Theoretical Population Biology 107 (2016) 4–13

Contents lists available at ScienceDirect

Theoretical Population Biology

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

Pedigrees or markers: Which are better in estimating relatedness and inbreeding coefficient?

Jinliang Wang *

Institute of Zoology, Zoological Society of London, London NW1 4RY, United Kingdom

ARTICLE INFO

Article history:
Received 15 January 2015
Available online 3 September 2015

Keywords:
Pedigree
SNPs
Genomic markers
Simulations
Inbreeding coefficient
Relatedness

ABSTRACT

Individual inbreeding coefficient (F) and pairwise relatedness (r) are fundamental parameters in population genetics and have important applications in diverse fields such as human medicine, forensics, plant and animal breeding, conservation and evolutionary biology. Traditionally, both parameters are calculated from pedigrees, but are now increasingly estimated from genetic marker data. Conceptually, a pedigree gives the expected F and r values, F_P and r_P, with the expectations being taken (hypothetically) over an infinite number of individuals with the same pedigree. In contrast, markers give the realised (actual) F and r values at the particular marker loci of the particular individuals, F_M and r_M. Both pedigree (F_P, r_P) and marker (F_M, r_M) estimates can be used as inferences of genomic inbreeding coefficients F_G and genomic relatedness r_G, which are the underlying quantities relevant to most applications (such as estimating inbreeding depression and heritability) of F and r. In the pre-genomic era, it was widely accepted that pedigrees are much better than markers in delineating F_G and r_G, and markers should better be used to validate, amend and construct pedigrees rather than to replace them. Is this still true in the genomic era when genome-wide dense SNPs are available? In this simulation study, I showed that genomic markers can yield much better estimates of F_G and r_G than pedigrees when they are numerous (say, 10^4 SNPs) under realistic situations (e.g. genome and population sizes). Pedigree estimates are especially poor for species with a small genome, where F_G and r_G are determined to a large extent by Mendelian segregations and may thus deviate substantially from their expectations (F_P and r_P). Simulations also confirmed that F_M, when estimated from many SNPs, can be much more powerful than F_P for detecting inbreeding depression in viability. However, I argue that pedigrees cannot be replaced completely by genomic SNPs, because the former allows for the calculation of more complicated IBD coefficients (involving more than 2 individuals, more than one locus, and more than 2 genes at a locus) for which the latter may have reduced capacity or limited power, and because the former has social and other significance for remote relationships which have little genetic significance and cannot be inferred reliably from markers.

1. Introduction

Relatedness (r) and inbreeding coefficients (F) are fundamental parameters in population and quantitative genetics (Wright, 1921, 1922), and have important applications in diverse fields such as human medicine, forensics, plant and animal breeding, conservation and evolutionary biology (for a review, see Weir et al., 2006). Originally, F and r were defined by Wright as the correlation between the two homologous genes at a locus within a diploid individual or taken at random from each of two individuals respectively. The correlation is due to the common ancestry or shared genealogy of the two parents (for F) or two pairs of parents (for r), and thus has the same expected value for any locus in the genome. Later, Malécot (1948) introduced an alternative definition in terms of the probability of identity by descent (IBD) of the two homologous genes at a locus within an individual (for F) or between two individuals (for r), where genes IBD are copies of the same ancestral allele. Both definitions have an implicit reference population against which F and r are measured and in which all homologous genes within and between individuals are assumed non-IBD or uncorrelated (or equivalently all individuals are assumed non-inbred and unrelated; Wang, 2014). Wright (1965) showed that the two definitions are equivalent in some simple cases, but the correlation definition is more general and can give meaningful negative

* Correspondence to: Institute of Zoology, Regent's Park, London NW1 4RY, United Kingdom.
E-mail address: jinliang.wang@ioz.ac.uk.

http://dx.doi.org/10.1016/j.tpb.2015.08.006
0040-5809/

values in some more complicated cases. For example, the inbreeding coefficient of a hybrid individual and the relatedness between a resident and an immigrant individual are negative according to the correlation definition, but are never smaller than zero in terms of IBD because probability is inherently non-negative.

Traditionally, both F and r were calculated from pedigrees as demonstrated by Wright (1922) with a cattle population, and had limited applications because pedigrees were usually unavailable or incomplete except in artificially well controlled populations such as those in plant and animal breeding. With the rapid developments of blood-protein markers, microsatellites and now genome wide dense SNPs, F and r are increasingly estimated from genetic marker data (Ritland, 1996; Lynch and Ritland, 1999) in wild and other not well studied populations. As a result, F and r have gained much broader applications in human medicine, forensics, conservation and evolutionary biology (Weir et al., 2006). Calculating F and r from marker data is faster, (arguably) easier and less expensive than from pedigree data, because it does not require collecting breeding records intensively, of all individuals in a population, and extensively, over a number of generations. It can even be carried out without observing the animals by non-invasive sampling (e.g. Lucchini et al., 2002). As a result, marker based F and r are now routinely calculated and used in studying mate choice, kin selection, mating system, inbreeding depression, and the inheritance of quantitative traits in natural populations (e.g. Garant and Kruuk, 2005; Foerster et al., 2006; Chandler and Zamudio, 2008; Cohas et al., 2008; Langen et al., 2011; Wang and Lu, 2011; Robinson et al., 2012; Forstmeier et al., 2012), and in designing captive breeding programs to minimise inbreeding and to maintain genetic diversity in endangered species (e.g. Wang, 2001; Fernandez et al., 2005). Indeed the possibility of calculating r and F from markers without pedigrees opens up new avenues of research on the genetics and evolution of wild populations in their natural habitats, and has been contributing tremendously to our understanding of the ecology and evolution of many species in the wild, such as the mating systems in birds (Griffith et al., 2002).

