can vary between less than 10 to 20 and can be even higher across natural populations of fish [56].

In this paper, we study the behaviour of the genetic distances between two isolated populations, denoted X and Y, diverging from a founder population P_0 for a small number of non-overlapping generations (*Short term evolution models*). The founder and derived populations are characterised by their allele frequencies $p_{0,i}$, $p_{X,i}$ and $p_{Y,i}$ (for $i = 1..k$) respectively at the ℓth loci (the indices ℓ varying from 1 to L were omitted).

For the sake of simplicity, the formulae of distances presented in the first section of the present paper are given assuming that the true allele frequencies are known. In practice, $p_{X,i}$ and $p_{Y,i}$ are estimated from a limited number of individuals: $x_i = \frac{m_{X,i}}{m_{X,\bullet}}$ and $y_i = \frac{m_{Y,i}}{m_{Y,\bullet}}$, where $m_{X,i}$ (resp. $m_{X,i}$) is the number of alleles i and $m_{X,\bullet}$ (resp. $m_{Y,\bullet}$) the total number of genes in sample X (resp. Y).

In the second section we will review the behaviour of genetic distances under the classical model of evolution of neutral markers assuming combined effects of mutation and genetic drift [28,29,38,41,52].

The negligible effect of mutations in a rather low divergence time allows us to consider in the third section the relationship between expectation and variance of distances and the *Wright-Malécot* inbreeding coefficient F [39] assuming genetic drift only. In order to guide the choice of distances, we will check their efficiency by computer simulations.

2. PRESENTATION OF DISTANCES

The apparent diversity of genetic distances may be structured into two or three main groups: the distances based on allele distributions of frequencies – Euclidean and angular distances – and the distances based on allele size distributions.

2.1. Distances based on allele frequency distributions

2.1.1. Euclidean and related distances

Denote by $\mathbf{X} = (p_{X,1}, \ldots, p_{X,k})$ and $\mathbf{Y} = (p_{Y,1}, \ldots, p_{Y,k})$ the vectors of allele frequencies of populations X and Y. The basis of distances overlooked in this paragraph is a norm $||\mathbf{X} - \mathbf{Y}||$. Gregorius [26] uses $||\mathbf{X} - \mathbf{Y}||_1$ the sum of absolute allele frequency differences to define the absolute distance D_{G}

$$D_{\mathrm{G}} = ||\mathbf{X} - \mathbf{Y}||_1 = \sum_i |p_{X,i} - p_{Y,i}|. \tag{1}$$

The sum of the squares of allele frequency differences, $||\mathbf{X}-\mathbf{Y}||_2$, usually called the Euclidean distance, has been directly used by Gower [25] and Goodman [23]

$$D_{\mathrm{E}} = ||\mathbf{X} - \mathbf{Y}||_2 = \sqrt{\sum_i (p_{X,i} - p_{Y,i})^2}. \tag{2}$$

Dividing (2) by $\sqrt{2}$, defines D_{Rog}, the Roger distance [58], and taking the square provides the minimum distance [46]

$$D_{\mathrm{m}} = \frac{1}{2}||\mathbf{X} - \mathbf{Y}||_2^2 = \frac{1}{2}\sum_i (p_{X,i} - p_{Y,i})^2. \tag{3}$$

According to the Nei notations [46] of gene identity $j, j_X = \sum_i p_{X,i}^2, j_Y = \sum_i p_{Y,i}^2$ (or expected homozygosity) and $j_{XY} = \sum_i p_{X,i} p_{Y,i}$ and diversity ($d = 1 - j$ or expected heterozygosity), D_{m} may be rewritten as the between populations gene diversity reduced by the average of the within population gene diversity

$$\begin{aligned} D_{\mathrm{m}} &= \frac{1}{2}(j_X + j_Y) - j_{XY} \\ &= d_{XY} - \frac{1}{2}(d_X + d_Y). \end{aligned} \tag{4}$$

Between two populations, G_{ST} [47] is generally expressed with the heterozygosity of the total population $H_{\mathrm{T}} = 1 - \sum_i \bar{p}_i{}^2$ (with $\bar{p}_i = (p_{X,i} + p_{Y,i})/2$) and the average of the expected heterozygosity within populations $\bar{H} = \frac{1}{2}(H_X + H_Y)$ ($H_X = 1 - j_X = d_X$ and $H_Y = 1 - j_Y = d_Y$)

$$G_{\mathrm{ST}} = \frac{H_{\mathrm{T}} - \bar{H}}{H_{\mathrm{T}}}. \tag{5}$$

It can be rewritten as

$$G_{\mathrm{ST}} = \frac{1}{4}\frac{\sum_i (p_{X,i} - p_{Y,i})^2}{\left(1 - \sum_i \bar{p}_i^2\right)} = \frac{1}{2}\frac{D_{\mathrm{m}}}{\left(1 - \sum_i \bar{p}_i^2\right)} \tag{6}$$

which is also called the distance of Morton [42].

Other variations of the minimum distance, γ_{L} and D_{R}, were used by Latter [31,32] and Reynolds [57] respectively

$$\gamma_{\mathrm{L}} = \frac{\sum_i (p_{X,i} - p_{Y,i})^2}{\left(\sum_i p_{X,i}^2 + \sum_i p_{Y,i}^2\right)} = \frac{2D_{\mathrm{m}}}{(j_X + j_Y)} \tag{7}$$

$$D_{\mathrm{R}} = \frac{1}{2} \frac{\sum_i (p_{X,i} - p_{Y,i})^2}{1 - \sum_i (p_{X,i} p_{Y,i})} = \frac{D_{\mathrm{m}}}{1 - j_{XY}} \cdot \tag{8}$$

In parallel, Balakrishnan and Sanghvi [1], and Barker [2] defined respectively

$$\chi^2 = \frac{1}{2} \sum_i \frac{(p_{X,i} - p_{Y,i})^2}{\bar{p}_i} \tag{9}$$

and

$$D_{\mathrm{B}} = \frac{1}{2} \sum_i \frac{(p_{X,i} - p_{Y,i})^2}{\bar{p}_i (1 - \bar{p}_i)} \cdot \tag{10}$$

2.1.2. Angular distances

These distances are defined on the basis of the cosine of the angle θ between the two vectors **X** and **Y**.

Nei [46,47,49] reformulated $\cos\theta$ as the normalised identity I between the two populations and derived its standard genetic distance from the logarithm of $\cos\theta$

$$D_{\mathrm{S}} = -\log \frac{j_{XY}}{\sqrt{j_X j_Y}} = -\log I. \tag{11}$$

It is noteworthy that D_{m} is turned into D_{S} after a logarithm transformation of the gene identity in (4).

With the square root of allele frequencies, which then have a unity norm, the cosine of θ can be rewritten as $\cos\theta_{\mathrm{EC}} = \sum_i \sqrt{p_{X,i} p_{Y,i}}$. Edwards and Cavalli-Sforza [5,6,12,13] defined D_{c}, the chord distance, and f_θ respectively as:

$$D_{\mathrm{c}} = Cste \sqrt{1 - \cos_{\mathrm{EC}}\theta} \tag{12}$$

$$f_\theta = 4 \frac{1 - \sum_i \sqrt{p_{X,i} p_{Y,i}}}{k - 1} \cdot \tag{13}$$

The values of *Cste* set the function support of chord distances (when $Cste = 1$, D_{c} varies from 0 to 1).

Since the number of rare alleles increases with the number of sampled individuals, f_θ underestimates the expected genetic differentiation that would

be obtained with an increased sample size [51]. For this reason, Nei advises using a corrected distance D_A (equal to the square of D_c for $Cste = 1$):

$$D_A = \left(1 - \sum_i \sqrt{p_{X,i} p_{Y,i}}\right) = \frac{k-1}{4} f_\theta. \tag{14}$$

2.2. Distances based on allele size distributions

We also consider genetic distances expressed with respect to the moments of allelic size distributions of markers exhibiting length polymorphism.

Denote by i and j the repeat numbers of alleles i and j respectively. Goldstein [20], derived a distance from the Average Square Difference between populations, D_1

$$D_1 = \sum_{i,j} p_{X,i} p_{Y,j} (i-j)^2 = (\mu_X - \mu_Y)^2 + V_X + V_Y \tag{15}$$

with μ_X, μ_Y, V_X and V_Y, the means and variances in allelic sizes within populations.

Denote by $\varphi_{i,j}$ a function of the difference $i - j$ (null when $i = j$ and > 0 otherwise). Introducing $\varphi_{i,j}$ in D_m (4) gives

$$\sum_{i,j} p_{X,i} p_{Y,j} \varphi_{i,j} - \frac{1}{2}\left(\sum_{i,j} p_{X,i} p_{X,j} \varphi_{i,j} + \sum_{i,j} p_{Y,i} p_{Y,j} \varphi_{i,j}\right). \tag{16}$$

The within population Average Square Difference $D_{0,X}$ is defined by $\sum_{i,j} p_{X,i} p_{X,j} (i-j)^2$ (idem for population Y) and is equal to $2V_X$. Then, equation (16) in which φ_{ij} is set to $(i-j)^2$ may be rewritten as the squared difference between the allele size means $(\mu_X - \mu_Y)^2$, usually called $(\delta\mu)^2$, the distance of Goldstein [21].

