BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 7 2014, pages 1027–1028
doi:10.1093/bioinformatics/btt652

Genetics and population analysis

Advance Access publication November 8, 2013

RelateAdmix: a software tool for estimating relatedness between admixed individuals

Ida Moltke[1,2] and Anders Albrechtsen[2,*]

[1]Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA and [2]Department of Biology, The Bioinformatics Centre, University of Copenhagen, 2200 Copenhagen N, Denmark

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Pairwise relatedness plays an important role in a range of genetic research fields. However, currently only few estimators exist for individuals that are admixed, i.e. have ancestry from more than one population, and these estimators fail in some situations.

Results: We present a new software tool, RelateAdmix, for obtaining maximum likelihood estimates of pairwise relatedness from genetic data between admixed individuals. We show using simulated data that it gives rise to better estimates than three state-of-the-art software tools, REAP, KING and Plink, while still being fast enough to be applicable to large datasets.

Availability and implementation: The software tool, implemented in C and R, is freely available from www.popgen.dk/software.

Contact: albrecht@binf.ku.dk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 16, 2013; revised on November 4, 2013; accepted on November 5, 2013

1 INTRODUCTION

Genetic relatedness between pairs of individuals plays an important role within several fields of genetic research including forensic and medical genetics. For example, most genome-wide association studies are based on the assumption that all individuals analyzed are unrelated and if related individuals are not removed or controlled for they can lead to highly inflated false-positive rates. Relatedness between a pair of individuals is usually described using the concept of identity-by-descent (IBD), which is genetic identity due to recent common ancestry. More specifically relatedness between two non-inbred individuals is often described by the fractions, k_0, k_1 and k_2, of the genome in which the two individuals share 0, 1 or 2 alleles IBD, respectively. Further, relatedness is sometimes summarized as the coefficient of relatedness r or the kinship coefficient $\theta = r/2$, which can both be calculated from k_0, k_1 and k_2 as $r = k_1/2 + k_2$.

There are numerous estimators for k_0, k_1, k_2, r and θ based on genetic data from individuals whose potential family relations are unknown, both method of moments estimators (Purcell *et al.*, 2007; Ritland, 1996) and maximum likelihood (ML) estimators (Milligan, 2003; Thompson, 1975). Additionally, there are several hidden Markov model-based methods that estimate IBD sharing locally along the genome (e.g. Albrechtsen *et al.*, 2009), which if applied to whole genomes also provide relatedness estimates. However, all of these estimators are based on an assumption that the analyzed individuals are from a homogeneous population, and it has been shown that when this assumption is violated it can lead to a marked misestimation of their relatedness (Rohlfs *et al.*, 2012; Thornton *et al.*, 2012). Motivated by this, Thornton *et al.* (2012) recently proposed a method of moments estimator for pairwise relatedness in admixed populations, implemented in the software REAP and an estimator for the kinship coefficient in structured populations was included in the software KING (Manichaikul *et al.*, 2010). Here we present a new ML-based software tool, RelateAdmix, for estimating k_0, k_1 and k_2 from single nucleotide polymorphism data, estimated admixture proportions and allele frequencies assuming no inbreeding. We obtain more accurate results than both REAP, KING and the often used software Plink (Purcell *et al.*, 2007), which does not take admixture into account.

2 METHOD

Assuming we have genotype data from M diallelic loci from two non-inbred individuals, ind1 and ind2, with ancestry from K source populations, we let $G^1 = (G^1_1, \ldots, G^1_M)$ and $G^2 = (G^2_1, \ldots, G^2_M)$ denote the two individuals' genotypes, $Q^1 = (Q^1_1, \ldots, Q^1_K)$ and $Q^2 = (Q^2_1, \ldots, Q^2_K)$ denote their ancestry/admixture proportions and $R = (k_0, k_1, k_2)$ denote the fractions in which ind1 and ind2 share 0, 1 and 2 alleles IBD. Using this notation, our goal is to estimate R from G^1, G^2, assuming Q^1, Q^2 and the allele frequencies in each of the K populations are known.

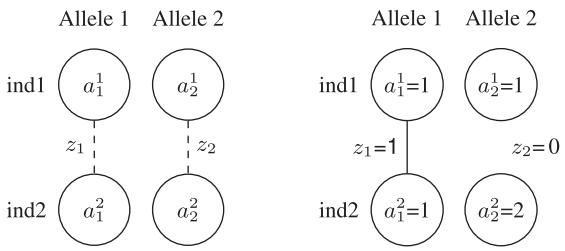

Fig. 1. Left: diagram of the unobserved variables, z_1, z_2 and a in a locus i. The circles represent the ancestral population of each allele, and lines indicate whether two alleles are IBD. Right: example configuration where the two individuals share one allele IBD originating from population 1, and the two other alleles originate from population 1 and 2, respectively, and are not IBD