Genetic markers and pedigrees, which give better estimates of F and r and thus should be preferred given the options? In the microsatellite era, typically 10–20 markers are used in estimating F and r or their surrogates such as multilocus heterozygosity (Slate et al., 2004; Szulkin et al., 2010). The estimates may be unbiased, but can be highly imprecise with a large sampling variance (e.g. Lynch and Ritland, 1999; Van de Casteele et al., 2001; Wang, 2002; Csilléry et al., 2006). This is not surprising because the actual F and r could have a high locus to locus variation due to Mendelian segregation (Hill and Weir, 2011). Take an individual from a full sib mating as an example. For a given locus, the two homologous genes in the individual have only two alternative IBD status, either IBD or non-IBD. The actual F at the locus is thus 1 and 0 with probabilities of 0.25 and 0.75, respectively, yielding an expected F of 0.25 as calculated from pedigree. The between-locus variance in actual F is extremely high, except for loci in close linkage. For independent loci, the variance of actual F is $F \times (1 - F) = 0.1875$, and the coefficient of variation (CV) is 1.73 for the individual. The actual variance and CV for the estimated F can be substantially larger, because of the additional estimation errors caused by limited marker information (or the difficulty in inferring IBD from the observed genes identical in state, IIS, or identical by state, IBS). Unsurprisingly, marker based estimators of r (Csilléry et al., 2006) and F (or its surrogate Slate et al., 2004; Balloux et al., 2004) have a depressingly low correlation with pedigree based estimates and explain only a tiny fraction of the variance of pedigree based values. For the same set of quantitative traits, inbreeding depression was detected by using pedigree-based F, but not by using marker-based F (Slate et al., 2004). It is logically concluded that pedigrees are much better than microsatellites in delineating relatedness and inbreeding (Pemberton, 2004), and markers should better be used to validate, amend and construct pedigrees (Pemberton, 2008) rather than to replace pedigrees completely.

Now in the genomic era with rapidly increasing applications of dense SNPs in model and non-model species, are pedigrees still preferable for calculating F and r? Under which conditions are pedigree based estimates better than dense SNPs based estimates, or *vice versa*? Answering these questions is important in optimising experimental design and in addressing many ecological and evolutionary issues such as inferring inbreeding depression more effectively and efficiently. For example, if dense SNPs yield equivalent or better estimates of F and r, then there is no need to make a tremendous effort in accumulating behaviour data over a long period of time to recover the pedigree with sufficient depth and width. This would be especially good news for studies of many wild species that are rare, elusive or difficult to observe, or that have a long generation interval or have a population too large to observe all individuals.

No obvious answers to the above questions can be derived from current knowledge about pedigree and marker based estimators. In almost all practical applications, the actual quantities of interest are the mean actual (realised) relatedness (r_G) and inbreeding coefficient (F_G). The mean is taken conceptually over all loci in a genome, whatever the loci are defined (Wang, 2012). Both pedigrees and markers just provide estimates of r_G and F_G. Previously, the accuracy of marker estimators is assessed against pedigree based F and r values (e.g. Lynch and Ritland, 1999; Van de Casteele et al., 2001; Wang, 2002; Balloux et al., 2004). The approach is justifiable as a good approximation when the number of markers used in the estimation is small such that marker based estimates, r_M and F_M, are expected to be inaccurate, and when the genome is large such that the expected values, r_P and F_P, provided by pedigrees are close to the gold standards of r_G and F_G. However, r_P and F_P may deviate from r_G and F_G substantially for realistic genomes, especially when they are small (Hill and Weir, 2011). Therefore, with a decreasing genome size and increasing number of genomic SNPs, it becomes increasingly possible that marker-based estimators outperform pedigree-based estimators of r_G and F_G. Both estimators should be evaluated against the same gold standards of r_G and F_G.

In this study, I use individual-based simulations to compare pedigrees and genomic SNPs in estimating r_G and F_G under various scenarios involving factors such as genome size, marker density, and pedigree width and depth, and to compare the powers of F_G and its pedigree and marker based estimators in detecting inbreeding depression. The results are discussed in the context of molecular ecology and conservation, and have implications for the experimental design, data analyses and interpretations in applications of F and r to these and other areas.

2. Methods

To avoid confusion, I will first clarify different concepts or estimators of F and r, and then describe the simulation procedures, data analyses methods, and the measurements of accuracy adopted by the study.

2.1. Different concepts and estimators of inbreeding and relatedness

The actual or realised inbreeding coefficient at a particular locus of a particular individual, F_g, is the probability of IBD of the two homologous genes at the locus of the individual. It takes two alternative values, either 1 (IBD) or 0 (non-IBD). The distribution of F_g across loci of the individual depends on both the pedigree and the segregation and recombination events involved in generating the individual's genome. In terms of pedigree, an individual whose parents have more common ancestors or/and more recent