The D_{SW} distance of Shriver [62] may be computed with (16) setting φ_{ij} equal to $|i-j|$.

Slatkin [63,64] argues to use D_1, $D_{0,X}$ and $D_{0,Y}$ in order to extend the G_{ST} calculation to length polymorphism

$$R_{ST} = \frac{D_1 - \bar{D}_0}{D_1 + \bar{D}_0} \tag{17}$$

with $\bar{D}_0 = \frac{1}{2}(D_{0,X} + D_{0,Y})$ [44].

2.3. Multiple loci

In practice, the estimation of distances is performed using the arithmetic mean over L loci.

Nevertheless, when at least one locus is fixed for the same allele in X and Y, D_R is undefined. So Latter [30] advises to use D_L computed as follows (PHYLIP package, [17])

$$D_L = \frac{\sum_\ell \sum_i (p_{X,\ell,i} - p_{Y,\ell,i})^2}{\sum_\ell (1 - \sum_i p_{X,\ell,i} p_{Y,\ell,i})}. \tag{18}$$

When at least one locus exhibits no allele shared between populations, the logarithm transformation $\log I$ is undefined ($I = 0$). So Nei advises rather to compute D_S with the arithmetic mean of gene identities

$$D_S = \frac{\sum_\ell j_{XY,\ell}}{\sqrt{\sum_\ell j_{X,\ell} \sum_\ell j_{Y,\ell}}}. \tag{19}$$

It is noteworthy that after removing loci with no shared alleles, taking the arithmetic mean of (11) (which is equivalent to using the geometric mean $\frac{1}{L} \prod_\ell j_\ell^{\frac{1}{L}}$) gives the maximum distance D_M of Nei [46]. Due to rare alleles within samples, the arithmetic mean of (11) is generally higher than (19).

Unbiased estimates of D_m called $\hat{D}_m$ (and derived distances), D_S called $\hat{D}_S$, (expectation of $\hat{D}_S$ is shown in Appendix A) and distances taking allelic sizes into account are computable with sampled allele frequencies x_i and y_i using an unbiased estimation of the within and between population gene identity [49]. The bias correction of $\widehat{\chi^2}$ given in [19] is also relevant for $\hat{D}_B$. So for the sake of simplicity, the expectations of distances under divergence models were computed assuming that true frequencies were known.

3. GENETIC DISTANCES UNDER GENETIC DRIFT AND MUTATION

The standard assumption that both derived populations, as well as the founder population, are in a mutation-drift equilibrium, implies that population divergence is due to the appearance of new mutants within populations. So distances can be used from a phylogenetic point of view, as estimators of divergence time.

3.1. Infinite allele mutation model

Due to the large number of variations a gene may theoretically exhibit, the number of possible new mutants is expected to be very large. The most appropriate mutation model for such markers is the infinite allele mutation model, *IAM* [28,38,65].

In this model, D_S is turned into a linear function of divergence time t and mutation rate β of markers:

$$E[D_{S(t)}] = 2\beta t. \tag{20}$$

Nei [45,46,49] advises to use D_S in order to construct phylogeny for closely related as well as for largely diverged populations. In contrast, the *IAM* expectation of D_m, exhibiting a finite maximal value, given the founder gene identity $j_{(0)}$ [51] is:

$$E(D_m) \approx j_{(0)}(1 - e^{-2\beta t}). \quad (21)$$

Derived distances (equations 5 to 10) as well as f_θ, D_c and D_A are not linear for all t values. Their behaviour (underestimation of divergence when t increases) disturbs their ability to distinguish a branching pattern between largely diverged populations. But for small divergence ($\beta t \ll 1$) they can be considered as quasi-linear functions of t. In addition γ_L, being independent of founder allele distributions, has the desirable advantage of being directly linked to the divergence time (expectation close to $2\beta t$ [31]).

Nevertheless, Takesaki and Nei [66] by simulations showed that D_S, exhibiting a larger variance than the non-linear distances, D_c or D_A, provides few correct tree topologies between populations within species.

Divergence is governed by βt implying that for a small divergence time, differences between populations measured with gene polymorphism and their confirmed low mutability (mutation rate of the α and β chains of insulin is estimated to be 10^{-7}/codon/generation, [48]) are expected to be small. The values of D_S are generally less than 0.01 or 0.02 between local breeds or subspecies [48]. So from a phylogenetic point of view assuming divergence by mutation, markers with a high mutability should enhance the precision of distance estimations for closely related populations. It was shown by Takesaki and Nei [66], *via* computer simulations, that markers with microsatellite characteristics give as many correct phylogeny when $t = 400$ as markers with low mutability when $t = 40\,000$.

3.2. Stepwise mutation model

Using microsatellites implies considering the Stepwise Mutation Model, SMM, [7,10,15,20,21,29,41,52,61,62,68] in which an allele carrying i repetitions can mutate to an allele carrying $j = i \pm 1$ repetitions. Due to reverse mutations yielding homoplasy phenomena [14], the expectation of D_S shows a great deviation from linearity [20,35], and therefore disturbs the phylogenetic reconstruction especially for large t values.

Shriver [62], Goldstein [20,21], Slatkin [64] and many others have developed linear statistics assuming infinite numbers of possible allelic scores. As D_1 and R_{ST} depend on the effective founder size, they are sensitive to bottlenecks and are not suited to deriving phylogenies [20,44].

Since under the assumption of an equilibrium between drift and mutation, the variance of allelic size converges [20,41,64], the growth of D_1 is only due

to the linear growth of the squared difference between the means (15) [21]:

$$E[(\delta\mu)_t^2] = 2\beta t. \tag{22}$$

Although there is no explicit formulae, Shriver [62] and Takesaki and Nei [66] showed by simulations that D_{SW} increases almost linearly (until 10 000 generations with $\beta = 0.0003$) with a slope different from 2β.

It is noteworthy that assuming alleles can mutate for more than 1 repeat, a generalised equation can be easily obtained substituting β by $\bar{w} = \frac{1}{L}\sum_\ell w_\ell$ [74] with $w_\ell = \beta_\ell \sigma_\ell^2$, when σ_ℓ^2 is the variance of the change in the number of repeats [64].

Between very closely related populations, Takesaki and Nei [66] by simulations showed that $(\delta\mu)^2$ and D_{SW} provide tree topologies of lower accuracy than non-linear distances (D_c or D_A). The dramatically bad results obtained with these statistics specifically developed for microsatellite evolution applications are due to their large variance. The coefficient of variation CV of $(\delta\mu)^2$, taking both biases and variance into account, is almost constant (distances exhibit linear standard deviation, [36,55,74]) and 5 times higher than those of non-linear distances. The CV of D_{SW} dramatically increases when t decreases with the consequence that these distances are the least appropriate for the estimation of phylogeny between breeds.

When the level of divergence increases, the efficiency of non-linear distances decreases (as predicted by theory) but they remain, however, the best methods to use with highly polymorphic markers [66].

3.3. Range constraints for microsatellites

Due to their high mutability, microsatellites are less convenient for the study of largely diverged groups. Takesaki and Nei [66] demonstrate that microsatellites perform better for $t = 400$ than for $t = 4\,000$. In [3], the tree between four species of primate (human, gorilla, chimpanzee and orang-utan) does not show any structure. The number of possible repeat scores converge to a maximum, denoted by R [3,20], with the consequence that $(\delta\mu)^2$ tends to a maximal value

$$\lim_{t\to\infty} (\delta\mu)^2 = \frac{R^2-1}{6} - 4(2N-1)\beta\left(1-\frac{1}{R}\right).$$

"As a consequence, mutation may be viewed as a homogenising factor" [44]. Feldman [16] and Pollock [55] propose linear corrections of $(\delta\mu)^2$ and more recently, Zhivotovsky [74] defines another linear statistics.

These distances introduced in order to improve estimation of large divergence times will not be described in more detail. Between closely related populations, they keep the same large variance suggesting that they are as inappropriate as D_{SW} and $(\delta\mu)^2$.

4. GENETIC DISTANCES UNDER GENETIC DRIFT

Focusing on the very early stages of evolution of populations allows us to consider that mutations can be neglected. As a consequence, fluctuations of allele frequencies are only due to genetic drift. Within populations, the genetic drift tends to reduce the genetic variability whereas differential loss of genes generates genetic diversity between populations.

In a diversity study of endangered breeds it is desirable to use distances which can be expressed as a function of the loss of the within population diversity. We will introduce the *Wright-Malécot* inbreeding coefficient in the calculus of drift expectation and variance of distances according to:

$$E(p_{X,i}) = p_{0,i}$$
$$E(p_{X,i}^2) = \Delta F p_{0,i} + (1 - \Delta F) p_{0,i}^2.$$

For the sake of simplicity, ΔF, the variation during t generations of the inbreeding coefficient from the founder population, which is equal to $1 - (1 - 1/2N)^t$, will be noted F with a subscript giving the name of the population, (F_X and F_Y for populations X and Y respectively) and called the inbreeding coefficient.

