Infection, Genetics and Evolution 33 (2015) 227–232

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Short communication

Genetic identities and local inbreeding in pure diploid clones with homoplasic markers: SNPs may be misleading

Thierry De Meeûs

IRD (INTERTRYP), UMR 177 IRD-CIRAD, Centre International de Recherche-Développement sur l'Elevage en zone Subhumide (CIRDES), 01 BP 454 Bobo-Dioulasso 01, Burkina Faso
IRD (INTERTRYP), UMR177 IRD-CIRAD, TA A-17/G, Campus International de Baillarguet, 34398 Montpellier Cedex 5, France

ARTICLE INFO

Article history:
Received 29 January 2015
Received in revised form 6 May 2015
Accepted 7 May 2015
Available online 8 May 2015

Keywords:
Population genetics
Heterozygosity
Genetic diversity
Homoplasy
Equilibrium

ABSTRACT

Expected values for observed heterozygosity, genetic diversity, and inbreeding of individuals relative to inbreeding of the population (F_{IS}) are derived in the case of one locus displaying homoplasy with K possible allelic states (KAM model) in a clonal diploid population. Heterozygosity (H_O) and genetic diversity (H_S) are substantially affected by homoplasy as long as the number of alleles $K \leqslant 10$, while F_{IS} remains weakly affected in any case. Simulations suggest that in big populations, or in case of maximum homoplasy ($K = 2$), expected values can appear far from the observed ones because equilibrium takes too many generations to be reached at homoplasic markers in clonally propagating populations. This raises some concern on the use of SNPs, at least in clonal populations.

1. Introduction

In diploid clones, model were derived in the infinite allele framework (Balloux et al., 2003; De Meeûs and Balloux, 2004, 2005; De Meeûs et al., 2006). In that case, it is expected a total absence of heterozygotes at equilibrium with the probability of within individual identity $Q_I = 1$ and consequently strongly negative inbreeding of individuals relative to sub-population's inbreeding as measured by Wright's F_{IS} (Wright, 1965). This was successfully used using microsatellite markers that are highly polymorphic, in particular in *Trypanosoma brucei gambiense* (Koffi et al., 2009; Simo et al., 2010). The effect of homoplasy was theoretically studied in sexually reproducing populations (e.g. (Rousset, 1996) and for different mutation models regarding microsatellites (reviewed in (Estoup et al., 2002)). These works have demonstrated a weak effect on Wright's F_{IS} and a moderate one on F_{ST}, considering that microsatellite loci rarely (if any) follow a strict stepwise model of mutation. Recent development of new sequencing techniques provides the availability of huge numbers of markers with great homoplasy, e.g. the SNPs. This is the case for many pathogenic microbes, in particular diploid parasites with a clonal propagation (e.g. (Abbey et al., 2011; Goodhead et al., 2013; Yeo et al., 2013; Rogers et al., 2014; Carnes et al., 2015)). Because point mutations are rare and transversions twice less likely than transitions, SNPs are classically considered as bi-allelic markers (Vignal et al., 2002; Rosenberg et al., 2010; Sere et al., 2014), and hence display maximum homoplasy. Other approaches, like the coalescent, consider such kind of markers through the infinite sites model that allows inferences to be made through genealogical relationships of haplotypes. The model developed here concern the classical population genetics approach through Wright's F_{IS} as can be found elsewhere (Wright, 1951; Cockerham, 1969, 1973; Rousset, 1996; Balloux et al., 2003; De Meeûs and Balloux, 2005; De Meeûs et al., 2007).

In this note, I derive equations for identity probabilities within and between individuals in an isolated population for different values of K (maximum possible number of alleles). This allows computing expected values for expected observed heterozygosity H_O and expected genetic diversity H_S and F_{IS}. I show that homoplasy has almost no effect on expected F_{IS}. I also show that homoplasy has a significant impact on expected values for H_O and H_S as long as $K \leqslant 10$. Simulations however suggest that the situation worsens in very big populations or for loci with maximum homoplasy ($K = 2$) for which equilibrium values may take so much time that no natural population is expected to display those. This raises concerns about the use of SNPs in clonal populations.