*To whom correspondence should be addressed

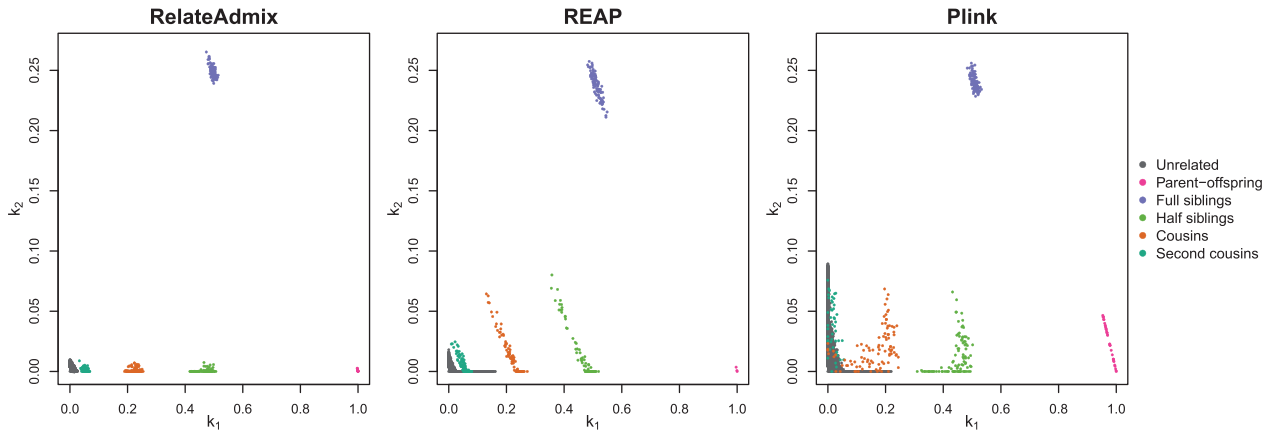

Fig. 2. Estimates of k_1 and k_2 from RelateAdmix, REAP and Plink, respectively. The true value of (k_1,k_2) is (0,0) for unrelated pairs, (1,0) for parent-offspring, (0.5,0.25) for full siblings, (0.5,0) for half siblings, (0.25,0) for first cousins and (0.0625,0) for second cousins

Assuming independence between loci we write the likelihood as

$$p(G^1, G^2|R, Q^1, Q^2) = \prod_{i=1}^{M} p(G_i^1, G_i^2|R, Q^1, Q^2)$$

$$= \prod_{i=1}^{M} \sum_{z_1, z_2} \sum_{a} p(G_i^1, G_i^2|Z_i = (z_1, z_2), A_i = a)$$

$$p(A_i = a|Z_i = (z_1, z_2), Q^1, Q^2) p(Z_i = (z_1, z_2)|R)$$

where for each locus z_1 indicates whether allele 1 in ind1 and ind2 are IBD, z_2 indicates whether allele 2 in ind1 and ind2 are IBD and $a = (a_1^1, a_2^1, a_1^2, a_2^2)$ represents the unobserved ancestral populations of the two individuals' two alleles, see Figure 1. Note that for convenience, we have restricted the IBD sharing patterns in the sense that allele 1 of ind1 can only be IBD with allele 1 of ind2, and allele 2 of ind1 can only be IBD with allele 2 of ind2. Because we sum over all values of a and we do not fix the allelic values of the specific alleles, this leads to the same likelihood as if we had not restricted the IBD sharing patterns. The three terms in the aforementioned likelihood are given in the supplementary data. The ML solution is found using an Expectation-Maximization algorithm (see Supplementary Data) which is accelerated using the squared iterative method S3 (Varadhan and Roland, 2008).

3 RESULTS AND DISCUSSION

To assess RelateAdmix, we simulated data and applied the tools RelateAdmix, REAP, KING and Plink. Using allele frequencies from 104 290 single nucleotide polymorphisms from a HapMap phase 3 European and African population, we simulated 1600 individuals with different relationships and varying degree of admixture (see Supplementary Data for details). We then used the software Admixture (Alexander *et al.*, 2009) to estimate admixture proportions for all the individuals and allele frequencies for the two populations and used these estimates and the genetic data as input to RelateAdmix and REAP. For Plink and KING, we only used the genetic data as input. Finally, we used RelateAdmix, REAP and Plink to estimate R between all pairs of individuals, in total 1 279 200 pairs, see Figure 2. Additional plots of the kinship coefficient, also including a plot for KING, which provides estimates of the kinship coefficient but not R, are shown in the Supplementary Material. The results show that based on R all the methods can identify the closely related individuals, but that Plink and REAP have problems separating the unrelated individuals from the more distantly related individuals. Based on the kinship estimates both REAP and RelateAdmix can distinguish all relationships, whereas Plink and KING still have problems. The results also show that RelateAdmix gives more accurate R estimates for all relationships.

Funding: The Danish Council of independent research and the Villum Foundation.

Conflict of Interest: none declared.

REFERENCES

Albrechtsen,A. *et al.* (2009) Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genet. Epidemiol.*, **33**, 266–274.

Alexander,D.H. *et al.* (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.*, **19**, 1655–1664.

Manichaikul,A. *et al.* (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics*, **26**, 2867–2873.

Milligan,B.G. (2003) Maximum-likelihood estimation of relatedness. *Genetics*, **163**, 1153–1167.

Purcell,S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

Ritland,K. (1996) Estimators for pairwise relatedness and individual inbreeding coefficients. *Genet. Res.*, **67**, 175–185.

Rohlfs,R.V. *et al.* (2012) Familial identification: population structure and relationship distinguishability. *PLoS Genet.*, **8**, e1002469.

Thompson,E.A. (1975) The estimation of pairwise relationships. *Ann. Hum. Genet.*, **39**, 173–188.

Thornton,T. *et al.* (2012) Estimating kinship in admixed populations. *Am. J. Hum. Genet.*, **91**, 122–138.

Varadhan,R. and Roland,C. (2008) Simple and globally convergent methods for accelerating the convergence of any EM algorithm. *Scand. J. Stat.*, **35**, 335–353.