The drift expectation of the minimum distance of Nei,

$$E(D_{\mathrm{m}}) = \bar{F}(1 - \sum_i p_{0,i}^2) = \bar{F}(1 - h_0), \tag{23}$$

depends on $\bar{F} = (F_X + F_Y)/2$, the average inbreeding coefficient (between populations) and on h_0, the homozygosity of the founder population. For a small divergence, the drift expectation of D_{S} calculated with a Taylor expansion, in which F_X^2, F_Y^2 and $F_X F_Y$ can be neglected is:

$$E(D_{\mathrm{S}}) \approx -\log\left(\frac{1}{\sqrt{(1 - 2\bar{F}) + \dfrac{2\bar{F}}{h_0}}}\right) + \left(\sum_i p_{0,i}^3 - (h_0)^2\right) \times \left[\frac{\bar{F}}{(h_0)^2} - \frac{F_X}{\big(h_0 + F_X(1 - h_0)\big)^2} - \frac{F_Y}{\big(h_0 + F_Y(1 - h_0)\big)^2}\right]. \tag{24}$$

In parallel, taking the limit of the general solution of recurrence of $(\delta\mu)^2$ when the mutation rate tends to 0, allows this distance to be equal to

$$\lim_{\beta \to 0} E[(\delta\mu)_t^2] = \left[1 - \left(1 - \frac{1}{2N_X}\right)^t\right] V_0 + \left[1 - \left(1 - \frac{1}{2N_Y}\right)^t\right] V_0 = 2\bar{F} V_0 \tag{25}$$

with V_0 the variance of allelic size in the founder population.

4.1. Estimation of the average inbreeding coefficient $\bar{F}$

For phylogeny purposes, the authors wish to use distances depending on divergence time only. In the present section, we focus on the distances allowing us to estimate the level of genetic diversity by way of the average inbreeding coefficient $\bar{F}$. In Section 3.3, we will test their accuracy by way of computer simulations.

Distances like D_{m}, D_{S} or $(\delta\mu)^2$ depend on the founder population parameters, and therefore cannot be directly linked to $\bar{F}$. A strategy to obtain an estimate of the average inbreeding coefficient considering S populations was developed by Wright [72] and Nei [47,51]. The mean and variance of the frequency of allele i between subpopulations are denoted by $\bar{p}_i = \frac{1}{S}\sum_{\mathrm{s}} p_{s,i}$ and $Var_{\mathrm{s}}(p_{s,i})$ respectively. F_{ST}, initially defined for dimorphic loci as the sum of the between population variance of alleles 1 and 2 weighted by $H_{\mathrm{T}} = 2\bar{p}_1\bar{p}_2$, an estimation of the founder heterozygosity H_0 [72], was extended to polymorphic loci by Nei [47] as the weighted variance G_{ST} given by:

$$G_{\mathrm{ST}} = \frac{\sum_i Var_{\mathrm{s}}(p_{s,i})}{\sum_i \bar{p}_i(1-\bar{p}_i)}.$$

The drift expectations of the numerator and denominator expressed with respect to the inbreeding coefficient of every sub-population, F_{s}, are

$$\sum_i Var(p_{s,i}) = \left(1 - \sum_i p_{0,i}^2\right)\left(\frac{S-1}{S^2}\sum_{\mathrm{s}} F_{\mathrm{s}}\right)$$

$$E\left[\sum_i \bar{p}_i(1-\bar{p}_i)\right] = \left(1 - \sum_i p_{0,i}^2\right)\left(1 - \frac{1}{S^2}\sum_{\mathrm{s}} F_{\mathrm{s}}\right)$$

with $p_{0,i}$ the allele frequency of the founder population common to the s subpopulations. Assuming, as in Nei and Chakravarty [50], that the ratio of expectations is within the same order as the expectation of the ratio, gives

$$E[G_{\mathrm{ST}}] \approx \frac{S-1}{S^2}\,\frac{\sum_{\mathrm{s}} F_{\mathrm{s}}}{1 - \frac{1}{S}\bar{F}}. \tag{26}$$

When S is large, $E[G_{\mathrm{ST}}]$ is approximately equal to the average inbreeding coefficient $\bar{F} = \frac{1}{S}\sum_{\mathrm{s}} F_{\mathrm{s}}$.

4.1.1. Euclidean distances

Considering two populations and taking $2G_{\mathrm{ST}}$ gives

$$E[2G_{\mathrm{ST}}] \approx \bar{F} + \frac{\bar{F}^2}{2-\bar{F}}. \tag{27}$$

Unfortunately, because of the biased estimation of H_0 provided by $\sum_i \bar{p}_i(1-\bar{p}_i)$, the estimation of $\bar{F}$ is positively biased, especially when divergence increases.

This strategy was extended to other distances by Reynolds [57], Balakrishnan and Sangvhi [1] and Barker [2]. Given that $E(1-\sum_i p_{X,i}p_{Y,i}) = 1-\sum_i p_{0,i}^2$, the Reynold's distance,

$$E(D_{\mathrm{R}}) \approx \bar{F} \tag{28}$$

is unbiased whatever the level of inbreeding.

Dividing each square allele differences $(p_{X,i}-p_{Y,i})^2$ by $\bar{p}_i(1-\bar{p}_i)$ and k in Barker's method and $\bar{p}_i$ and $(k-1)$ in Sanghvi's method [19] allows a rather long and fastidious computation of their expectations for polymorphic loci. However for dimorphic loci, these distances together with $2G_{\mathrm{ST}}$ can be rewritten as

$$\frac{(p_{X,1}-p_{Y,2})^2}{\bar{p}_1\bar{p}_2} \tag{29}$$

and have the same expectation as in (27). For polymorphic loci with uniformly distributed founder frequencies $p_{0,i} \approx 1/k$, approximate calculus (expectation of a ratio is approximated by the ratio of expectations) giving

$$E\left(\frac{1}{k}D_{\mathrm{B}}\right) \approx \bar{F} + \frac{\bar{F}^2}{2-\bar{F}} \tag{30}$$

$$E\left(\frac{1}{k-1}\chi^2\right) \approx \bar{F} \tag{31}$$

shows that these distances might be used as estimators of $\bar{F}$.

4.1.2. Angular distances

Given that neglecting F_X^2, F_Y^2, F_XF_Y and assuming uniformly distributed founder frequencies $p_{0,i} \approx 1/k$

$$E\left[\sqrt{p_{X,i}p_{Y,i}}\right] \approx p_{0,i} - \frac{1}{4}\bar{F}(1-p_{0,i}), \tag{32}$$

the drift expectation of f_θ calculated with the Taylor expansion is

$$E[f_\theta] \approx \bar{F}\frac{1}{k-1}\sum_i (1-p_{0,i}). \tag{33}$$

Rearranging (33) gives

$$E[f_\theta] \approx \bar{F}. \tag{34}$$

The distance f_θ, considered as nearly unbiased for small $\bar{F}$, will be biased when the number of alleles and the population divergence increases (for example when $\bar{F}$ is large, a term depending on $F_X F_Y$, which is equal to $-\frac{1}{16}F_X F_Y(k-1)$, cannot be neglected longer).

In the present work we focused on f_θ rather than D_A which was no longer directly linked to the inbreeding coefficient (its expectation can be directly deduced from (33) ignoring $4/(k-1)$). As a consequence, the chord distances equal to the square root of D_A were not kept for further analysis.

4.2. Variance of unbiased estimates of D_R

Variance of $\widehat{G_\mathrm{ST}}$ was given in Nei and Chakravarty [50]. Foulley and Hill [19], compute the variance of $\widehat{\chi^2}$, assuming Gaussian distribution of true allele frequencies and equal sample sizes, $m_{X,\bullet} = m_{Y,\bullet} = m$.

In this paper, approximate standard deviation of $\hat{D}_\mathrm{m}$ and $\widehat{D}_\mathrm{R}$ corrected for sample size were computed under drift divergence assuming $F_X \neq F_Y$ and $m_{X,\bullet} \neq m_{Y,\bullet}$ (Appendix B). In order to provide understandable formulas, approximated standard deviations may be easily rewritten assuming L independent loci, each one exhibiting k_0 uniformly distributed founder frequencies ($p_{0,\ell,i} = 1/k_{0,\ell}$ and $k_{0,1} = k_{0,\ell} = k_{0,L} = k_0$):

$$\sigma(\hat{D}_\mathrm{m}) \approx \sqrt{\frac{2}{L}\frac{(k_0-1)}{k_0^2}}\left(\bar{F} + \left(\frac{1}{2m_{X,\bullet}} + \frac{1}{2m_{Y,\bullet}}\right)\right) \tag{35}$$

$$\sigma(\hat{D}_\mathrm{R}) \approx \sqrt{\frac{2}{L(k_0-1)}}\left(\bar{F} + \left(\frac{1}{2m_{X,\bullet}} + \frac{1}{2m_{Y,\bullet}}\right)\right). \tag{36}$$

In the following section the validity of the approximated formulae (36) will be checked by way of computer simulations.

4.3. Comparison of several estimators of $\bar{F}$

The accuracy of distances estimating $\bar{F}$ was compared by computer simulations performed under pure genetic drift divergence of two isolated populations X and Y.

4.3.1. Simulation procedure

The change in allele frequencies between two generations was simulated as a Multinomial sampling scheme according to the *Wright-Ficher* model of population evolution. Twenty genetically independent loci were considered, a number frequently found in diversity studies [33, 37, 40].