2. Equations

I follow the same approach as in previous population genetics papers (Rousset, 1996; Balloux et al., 2003; De Meeûs and Balloux, 2005). Wright's F_{IS} can be defined as the inbreeding of

E-mail address: thierry.demeeus@ird.fr

http://dx.doi.org/10.1016/j.meegid.2015.05.008
1567-1348/

individuals relative to the inbreeding of sub-populations. It is a measure of deviation of the local genotypic distribution from the panmictic expectations and can be defined according to Cockerham (1969, 1973) as a function of individual inbreeding Q_I (probability to randomly draw the same allele within an individual) and subpopulation inbreeding Q_S (probability to randomly draw two identical alleles from two different individuals of the same subpopulation):

$$F_{IS} = \frac{Q_I - Q_S}{1 - Q_S} \tag{1}$$

I now assume a locus with K alleles mutating at rate u (with change in state into one of the $K-1$ other available alleles) in a purely clonal isolated population of size N with non-overlapping generations. At such a locus, probability of identity by state within individuals $Q_{I(t)}$ at generation t will evolve from one generation to the next as:

$$Q_{I(t+1)} = (1-u)^2[Q_{I(t)}] + 2u(1-u)\left[(1-Q_{I(t)})\frac{1}{K-1}\right] + u^2\left[Q_{I(t)}\frac{1}{(K-1)^2} + (1-Q_{I(t)})\frac{K-2}{(K-1)^2}\right] \tag{2}$$

Indeed, when no allele mutates, with probability $(1-u)^2$, then these alleles are identical only if they already were identical at the previous generation with probability $Q_{I(t)}$. Now, if one allele mutates and not the other, with probability $u(1-u)+(1-u)u$, only those that were different (with probability $1-Q_{I(t)}$) can become identical and the one that mutates has then a probability $1/(K-1)$ of becoming identical to the one that does not mutate. Finally, if both mutate (with probability u^2), and thus change of state, if already identical (with probability $Q_{I(t)}$) they become identical again with probability $1/(K-1)^2$ and, if they were different (with probability $1-Q_{I(t)}$), only $K-2$ alleles can offer the possibility of identity, with individual probability $1/(K-1)^2$. We can now rearrange Eq. (2) into:

$$Q_{I(t+1)} = Q_{I(t)}\left[(1-u)^2 - \frac{2u(1-u)}{K-1} + \frac{u^2}{(K-1)^2} - u^2\frac{K-2}{(K-1)^2}\right] + \frac{2u(1-u)}{K-1} + u^2\frac{K-2}{(K-1)^2} \tag{3}$$

At equilibrium $Q_{I(t+1)} = Q_{I(t)} = \widehat{Q_I}$, hence:

$$\widehat{Q_I} \approx \widehat{Q_I}\left[(1-u)^2 - \frac{2u(1-u)}{K-1} + \frac{u^2}{(K-1)^2} - u^2\frac{K-2}{(K-1)^2}\right] + \frac{2u(1-u)}{K-1} + u^2\frac{K-2}{(K-1)^2} \tag{4}$$

$$\Longleftrightarrow$$

$$\widehat{Q_I} = \frac{\frac{(K-1)2u(1-u)+u^2(K-2)}{(K-1)^2}}{\frac{(K-1)^2(1-1+2u-u^2)+(K-1)2u(1-u)-u^2+u^2(K-2)}{(K-1)^2}} \tag{5}$$

$K > 1$ and $u > 0$, hence:

$$\widehat{Q_I} = \frac{2(K-1)(1-u)+u(K-2)}{(K-1)^2(2-u)+2(K-1)(1-u)-u+u(K-2)} \tag{6}$$

$$\Longleftrightarrow$$

$$\widehat{Q_I} = \frac{K(2-u)-2}{K(K-1)(2-u)-2u} \tag{7}$$

Knowing that for reasonable small $K > 1$ $u \ll 1$, we can approximate:

$$\widehat{Q_I} \approx \frac{2(K-1)}{2K(K-1)} = \frac{1}{K} \tag{8}$$

which is the probability to randomly draw twice the same allele among the K possible ones.

