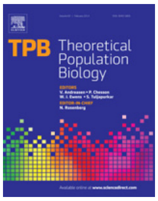
Theoretical Population Biology 107 (2016) 31–38

Contents lists available at ScienceDirect

Theoretical Population Biology

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



Relationship inference from the genetic data on parents or offspring: A comparative study

Steven Gazal[a,b,c], Emmanuelle Génin[d,e,f], Anne-Louise Leutenegger[g,h,*]

[a] Inserm, UMR 1137, IAME, Paris, France
[b] Université Paris Diderot, Sorbonne Paris Cité, UMR 1137, Paris, France
[c] Plateforme de Génétique constitutionnelle-Nord (PfGC-Nord), Paris, France
[d] Inserm, UMR 1078, Brest, France
[e] Université Bretagne Occidentale, Brest, France
[f] Centre Hospitalier Régional Universitaire, Brest, France
[g] Inserm, U946, Genetic Variation and Human Diseases Lab, Paris, France
[h] Université Paris Diderot, Sorbonne Paris Cité, Institut Universitaire d'Hématologie, UMR 946, Paris, France

ARTICLE INFO

Article history:
Received 3 March 2015
Available online 30 September 2015

Keywords:
Relationship inference
Mating habit
Identity-by-descent
Kinship coefficient
Inbreeding coefficient
Genome sharing

ABSTRACT

Relationship inference in a population is of interest for many areas of research from anthropology to genetics. It is possible to directly infer the relationship between the two individuals in a couple from their genetic data or to indirectly infer it from the genetic data of one of their offspring. For this reason, one can wonder if it is more advantageous to sample couples or single individuals to study relationships of couples in a population. Indeed, sampling two individuals is more informative than sampling one as we are looking at four haplotypes instead of two, but it also doubles the cost of the study and is a more complex sampling scheme.

To answer this question, we performed simulations of 1000 trios from 10 different relationships using real human haplotypes to have realistic genome-wide genetic data. Then, we compared the genome sharing coefficients and the relationship inference obtained from either a pair of individuals or one of their offspring using both single-point and multi-point approaches.

We observed that for relationships closer than 1st cousin, pairs of individuals were more informative than one of their offspring for relationship inference, and kinship coefficients obtained from single-point methods gave more accurate or equivalent genome sharing estimations. For more remote relationships, offspring were more informative for relationship inference, and inbreeding coefficients obtained from multi-point methods gave more accurate genome sharing estimations.

In conclusion, relationship inference on a parental pair or on one of their offspring provides complementary information. When possible, sampling trios should be encouraged as it could allow spanning a wider range of potential relationships.

1. Introduction

Inferring the relationship that exists between the two partners in a couple is of interest for many areas of research from anthropology to genetics. It is informative of the mating habits and marriage patterns in a given population and allows comparative studies between populations (Romeo and Bittles, 2014). Several such studies have been performed in different human and animal populations based on pedigree records (see for example a recent work by Zlotogora and Shalev, 2014 in a Muslim village) or population surveys of the number of marriages between relatives based on church records (Sutter and Goux, 1964).

Recent advances in molecular genetics have made it possible to obtain genotype information for hundreds of thousands of markers spanning the whole genome. This genetic information can be used to estimate the kinship coefficients between pairs of individuals. This is now routinely done as a quality control step to identify related individuals in a sample and discard them to avoid false

[*] Correspondence to: Genetic Variation and Human Diseases Lab, Inserm U946, Fondation Jean Dausset-CEPH, 27 rue Juliette Dodu, 75010 Paris, France.
E-mail address: anne-louise.leutenegger@inserm.fr (A.-L. Leutenegger).

http://dx.doi.org/10.1016/j.tpb.2015.09.002

positives in case–control association studies (Voight and Pritchard, 2005). Unknown relatedness between individuals might then be discovered as it was, for example in the Hapmap data (Pemberton et al., 2010). When genetic information is available on spouses, it is then possible to get an overview of the realized relationships between them. Indeed, the pedigree only gives the expected relatedness and does not directly provide the true proportion of their genome that they really share (Speed and Balding, 2015). This was recently illustrated in both human and animal data (Colonna et al., 2007; Wang et al., 2014). Knowing this realized relationship might be of interest to identify regions of the genome that could harbor disease related genes.

Several different methods have been developed to estimate kinship coefficients between two individuals and infer their possible relationship or even reconstruct pedigrees from genetic data. These methods aim to identify regions of the genome that were inherited by the two individuals from a common ancestor and that are therefore identical-by-descent (IBD). They can be divided into two groups: single point methods that use the information at each marker independently and multipoint methods that take into account linkage between markers (see Browning and Browning, 2012 for a review). The latter methods have been shown to allow a better detection of distant relationships between individuals.