The founder frequencies of the founder population of X and Y were generated as follows. An initial simulated population of size $N = 500$ was first considered, with allele frequencies $p_{00,i}$ (for $i = 1, \ldots, k$), was submitted 1 000 times to a genetic drift process during five generations. This process generates 1 000 quasi-independent populations used as starting points of simulation runs. Each one of these 1 000 populations, described by its founder frequencies, $\mathbf{p_0}$, was submitted to a pure genetic drift divergence generating the populations X and Y, which have constant diploid effective sizes equal to $N = 100$ and $N = 400$ respectively during 22 non-overlapping generations.

In order to provide estimations of increasing values of $\bar{F}$ (ranging from 0.025 to 0.3), gene samplings ($m_{X,\bullet} = m_{Y,\bullet} = 50$ genes) were computed every five generations from the divergence.

4.3.2. Results

The performances of the F-estimates established using the following statistics averaged over 1 000 replications, the relative bias B_r (expressed in percent of the true value of $\bar{F}$), the standard error SE and the squared root of the mean square error $\sqrt{MSE} = \sqrt{bias^2 + SE^2}$ are presented in Figures 1, 2 and 3 respectively.

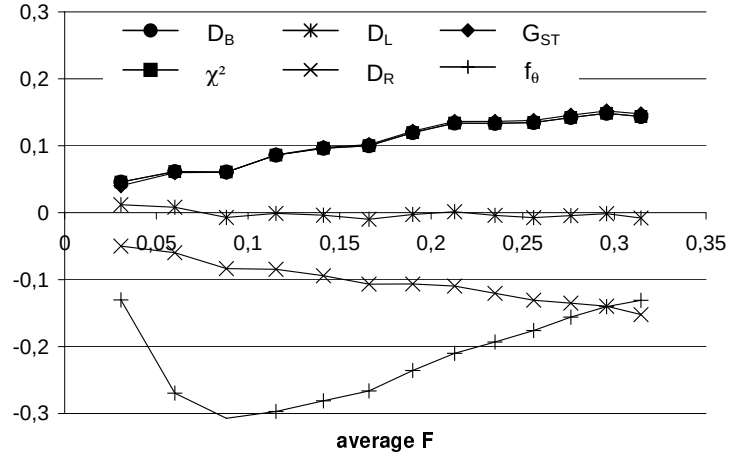
Uniform founder frequencies

Two sets of 1 000 simulations, in which allele frequencies of the initial population were set to $p_{00,i} = 1/k$, were performed with $k = 2$ and $k = 8$ alleles. Estimations of $\hat{G}_{ST}$, $\hat{D}_R$, $\hat{D}_B$ and $\widehat{\chi^2}$ – corrected for sample sizes – were performed using the arithmetic mean across loci. We also introduce the distance of Latter $\hat{D}_L$ [30], equation (18), and $\hat{f}_\theta$.

Relative bias (Fig. 1): As expected, with two (Fig. 1a) or eight (Fig. 1b) alleles per locus, $\hat{G}_{ST}$ exhibits a positive bias, this increases with the level of divergence (this bias is well predicted by equation (27)). By contrast, $\widehat{\chi^2}$ expected to be unbiased (31) and $\hat{D}_B$ expected to be of the order of magnitude of $\hat{G}_{ST}$ (30), are negatively biased as $\hat{f}_\theta$. In parallel $\hat{D}_L$ and $\hat{D}_R$ are the least biased distances (constant bias whatever the divergence level) for diallelic or more polymorphic loci. It is noteworthy that estimations given by $\hat{D}_L$ (weighted by estimates of founder heterozygosity computed with all loci) provide lower bias than estimations given by $\hat{D}_R$ (weighted for each locus by an estimate of founder heterozygosity).

Standard deviation (Fig. 2): With two alleles per locus (Fig. 2a), the Reynolds distance exhibits the smallest standard error when $\bar{F}$ increases. Otherwise, with eight alleles per loci (Fig. 2b) $\hat{f}_\theta$, $\hat{D}_B$ and $\widehat{\chi^2}$ show the smallest standard errors. The strait line computed from (36) shows the validity of the approximated standard error neglecting power of F higher than 2 (as expected, formula

a



b

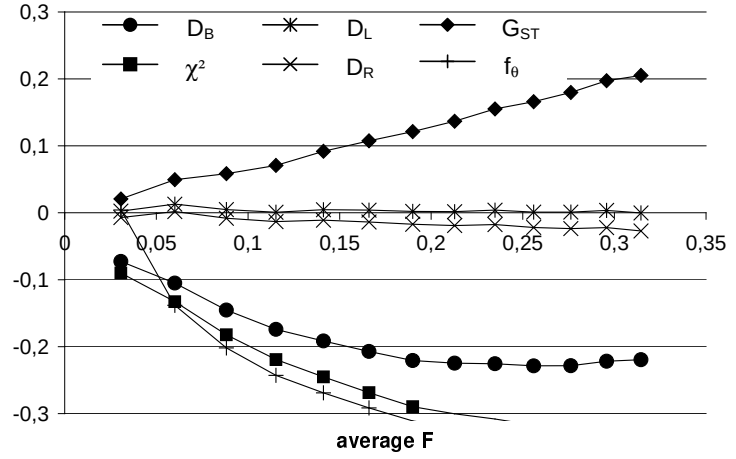
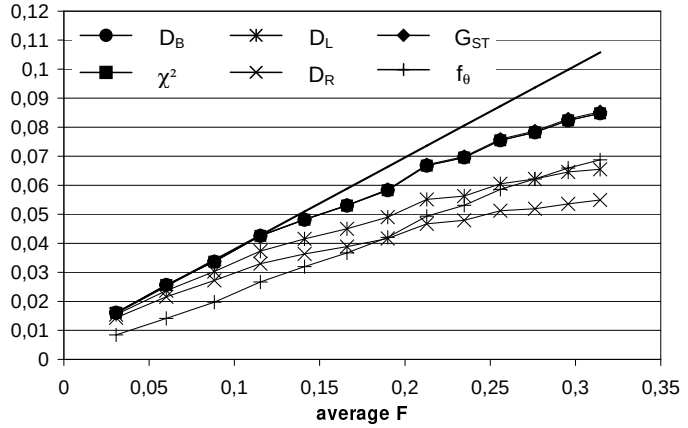


Figure 1. Relative biases of distances as a function of the increase of the average inbreeding coefficient $\bar{F}$.

The estimations of $\bar{F}$ were computed over 20 loci and 1000 replications performed with two populations with effective sizes $N = 100$ and $N = 400$ respectively evolving during 22 non-overlapping generations. The sample sizes, drawn every two generations, are set to 25 individuals. The distances D_B, χ^2 and G_{ST} are plotted with black circles, squares and lozenges respectively. The distances D_R and D_L are plotted with crosses and stars respectively. The distance f_θ is plotted with plus symbols. Part (**a**) shows the results obtained with the diallelic markers. In this case the distances D_B and χ^2 give identical numerical results. Part (**b**) shows the results obtained with the markers exhibiting eight alleles.

a



b

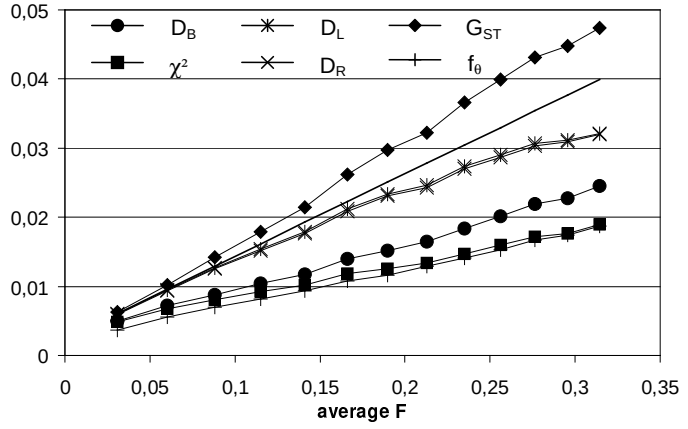
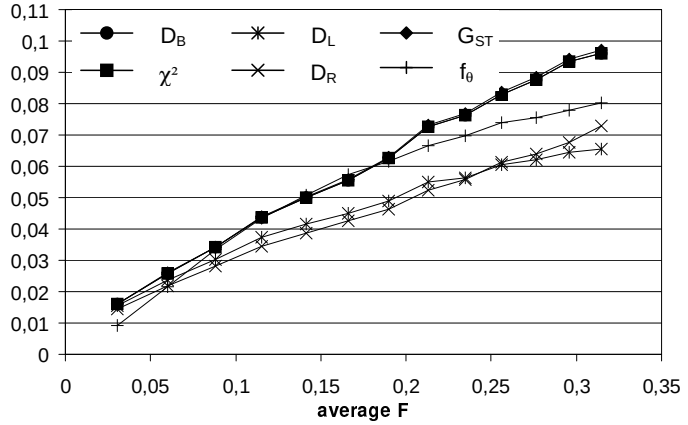


Figure 2. Standard errors of distances as a function of the increase of the average inbreeding coefficient $\bar{F}$.
In this figure we kept the same symbols as in Figure 1. The strait line was computed with the expected value of standard deviations (equation (36)). Part (**a**) shows the results obtained with the diallelic markers. In this case the distances D_B and χ^2 give identical numerical results. Part (**b**) shows the results obtained with the markers exhibiting eight alleles.