It can be noted that Q_I is the reversed of the probability of finding a heterozygous individual (within individual probability to find different alleles) also known as observed heterozygosity $H_O = 1 - Q_I$. Hence, it is easily seen that, with $K = 2$, $H_O = 0.5$. This means that half of SNPs sites will be expected to be found heterozygous in a clonal population, while the other half will be found as fixed homozygous individuals.

The evolution of identity by state between individuals $Q_{S(t)}$ will be:

$$Q_{S(t+1)} = (1-u)^2\left\{\frac{1}{N}\left[Q_{I(t)} + (1-Q_{I(t)})\times\frac{1}{2}\right] + \left(1-\frac{1}{N}\right)Q_{S(t)}\right\} + 2u(1-u)\left\{\begin{array}{c}\frac{1}{N}\left[(1-Q_{I(t)})\frac{1}{K-1}\right]\\ +\left(1-\frac{1}{N}\right)\left[(1-Q_{S(t)})\frac{1}{K-1}\right]\end{array}\right\} + u^2\left\{\begin{array}{c}\frac{1}{N}\left[Q_{I(t)}\frac{1}{(K-1)^2} + (1-Q_{I(t)})\frac{K-2}{(K-1)^2}\right]\\ +\left(1-\frac{1}{N}\right)\left[Q_{S(t)}\frac{1}{(K-1)^2} + (1-Q_{S(t)})\frac{K-2}{(K-1)^2}\right]\end{array}\right\} \tag{9}$$

When both alleles mutate, with probability $(1-u)^2$, and when both alleles come from the same ancestor, with probability $1/N$, these two alleles can be identical because they already were in the common ancestor, with probability $Q_{I(t)}$, or they were not identical $(1-Q_{I(t)})$ but the same allele is sampled twice, with probability ½, or the two alleles come from two different individuals of generation t (with probability $1-1/N$) and the probability that they are identical is then $Q_{S(t)}$. If one allele mutate and not the other, with probability $2u(1-u)$, and the two alleles come from the same ancestor individual $(1/N)$, the two alleles can become identical if they were not identical in that ancestor $(1-Q_{I(t)})$ and the mutant becomes identical to the other allele that does not mutate, with probability $1/(K-1)$, or if the two alleles came from two different ancestor individuals (with probability $1-1/N$) the two alleles can become identical if they were not identical at time t $(1-Q_{S(t)})$ and the mutant becomes identical to the other $(1/(K-1))$. Finally, in case both alleles mutate, with probability $(1-u)^2$, and if they come from the same ancestor individual $(1/N)$, if the two alleles were identical in that individual $(Q_{I(t)})$, they mutate into the same state with probability $1/(K-1)^2$, or if they were different in that ancestor $(1-Q_{I(t)})$ there are $K-2$ possibilities that both mutate into the same state with probability $1/(K-1)^2$, or the two alleles come from two different ancestor individuals $(1-1/N)$ in which they were identical $(Q_{S(t)})$ and both mutate into the same state $(1/(K-1)^2)$ or not identical $(1-Q_{S(t)})$ and both mutate into the $K-2$ remaining state and become identical $((K-2)/(K-1)^2)$. Then Eq. (9) can be rearranged into:

$$Q_{S(t+1)} = (1-u)^2\left\{\frac{1+Q_{I(t)}}{2N} + \left(1-\frac{1}{N}\right)Q_{S(t)}\right\} + \frac{2u(1-u)}{K-1}\left\{\frac{1}{N}(1-Q_{I(t)}) + \left(1-\frac{1}{N}\right)(1-Q_{S(t)})\right\} + \frac{u^2}{(K-1)^2}\left\{\frac{1}{N}\left[Q_{I(t)} + (1-Q_{I(t)})(K-2)\right] + \left(1-\frac{1}{N}\right)\left[Q_{S(t)} + (1-Q_{S(t)})(K-2)\right]\right\} \tag{10}$$