In parallel, similar methods have been developed to estimate inbreeding coefficients and identify genomic regions shared homozygous-by-descent (HBD) by a single individual (Leutenegger et al., 2003). These methods have mostly been used in the context of homozygosity mapping to identify genes involved in rare recessive monogenic diseases (Leutenegger et al., 2006) or genomic regions potentially harboring rare recessive variants involved in complex diseases (Génin et al., 2012). However, it is also possible to exploit the realized inbreeding in a population to learn about the mating habits in this population. Indeed, since the inbreeding coefficient of an individual is the same as their parents' kinship coefficient (Malécot, 1948), one can infer parental relationships using one of their offspring. We have recently proposed to do so with the individuals from the Human Genome Diversity Panel (HGDP-CEPH) to infer the mating habits of world-wide populations (Leutenegger et al., 2011). We have developed software that allows inferring the most likely relationship of the parents from the available genetic data of the offspring (Gazal et al., 2014a). Such indirect inference, based on offspring data, presents the advantage of being much simpler in terms of sampling than a direct inference from the parents. Indeed, sampling couples could be difficult and perhaps more prone to ascertainment bias than sampling isolated individuals. The cost is also double since two individuals need to be genotyped to estimate the kinship coefficient, compared to only one when estimating inbreeding. However, the available information in a couple is richer than in a single individual as we are then looking at four haplotypes instead of only two haplotypes. Finally, inference from a couple tells us about *potential* mating in the population but inference from an individual is informative about *realized* mating in the population.

Genome-based kinship and inbreeding coefficients are only equal in expectation with a large variability around this expected value (Donnelly, 1983; Leutenegger et al., 2003). To date however, no study has compared the estimates obtained from the genetic data on couples and from the genetic data on one of their offspring except for some studies that focused on assortative mating and aimed at identifying regions of the genome where offspring were either more or less similar than expected given the kinship of their parents (Laurent et al., 2012; Laurent and Chaix, 2012).

In this paper, we are interested in comparing the relationship inference obtained from the genetic data on either a pair of individuals or one of their offspring and the estimation of the proportion of genome shared IBD (kinship coefficient of the pair) or HBD (inbreeding coefficient of the offspring). To do so, we performed a simulation study on trio data with different parental relationships and compare (1) for the relationship inference, the results obtained using RELPAIR for a pair of individuals (Epstein et al., 2000) and using FSuite for a single individual, and (2) for the estimation of genome sharing proportions, the results of PLINK (Purcell et al., 2007), GIBDLD (Han and Abney, 2013) and FSuite (Gazal et al., 2014a).

2. Materials and methods

2.1. Estimating genomic kinship and inbreeding coefficients

Approaches to estimate the genomic kinship and inbreeding coefficients can be organized into three main categories. The first category of approaches rely on the allele frequencies at each marker considered independently (single-point). They can either use method of moments (MoM) estimation (Purcell et al., 2007; Yang et al., 2011) or maximum likelihood estimation (Thompson, 1975; Milligan, 2003; Polasek et al., 2010). The second category of approaches rely on the segmental nature of IBD (Purcell et al., 2007; Gusev et al., 2009). Finally, the third category of approaches that rely on both the marker allele frequencies and the segmental nature of IBD through hidden Markov models (HMM) (Leutenegger et al., 2003; Browning, 2008; Browning and Browning, 2010; Han and Abney, 2011; Brown et al., 2012; Han and Abney, 2013; Gazal et al., 2014a). Here, we focus on the single-point MoM approaches as implemented in PLINK (Purcell et al., 2007) and the multi-point approaches as implemented in GIBDLD (Han and Abney, 2013) and FSuite (Gazal et al., 2014a).

PLINK option—het allows the estimation of the genomic inbreeding coefficient F_{PLINK} as the genome-wide excess homozygosity. It is obtained as a function of the number of observed homozygous loci and the allele frequencies.

PLINK option—genome allows the estimation of the genomic kinship coefficient K_{PLINK}. PLINK provides both $\hat{\pi}$, which is twice the kinship coefficient, and the probabilities k_i of sharing i alleles IBD between two individuals, with the following relation between these different quantities: $K_{PLINK} = \hat{\pi}/2 = 0.5*k_2+0.25*k_1$. Note that when neither individual in the pair is inbred, k_i probabilities are also referred to as Cotterman's k coefficients (Cotterman, 1940). The k's are a function of the number of loci with 0, 1 or 2 alleles identical-by-state and the allele frequencies.

For the multi-point approaches, FSuite provides the maximum likelihood estimate (MLE) of the genomic inbreeding coefficient F_{FSUITE}. Let X_k denote the HBD state (i.e., $X_k = 1$ if the 2 alleles at locus k within the individual are IBD, 0 otherwise), and Y_k the genotype of the individual at locus k ($k = 1$ to N the total number of loci). The HBD process of an individual is approximated by a Markov chain, the transition probabilities $P(X_k|X_{k-1})$ depending on F the inbreeding coefficient, A the rate of HBD state change per cM and t_k the genetic distance between adjacent loci. The model requires the specification of the transition probabilities between the different HBD states at adjacent markers. These different transition probabilities are given in Leutenegger et al. (2003). For example, the probability for staying HBD at marker k given HBD at marker $k-1$ is: $P(X_k = 1|X_{k-1} = 1) = \left(1 - e^{-At_k}\right) F + e^{-At_k}$. The model also requires the specification of emission probabilities $P(Y_k|X_k)$ that depend on the allele frequencies at locus k. These allele frequencies can be estimated on the studied sample, or on a reference panel (such as HGDP-CEPH or HapMap panels) if the studied sample is too small to estimate them. Parameters F and A are then estimated by maximum likelihood.

GIBDLD for the estimation of the genomic kinship coefficient K_{GIBDLD} between two individuals relies on a similar model. The observed data Y_k are the unphased genotypes of the two