(36) is a better approximation for small $\bar{F}$, lower than 0.15, than for large $\bar{F}$). The deviation from the expected value of the standard errors of $\hat{D}_B$ and $\hat{\chi}^2$ (for small and large $\bar{F}$) is certainly due to their large negative biases allowing the variance of estimation to be decreased.

a



b

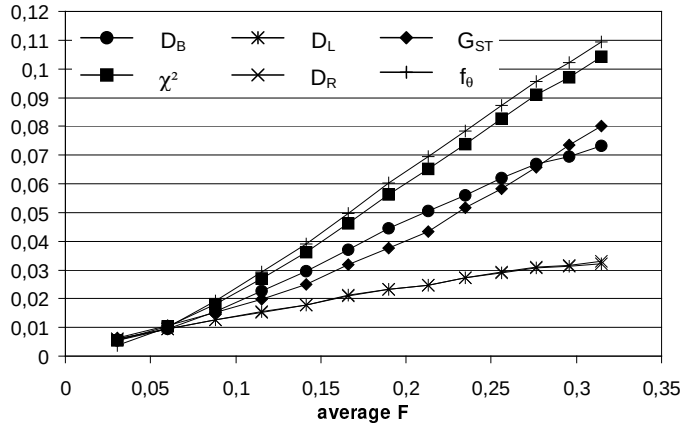


Figure 3. Square root of the mean square errors of distances as a function of the increase of the average inbreeding coefficient $\bar{F}$.
In this figure we kept the same symbols as in Figure 1. Part (**a**) shows the results obtained with the markers exhibiting two alleles. In this case the distances D_B and χ^2 gave identical numerical results. Part (**b**) shows the results obtained with the markers exhibiting eight alleles.

Mean square error (Fig. 3): When the bias is rather small with respect to the standard error, $\sqrt{MSE}$ is expected to be close to the standard error. With two alleles per loci the method with the smallest standard error $\hat{D}_R$ and $\hat{D}_L$ give the smallest $\sqrt{MSE}$ whatever the value of the inbreeding coefficient. With eight alleles per locus and when the level of divergence increases, methods

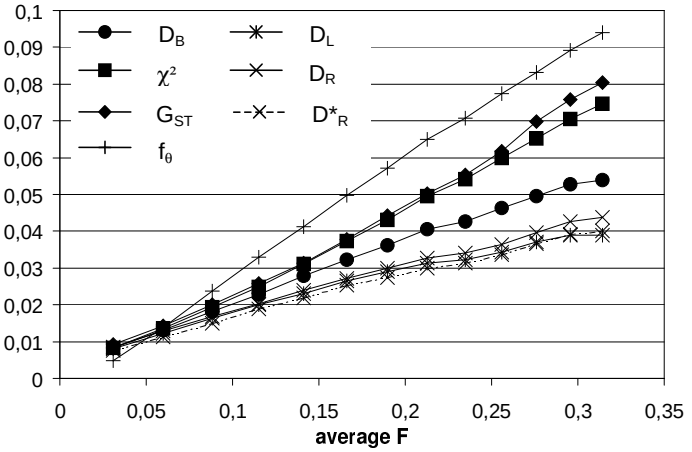


Figure 4. Square root of the mean square errors of distances computed with microsatellites exhibiting different allele numbers.
The estimations of $\bar{F}$ were computed over 20 microsatellites and 1000 replications performed with two populations with effective sizes $N = 100$ and $N = 400$ respectively evolving during 22 non-overlapping generations. The sample sizes, drawn every two generations, are set to 25 individuals. In this figure we kept the same symbols as in Figure 1. The distance D_{R}^{*} (equation (37)) is plotted with dotted lines.

with the smallest biases ($\hat{D}_{\mathrm{R}}$ and $\hat{D}_{\mathrm{L}}$) give the smallest $\sqrt{MSE}$ although they do not exhibit the smallest standard errors. On the basis of an accuracy criterion combining the bias and the standard error of estimations, $\hat{D}_{\mathrm{R}}$ and $\hat{D}_{\mathrm{L}}$ are the most accurate distances whatever the polymorphism of the marker used.

Microsatellite founder frequencies

One set of 1 000 simulations was performed, in which allele frequencies $p_{00,i}$ in the initial populations were set to microsatellite marker frequencies published in [33]. The number of alleles varied between loci (the mean number of alleles is close to 6). In this case the distances D_{L} and D_{R} were still the most accurate methods considering the $\sqrt{MSE}$ criterion (Fig. 4, [34]).

On the basis of $\sqrt{MSE}$ we also compared the distance D_{L} and the distance D_{R} computed using the arithmetic mean over loci with another estimate of the distance D_{R} computed using the following formula [34]

$$\hat{D}_{\mathrm{R}}^{*} = \frac{\sum_{\ell} (n_{XY,\ell} - 1)\widehat{D}_{\ell}}{\sum_{\ell} (n_{XY,\ell} - 1)}. \tag{37}$$

This formula takes the heterogeneity of the marker polymorphism into account with $n_{XY,\ell}$ which is the number of alleles present both in the sample of X and Y.

In this case, the standard error of the weighted Reynolds distance is equal to

$$\sigma(\hat{D}^*_{\mathrm{R}}) \approx \sqrt{\frac{2}{\sum_\ell (k_{0,\ell} - 1)}} \left(\bar{F} + \left(\frac{1}{m_{X,\bullet}} + \frac{1}{m_{Y,\bullet}} \right) \right). \tag{38}$$

Using the weighted estimate did not yield a significant gain of accuracy. The $\sqrt{MSE}$ of D^*_{R} was nearly identical to the $\sqrt{MSE}$ of D_{L} (Fig. 4).

5. DISCUSSION AND CONCLUSION

Under the assumption of equilibrium between drift and mutation, the power of different distance estimation methods for constructing phylogenetic trees is well discussed in Takesaki and Nei [66]. Their work points out that the quest for linearity at the cost of variance is not an efficient strategy. Increasing functions of time (non-necessarily linear but with a slope large enough to discriminate closely related populations) with small variances provide correct phylogeny with higher levels of confidence than linear distances do. It is clear that with such distances the length of branches is not representative of divergence time. However, this question seems of minor importance with regards to that of a correct branching pattern. Perez-Lezaun [54] compared human populations using 20 microsatellite loci on the basis of D_{R}, R_{ST}, D_{SW} and $(\delta\mu)^2$. As expected, D_{R} gives trees with the highest bootstrap values and the best topology with regards to our knowledge of human history.

Goldstein and Pollock [22] argued that the misunderstanding of mutation processes also explains the poor efficiency of these distances. D_{SW} and $(\delta\mu)^2$ were defined assuming equal probabilities of insertion and deletion of repeats whereas observed microsatellite distributions clearly show evidence in favour of asymmetric mutation processes [27,73]. Taking the mutation process of microsatellites into account should be more efficient when using methods with a small variance such as likelihood based approaches, rather than for distances based on a simple difference between allele size means.

In the second section of the present work we assumed that for very closely related breeds the number of mutations cannot explain the observed genetic variation even when highly mutable DNA sequences are used. For populations of small size, $N = 50$, and a mutation rate of $\beta = 10^{-3}$, mutations can be neglected during 200 generations: the difference between the values of inbreeding coefficients computed assuming or neglecting mutation is small, being less than 7 percent of the true value [34].

The genetic drift allows genetic distances computed with allele frequencies to be strongly dependent on the number of generations since divergence, t, and on the value of the effective sizes of breeds, N_X and N_Y [43]. The values

of distances increases with the parameters $1-(1-1/2N_X)^t \approx t/2N_X$ and $1-(1-1/2N_Y)^t \approx t/2N_Y$ which represent the increase of the inbreeding coefficients during t generations. Since $t/2N_X$ can be viewed as the *evolution rate* in population X, no phylogeny can be inferred from the tree in cases of very closely related breeds exhibiting different effective sizes. Indeed the location on the tree of the most recent common ancestor cannot be exactly determined when evolution rates vary between lineages (*e.g.* when a bottleneck does occur within a breed). In order to infer the true history of populations, it is necessary to root the tree using an *outgroup*.

This work points out that, under the drift assumption, the major part of the genetic distances (the Nei distances D_m and D_S for example) also depends on unknown parameters, the founder frequencies. For example the expected value of the minimum distance of Nei depends on the heterozygosity H_0 of the founder population. With such a distance we cannot separate the effect of the genetic drift occurring in each population and the ancient history of the founder population. So this fact can also disturb the phylogeny reconstruction, mainly when migration or admixture does occur between founder populations.

As in [11], we privileged distances which can be expressed with the increase during t generations of the inbreeding coefficient alone (or equivalently the increase of the kinship coefficient). This parameter is of importance to analyse the genetic diversity of breeds. It allows us to measure the loss of the within population diversity due to the drift process [34]. Eding [11] argues that, in terms of kinship, a generic formula of distance can be written as $d(X,Y) = f_Y + f_Y - 2f_{XY} = \Delta f_X + \Delta f_Y$, with f_X, f_Y the within breeds kindship coefficient, f_{XY} the kindship coefficients between breeds and $\Delta f_X = F_X$, $\Delta f_Y = F_Y$ the increase since divergence of f_Y and f_Y respectively. $d(X,Y)/2$ is therefore equal to the average inbreeding coefficient $\bar{F}$. This shows that using the Reynolds distance is equivalent to using a distance giving a measure of the within breed diversity (f_X and f_Y) corrected by the between breed diversity (f_{XY}).

As a by product, this suggests an important fact when considering very closely related breeds. Since distances computed with allele frequencies of neutral markers are expressed as a function of the loss of the genetic diversity methods, such criteria as the Weitzman one [67,71] which advises conserving most of the diversity of the whole set by conserving the most distant breeds, are not appropriate in this case [34]. Indeed if we consider a set involving large populations and a totally inbred breed ($F = 1$) which has no original allele, the Weitzman approach will suggest conserving the inbred breed.

Although expected values of distances are quasi independent of the sampling process, a part of their standard deviation depends on sample size. From (36) $\sigma/\bar{F}$ is proportional to $1/m\bar{F}$, showing that when divergence is low, the accuracy of distances when building trees is sensitive to sample size. It is impossible to get accurate estimations when divergence tends to 0.

By contrast when the divergence increases the sample size does not make much differences in the accuracy of distance estimations. Therefore, for intermediate inbreeding values, the accuracy of distance estimations mainly depends on the number and on the degree of polymorphism of the markers used. The variance of distances is inversely proportional to the number of alleles per locus within the founder population. This strongly advocates in favour of the present use of markers such as microsatellites rather than gene polymorphism, which is expected to be less variable within populations.

Nevertheless, distances such as χ^2 or D_B are more biased with eight founder alleles than with two founder alleles. For such low polymorphism values, the bias of D_B, χ^2 and $2G_{ST}$ behaves as predicted by equation (27). The dependency of their biases on the value of inbreeding and on the number of founder alleles suggests that these distances are sensitive to rare alleles present within the founder and derived populations (the most frequently eliminated when the level of drift increases and forgotten when sample size is small).

The estimations computed with five loci and eight founder alleles show biases close to those observed with 20 loci (data not shown). For small $\bar{F}$ (between 0.03 and 0.1), the $\sqrt{MSE}$ are within the order of magnitude of the standard error making D_B and χ^2 slightly more accurate than the less biased distances D_L and D_R, whereas all distances show the same performances when the number of loci is equal to 20. For $\bar{F}$ higher than 0.1 and for a small number of loci as well as for a number of loci close to that observed in the literature [33, 40], more than 20, the conclusions are different. As shown by the difference of $\sqrt{MSE}$ with respect to the standard error as long as $\bar{F}$ increases, the reduction of the accuracy due to bias largely counterbalances the gain in variance due to the number of loci and high polymorphisms when we consider distances such as D_B or χ^2. This suggests that unbiased distances, such as D_L in all cases presented and D_R with high polymorphisms, should be privileged mainly when the number of markers used is larger than 20.

For $\bar{F}$ higher than 0.3, D_L and D_R should behave quite better than the other distances, mainly when the polymorphism of markers is high (microsatellites and eight alleles per locus, data not shown).

The weighted estimate of the Reynolds distance (37), taking the difference between the number of alleles observed into account, do not give a significant gain in accuracy. This formula is deduced from the expected standard deviation of the Reynolds distance (36) which depends on $k_{0,\ell}$ the number of alleles within the founder population. When this number is approximately known (for example when a sample of the founder population is available), using the weighted estimate of the Reynolds distance computed between the founder and the derived population X yields an important gain in accuracy [34]. Since the founder alleles can be lost because of the genetic drift process $n_{XY,\ell}$ is a bad estimator of $k_{0,\ell}$ as far as the inbreeding coefficient increases.

To conclude this work it seems that, among distances estimating $\bar{F}$ when drift is assumed, the Latter and Reynolds distances (D_L and D_R) have to be privileged whatever the polymorphism of markers used. It is necessary to keep in mind that, because of the drift process, the obtained trees do not represent true phylogenetic relationships when the effective sizes are different between breeds. Since the distances depend on the increase of the inbreeding coefficient of each breed, F_X and F_Y [11,34], these trees can be viewed as a representation of the loss of the within breed genetic diversity due to the genetic drift process.

However F_X and F_Y can be separately estimated using a statistics directly derived from the Reynolds distance [69] or using a more accurate method based on a Monte Carlo Markov Chain algorithm [34]. Since all $t/2N$ can be measured in all couples of breeds by these approaches, new methods allowing to locate the most recent common ancestor on trees, and therefore to retrieve the true evolutionnary relationships when no outgroup is available, could be proposed.

ACKNOWLEDGEMENTS

We thank Jean-Marie Cornuet and John William James for motivating remarks and Alain Vignal for the English revision of the manuscript.

REFERENCES

[1] Balakrishnan V., Sanghvi L.D., Distance between populations on the basis of attribute data, Biometrics 24 (1968) 859–865.

[2] Barker J.S.F., Hill W.G., Bradley D., Nei M., Fries R., Wayne R.K., Measurement of domestical animal diversity (MoDAD): Original working group report, FAO, Rome, 1998.

[3] Bowcock A.M., Ruiz-Linares A., Tomfohrde J., Minch E., Kidd J.R., Cavalli-Sforza L.L., High resolution of human evolutionary trees with polymorphic microsatellites, Nature 368 (1994) 455–457.

[4] Buchanan F.C., Adams L.J., Littlejohn R.P., Maddox J.F., Crawford A.M., Determination of evolutionary relationships among sheep breeds using microsatellites, Genomics 22 (1994) 397–403.

[5] Cavalli-Sforza L., Edwards A.W.F., Phylogenetic analysis models and estimation procedure, Evolution 21 (1967) 550–570.

[6] Cavalli-Sforza L.L., Zonta L.A., Nuzzo F., Bernini L., De Jong W.W.W., Meera Khan P., Ray A.K., Went L.N., Siniscalco M., Nijenhuis L.E., Van Loghem E., Modiano G., Studies on African pygmies. I. A pilot investigation of Babinga pygmies in the central Africa republic (with an analysis of genetic distances), Am. J. Hum. Genet. 21 (1969) 252–274.

[7] Chakravarthy R., Nei M., Bottleneck effects on average heterozygosity and genetic distances with the stepwise mutation model, Evolution 31 (1977) 347–356.

[8] Chevalet C., Gillois M., Valeurs approchées des coefficients d'identité dans les populations panmictiques, Lecture Notes in Biomathematics, Modèles Mathématiques en Biologie 41 (1978) 128–136.

[9] Chiampolini R., Moazami-Goudarzi K., Vaiman D., Dillman C., Mazzanti E., Foulley J.-L., Leveziel H., Cianci D., Individual multilocus genotypes using microsatellites polymorphisms to permit the analysis of the genetic variability within and between Itatlian beef cattle breeds, J. Anim. Sci. 73 (1995) 3259–3268.

[10] Di Rienzo A., Peterson A.C., Garza J.C., Valdes A.M., Slatkin M., Freimer N.B., Mutational processes of simple-sequence repeat loci in human populations, Proc. Natl. Acad. Sci. USA 91 (1994) 3166–3170.

[11] Eding H., Meuwissen T.H.E., Marker-based estimates of between and within population kinships for the conservation of genetic diversity, J. Anim. Breed. Genet. 118 (2001) 141–159.

[12] Edwards A.W.F., Distances between populations on the basis of gene frequencies, Biometrics 27 (1971) 873–881.

[13] Edwards A.W.F., Cavalli-Sforza L.L., Reconstruction of evolutionary trees, in: Phenetic and Phylogenetic classification, Systematics Association, London 6 1964 pp. 67–76.

[14] Estoup A., Tailliez C., Cornuet J.M., Solignac M., Size homoplasy and mutationnal process of interrupted microsatellites in two bee species, *Apis mellifera* and *Bombus terrestris* (Apidae), Mol. Biol. Evol. 12 (1995a) 1074–1084.

[15] Estoup A., Garnery L., Solignac M., Cornuet J.M., Microsatellite variation in honey bee (*Apis mellifera* L.) populations: hierarchical genetic structure and test of the infinite allele and stepwise mutation models, Genetics 140 (1995b) 679–695.

[16] Feldman M.W., Bergman A., Pollock D.D., Goldstein D.B., Microsatellite genetic distances with range constraints: analytic description and problems of estimation, Genetics 145 (1997) 207–216.

[17] Felsenstein J., PHYLIP (Phylogeny Inference Package) Version 3.5, Departement of genetics, University of Washington, Seattle, 1993.

[18] Forbes H.S., Hogg J.T., Buchanan F.C., Crawford A.M., Allendorf F.W., Microsatellite evolution in congeneric mammals: domestic and bighorn sheep, Mol. Biol. Evol. 16 (1995) 1106–1113.

[19] Foulley J.-L., Hill W.G., On the precision of estimation of genetic distance, Genet. Sel. Evol. 31 (1999) 457–464.

[20] Goldstein D.B., Linares A.R., Feldman M.W., An evaluation of genetic distances for use with microsatellite loci, Genetics 139 (1995a) 463–471.

[21] Goldstein D.B., Linares A.R., Cavalli-Sforza L.L., Feldman M.W., Genetic absolute dating based on microsatellites and the origin of modern humans, Proc. Natl. Acad. Sci. USA 92 (1995b) 6723–6727.

[22] Goldstein D.B., Pollock D.D., Launching microsatellites: a review of mutation processes and methods of hylogenetic inference, J. Hered. 88 (1997) 335–342.

[23] Goodman M.M., Genetic distances: measuring dissimilarity among populations, Yearbook of physical anthropology 17 (1973) 1–38.

[24] Gottelli D., Sillero-Zubiri C., Applebaum G.D., Roy M.S., Girman D.J., Garcia-Moreno J., Ostrander E.A., Wayne R.K., Molecular genetics of the most

endangered canid: the Ethiopian wolf *Canis simiensis*, Mol. Ecol. 3 (1994) 301–312.
[25] Gower J.C., Measures of taxonomic distances between populations based on gene frequencies, in: The assessment of population affinities in man, J.S. Weiner and J. Huizing Edition, Oxford University Press, 1972.
[26] Gregorius H.R., On the concept of genetic distances between populations based on gene frequencies, Proceeding, Joint IUFRO Meeting, S. 02.04.1–3, Stockholm, Session I, 17–26, 1974.
[27] Jin L., Macaubas C., Hallmayer J., Kimura A., Mignot E., Mutation rate varies among alleles at a microsatellite loci: phylogenetic evidence, Proc. Natl. Acad. Sci. USA 93 (1996) 15285–15288.
[28] Kimura M., Crow J.F., The number of alleles that can be maintained in a finite population, Genetics 49 (1964) 725–738.
[29] Kimura M., Ohta T., Stepwise mutation model and distribution of allelic frequencies in a finite population, Proc. Natl. Acad. Sci. USA 75 (1978) 2868–2872.
[30] Latter B.D.H., The island model of population differentiation: a general solution, Genetics 73 (1972a) 147–157.
[31] Latter B.D.H., Selection in finite populations with multiple alleles. III. Genetic divergence with centripetal selection and mutation, Genetics 70 (1972b) 475–490.
[32] Latter B.D.H., The estimation of genetic divergence between populations based on gene frequency data, Amer. J. Hum. Genet. 25 (1973) 247–261.
[33] Laval G., Iannuccelli N., Legault C., Milan D., Groenen M.A.M., Giuffra E., Andersson L., Nissen P.H., Jorgensen C.B., Beeckman P., Geldermann H., Foulley J.-L., Chevalet C., Ollivier L., Genetic diversity of eleven European pig breeds, Genet. Sel. Evol. 32 (2000) 187–203.
[34] Laval G., Éléments de choix des marqueurs et des méthodes dans l'analyse de la diversité génétique intra spécifique : cas des races animales domestiques, Institut national agronomique Paris-Grignon Ph.D. thesis, 2001.
[35] Li W.H., Simple method for constructing phylogenetic trees from distances matrix, Proc. Natl. Acad. Sci. USA 78 (1981) 1085–1089.
[36] Li W.H., Nei M., Drift variances of heterozygosity and genetic distance in transient states, Genet. Res. Camb. 25 (1975) 229-248.
[37] MacHugh D.E., Loftus R.T., Cunninghanm P., Bradley D.G., Genetic structure of seven European cattle breeds assessed using 20 microsatellite markers, Anim. Genet. 29 (1998) 333–340.
[38] Malécot G., La consanguinité dans une population limitée, C. R. Acad. Sci. Paris 222 (1946) 841–843.
[39] Malécot G., Les mathématiques de l'hérédité, Masson, Paris, 1948.
[40] Moazami-Goudarzi K., Laloë D., Furet J.P., Grosclaude F., Analysis of genetic relationships between 10 cattle breeds with 17 microsatellites, Anim. Genet. 28 (1997) 338–345.
[41] Moran P.A.P., Wandering distributions and the electrophoretic profile, Theor. Popul. Biol. 8 (1975) 318–330.
[42] Morton N.E., Yee S., Harris D.E., Lew R., Bioassay of kindship, Theor. Popul. Biol. 2 (1971) 507–524.
[43] Nagamine Y., Higuchi M., Genetic distance and classification of domestic animals using genetic markers, J. Anim. Breed. Genet. 118 (2001) 101–109.

[44] Nauta M.J., Weissing F.J., Constraints on allele size at microsatellite loci: implication for genetic differentiation, Genetics 143 (1996) 1021–1032.

[45] Nei M., Interspecific gene differences and evolutionary time estimated from electrophoretic data on protein identity, Am. Naturalist 105 (1971) 385–398.

[46] Nei M., Genetic distance between populations, Am. Naturalist 106 (1972) 283–292.

[47] Nei M., Analysis of gene diversity in subdivided populations, Proc. Natl. Acad. Sci. USA 70 (1973) 3321–3323.

[48] Nei M., Molecular population genetics and evolution, North-Holland Publishing Company, Amsterdam, Oxford, 1975.

[49] Nei M., Estimation of average heterozygosity and genetic distance from a small number of individuals, Genetics 89 (1978) 583–590.

[50] Nei M., Chakravarti A., Drift variance of FST and GST statistic obtained from a finite number of isolated populations, Theor. Popul. Biol. 11 (1977) 307–325.

[51] Nei M., Tajima F., Tateno Y., Accuracy of estimated phylogenetic trees from molecular data, J. Mol. Evol. 19 (1983) 153–170.

[52] Ohta T., Kimura M., A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population, Genet. Res. Camb. 22 (1973) 201–204.

[53] Paszek A.A., Flickinger G.H., Fontanesi L., Beattie C.W., Rohrer G.A., Alexander L., Schook L.B., Evaluating evolutionary divergence with microsatellites, J. Mol. Evol. 46 (1998) 121–126.

[54] Pérez-Lezaun A., Calafell F., Mateu E., Comas D., Ruiz-Pacheco R., Bertranpetit J., Microsatellite variation and the differentiation of modern humans, Hum. Genet. 99 (1997) 1–7.

[55] Pollock D.D., Bergman A., Feldman M.W., Goldstein D.B., Microsatellite with range constraints: parameter estimation and improved distances for use in phylogenetic reconstruction, Theor. Popul. Biol. 53 (1998) 256–271.

[56] Poteaux C., Bonhomme F., Berrebi P., Microsatellite polymorphism and genetic impact of restocking in mediterranean brown trout, Heredity 82 (1999) 645–653.

[57] Reynolds J., Weir B.S., Cockerham C.C., Estimation of the coancestry coefficient: basis for a short-term genetic distance, Genetics 105 (1983) 767–779.

[58] Rogers J.S., Measures of genetic similarity and genetic distances, Univ. of Texas Publ., 1972.

[59] Saitbekova N., Gaillard C., Obexer-Ruff G., Dolf G., Genetic diversity in Swiss goat breeds based on microsatellites analysis, Anim. Genet. 30 (1999) 36–41.

[60] Santos E.J.M., Epplen J.T., Epplen C., Extensive gene flow in human populations as revealed by protein and microsatellites DNA markers, Hum. Hered. 47 (1997) 165–172.

[61] Shriver M.D., Jin L., Chakraborty R., Boerwinkle E., VNTR allele frequency distribution under the stepwise mutation model: a computer simulation approach, Genetics 134 (1993) 983–993.

[62] Shriver M.D., Jin L. , Boerwinkle E., Deka R., Ferrell R.E., A novel measure of genetic for highly polymorphic tandem repeat loci, Mol. Biol. Evol. 12 (1995) 914–920.

[63] Slatkin M., Inbreeding coefficients and coalescent time, Genet. Res. Camb. 58 (1994) 167–175.
[64] Slatkin M., A measure of population subdivision based on microsatellite allele frequencies, Genetics 139 (1995) 457–462.
[65] Tajima F., Infinite-allele model and infinite-site model in population genetics, J. Genet. 75 (1996) 27–31.
[66] Takezaki N., Nei M., Genetic distances and reconstruction of phylogenetic trees from microsatellite data, Genetics 144 (1996) 389–399.
[67] Thaon d'Arnoldi C., Foulley J.-L., Ollivier L., An overview of the Weitzman approach to diversity, Genet. Sel. Evol. 30 (1998) 149–161.
[68] Valdes A.M., Slatkin M., Freimer N.B., Allele frequencies at microsatellite loci: the stepwise mutation model revisited, Genetics 133 (1993) 737–749.
[69] Vitalis R., Dawson K., Boursot P., Interpretation of variation across marker loci as evidence of selection, Genetics 158 (2001) 1811–1823.
[70] Wiegand P., Meyer E., Brinkmann B., Microsatellites structures in the context of human evolution, Electrophoresis 21 (2000) 889–895.
[71] Weitzman M.L., What to preserve? An application of diversity theory to crane conservation, Quart. J. Econ. 108 (1993) 157–183.
[72] Wright S., The genetical structure of populations, Ann. Eugenics 15 (1951) 323–354.
[73] Zhivotovsky L.A., Feldman M.W., Grishechkin S.A., Biased mutation and microsatellites variation, Mol. Biol. Evol. 14 (1997) 926–933.
[74] Zhivotovsky L.A., A new genetic distance with application to constrained variation at microsatellite loci, Mol. Biol. Evol. 16 (1999) 467–471.

APPENDIX A

Denote by E_e and Var_e the sampling expectation and variance respectively. Setting $\mu_{X,k} = \sum_i x_i^k$ and $\mu_{X,k,k'} = \sum_{i\neq j}^{k} x_i^k x_j^{k'}$ (idem for population Y) and $\nu_{k,k'} = \sum_i x_i^k y_i^{k'}$, the sampling expectation of $\hat{D}_S$ calculated with $m_{X,\bullet} = m_{Y,\bullet} = m$ and a Taylor expansion of the second order around unbiased estimates of j is

$$E_e(\hat{D}_S) = D_S + \frac{1}{2m}\frac{\nu_{1,2} + \nu_{2,1} - (\nu_{1,1})^2}{(\nu_{1,1})^2} - \frac{1}{m}\left(\frac{\mu_{X,3} - (\mu_{X,2})^2}{(\mu_{X,2})^2}\right) - \frac{1}{m}\left(\frac{\mu_{Y,3} - (\mu_{Y,2})^2}{(\mu_{Y,2})^2}\right) + O\left(\frac{1}{m^2}\right).$$

APPENDIX B

One locus variance of D_m assuming genetic drift only

Denote by E_d and Var_d the drift expectation and variance respectively. The total variance of $\hat{D}_m$ estimated with sampled allele frequencies and unbiased estimation of $\sum_i p_{X,i}^2$ and $\sum_i p_{Y,i}^2$, may be decomposed into

$$Var(\hat{D}_m) = E_d[Var_e(\hat{D}_m)] + Var_d[E_e(\hat{D}_m)].$$

Assuming that $\left(1-\frac{1}{2N}\right)\left(1-\frac{2}{2N}\right)$, $\left(1-\frac{1}{2N}\right)\left(1-\frac{2}{2N}\right)\left(1-\frac{3}{2N}\right)$, c_1 and c_2 in Nei and Chakravarty [50], may be respectively approximated by $(1-F)^3$, $(1-F)^6$, $\frac{9}{5}$ and $\frac{1}{5}$ in Chevalet and Gillois [8] and setting $\mu_k = \sum_i p_{0,i}^k$ and $\mu_{k,k'} = \sum_{i\neq j}^k p_{0,i}^k p_{0,j}^{k'}$,

$$
\begin{aligned}
Var_{\mathrm{d}}[E_{\mathrm{e}}(\hat{D}_{\mathrm{m}})] &= \frac{1}{2}(F_X^2+F_Y^2)\big[(\mu_2+(\mu_2)^2-2\mu_3)-6(\mu_3-(\mu_2)^2)\big] \\
&\quad + F_X F_Y(\mu_2-(\mu_2)^2+(\mu_2)^3-\mu_4)+O(\epsilon^3) \\
E_{\mathrm{d}}[Var_{\mathrm{e}}(\hat{D}_{\mathrm{m}})] &= \left(\frac{1}{m_{X,\bullet}}+\frac{1}{m_{Y,\bullet}}\right)(F_X+F_Y)\big(\mu_2+(\mu_2)^2-2\mu_3\big) \\
&\quad + \left(\frac{1}{2m_{X,\bullet}^2}+\frac{1}{2m_{Y,\bullet}^2}+\frac{1}{m_{X,\bullet}m_{Y,\bullet}}\right) \\
&\quad \times \big(\mu_2+(\mu_2)^2-2\mu^3\big)+O(\epsilon^3)
\end{aligned}
$$

where $O(\epsilon^3)$ means $O(F^3)+O\left(\frac{1}{m^3}\right)+O\left(\frac{F}{m^2}\right)+O\left(\frac{F^2}{m}\right)\cdot$

Assuming that the founder frequencies are equally distributed within loci, $p_{0,i}=\frac{1}{k_0}$, the factors $\mu_2+(\mu_2)^2-2\mu_3$ and $\mu_2-(\mu_2)^2+(\mu_2)^3-\mu_4$ are equal to $(k_0-1)/k_0^2$, whereas $\mu_3-(\mu_2)^2$ is equal to 0. Then the approximated standard deviation of D_{m} can be simplified to

$$
\begin{aligned}
\sigma(\hat{D}_{\mathrm{m}}) &= \sqrt{2\frac{(k_0-1)}{k_0^2}\left(\bar{F}+\left(\frac{1}{2m_{X,\bullet}}+\frac{1}{2m_{Y,\bullet}}\right)\right)^2+O(\epsilon^3)} \\
&\approx \sqrt{2\frac{(k_0-1)}{k_0^2}}\left(\bar{F}+\left(\frac{1}{2m_{X,\bullet}}+\frac{1}{2m_{Y,\bullet}}\right)\right)\cdot
\end{aligned}
\tag{A.1}
$$

One locus variance of D_{R} assuming genetic drift only

With unbiased estimation of $\sum_i p_{X,i}^2$ and $\sum_i p_{Y,i}^2$,

$$
E_{\mathrm{e}}(\hat{D}_{\mathrm{R}}) = E_{\mathrm{e}}\left[\frac{\hat{D}_{\mathrm{m}}}{1-\sum_i x_i y_i}\right] \approx \frac{D_{\mathrm{m}}}{1-j_{XY}}
$$

with D_{m}, the distance computed with true allele frequencies (equation (3)).

$$
Var(\hat{D}_{\mathrm{m}}) = E_{\mathrm{d}}[Var_{\mathrm{e}}(\hat{D}_{\mathrm{R}})] + Var_{\mathrm{d}}\left[\frac{D_{\mathrm{m}}}{1-j_{XY}}\right]
$$

$$Var_{\mathrm{d}}\left[\frac{D_{\mathrm{m}}}{1-j_{XY}})\right] \approx \frac{Var_{\mathrm{d}}(D_{\mathrm{m}})}{E_{\mathrm{d}}^2(1-j_{XY})} - 2\frac{Cov_{\mathrm{d}}(D_{\mathrm{m}}, 1-j_{XY})E_{\mathrm{d}}(D_{\mathrm{m}})}{E_{\mathrm{d}}^3(1-j_{XY})} + \frac{E_{\mathrm{d}}^2(D_{\mathrm{m}})Var_{\mathrm{d}}(1-j_{XY})}{E_{\mathrm{d}}^4(1-j_{XY})}$$

$$E_{\mathrm{d}}\left[Var_{\mathrm{e}}\left(\frac{\hat{D}_{\mathrm{m}}}{1-\sum_i x_i y_i}\right)\right] \approx \frac{E_{\mathrm{d}}[Var_{\mathrm{e}}(\hat{D}_{\mathrm{m}})]}{E_{\mathrm{d}}\left[E_{\mathrm{e}}^2\left(1-\sum_i x_i y_i\right)\right]} - 2\frac{E_{\mathrm{d}}\left[Cov_{\mathrm{e}}\left(\hat{D}_{\mathrm{m}}, 1-\sum_i x_i y_i\right)E_{\mathrm{e}}(\hat{D}_{\mathrm{m}})\right]}{E_{\mathrm{d}}\left[E_{\mathrm{e}}^3\left(1-\sum_i x_i y_i\right)\right]} + \frac{E_{\mathrm{d}}\left[E_{\mathrm{e}}^2(\hat{D}_{\mathrm{m}})Var_{\mathrm{e}}\left(1-\sum_i x_i y_i\right)\right]}{E_{\mathrm{d}}\left[E_{\mathrm{e}}^4\left(1-\sum_i x_i y_i\right)\right]}.$$

As in Nei and Chakravarty [50], we do not take into account the second and third terms of the Taylor expansion (in order to compute the second term we need to know the moment of the order 5 of frequency distributions under genetic drift).

$$\frac{Var_{\mathrm{d}}(D_{\mathrm{m}})}{E_{\mathrm{d}}^2(1-j_{XY})} = \frac{1}{2}(F_X^2+F_Y^2)\left[\left(\frac{\mu_2+(\mu_2)^2-2\mu_3}{(1-\mu^2)^2}\right)-6\frac{\mu_3-(\mu_2)^2}{(1-\mu^2)^2}\right] + F_X F_Y\left(\frac{\mu_2-(\mu_2)^2+(\mu_2)^3-\mu_4}{(1-\mu^2)^2}\right)+O(\epsilon^3)$$

$$\frac{E_{\mathrm{d}}[Var_{\mathrm{e}}(\hat{D}_{\mathrm{m}})]}{E_{\mathrm{d}}\left[E_{\mathrm{e}}^2\left(1-\sum_i x_i y_i\right)\right]} \approx \left(\frac{1}{m_{X,\bullet}}+\frac{1}{m_{Y,\bullet}}\right)(F_X+F_Y)\frac{\mu_2+(\mu_2)^2-2\mu_3}{1-\mu_2} + \left(\frac{1}{2m_{X,\bullet}^2}+\frac{1}{2m_{Y,\bullet}^2}+\frac{1}{m_{X,\bullet}m_{Y,\bullet}}\right) \times \frac{\mu_2+(\mu_2)^2-2\mu_3}{1-\mu_2}+O(\epsilon^3).$$

With the same approximations as in the previous section, the approximated standard deviation of $\hat{D}_{\mathrm{R}}$ can be simplified to

$$\sigma(\hat{D}_{\mathrm{R}}) = \sqrt{\frac{2}{k_0-1}\left(\bar{F}+\left(\frac{1}{2m_{X,\bullet}}+\frac{1}{2m_{Y,\bullet}}\right)\right)^2+O(\epsilon^3)} \approx \sqrt{\frac{2}{k_0-1}}\left(\bar{F}+\left(\frac{1}{2m_{X,\bullet}}+\frac{1}{2m_{Y,\bullet}}\right)\right). \quad \text{(A.2)}$$

The validity of this approximated formula has been checked by way of computer simulations in Section 4.3.