NgsRelate

a software tool for estimating pairwise relatedness from next-generation sequencing data
Korneliussen, Thorfinn Sand; Moltke, Ida

Published in:
Bioinformatics (Oxford, England)

DOI:
10.1093/bioinformatics/btv509

Publication date:
2015

Document version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Motivation: Pairwise relatedness estimation is important in many contexts such as disease mapping and population genetics. However, all existing estimation methods are based on called genotypes, which is not ideal for next-generation sequencing (NGS) data of low depth from which genotypes cannot be called with high certainty.

Results: We present a software tool, NgsRelate, for estimating pairwise relatedness from NGS data. It provides maximum likelihood estimates that are based on genotype likelihoods instead of genotypes and thereby takes the inherent uncertainty of the genotypes into account. Using both simulated and real data, we show that NgsRelate provides markedly better estimates for low-depth NGS data than two state-of-the-art genotype-based methods.

Availability: NgsRelate is implemented in C++ and is available under the GNU license at www.popgen.dk/software.

Contact: ida@binf.ku.dk

Supplementary information: Supplementary data are available at Bioinformatics online.
and weighing each according to their GLs. is taken into account by summing over all possible true genotypes ML method in Choi all but the true genotype and in that case the method reduces to the that if the genotypes are known with certainty the GLs will be 0 for minimized then becomes a composite likelihood function. We also note that if the assumption of independence between loci is violated, since the function that is optimized then becomes a composite likelihood function. Like all other ML estimators, this estimator is consistent and we note that this is also true if the assumption of independence between loci is violated, since the function that is optimized then becomes a composite likelihood function. We also note that if the genotypes are known with certainty the GLs will be 0 for all but the true genotype and in that case the method reduces to the ML method in Choi et al. (2009). In all other cases the uncertainty is taken into account by summing over all possible true genotypes and weighing each according to their GLs.

3 Results and discussion

To test NgsRelate we used both simulated and real data. We first simulated NGS data for 100 000 diallelic loci from 100 pairs of individuals from each of the relationships: parent–child, full siblings, half-siblings, first cousins and unrelated individuals. To make it possible to assess how NgsRelate’s performance depends on average sequencing depth we simulated such data for five different average depths ranging from low (1, 2 and 4x) over medium (8x) to relatively high depth (16x). From the simulated data we calculated GLs, which we applied NgsRelate to. We also called genotypes based on the maximum GLs and applied the genotype-based ML method from Choi et al. (2009) and PLINK (Purcell et al. 2007) to these called genotypes. See Supplementary Data for details. The simulations showed that all three methods perform well on high-depth data, but that the two genotype-based methods did not provide accurate estimates of R for the related pairs based on low- and medium-depth data (Fig. 1). Further inspection of the results revealed that for all the related pairs these two methods tend to overestimate k0 and thereby make the pairs look less related (Supplementary Figs S1–S5). NgsRelate on the other hand performs well on medium and low-depth data down to 4x (Fig. 1). Even for 2x data it is only slightly biased (Supplementary Figs S1–S5) and for 1x it has large variance, yet it still performs markedly better than the other two methods (Fig. 1). Hence, the simulations suggest that for low-depth NGS data NgsRelate outperforms the two genotype-based methods.

To assess if this holds true for real data we then applied the three methods to low-depth (~4x) NGS data from six genomes from the 1000 Genomes Project Consortium (2012). These individuals have also been SNP chip genotyped (International HapMap 3 Consortium, 2010), and six of the pairs have been reported to be related. We applied NgsRelate to GLs calculated from the low-depth NGS data using ANGSD and applied the two other methods to genotypes called from these GLs. To limit the amount of genotype calling errors only data from sites with depth above 2 in both genomes and a minor allele frequency above 0.05 were included in the genotype-based analyses. For all six-related pairs the estimates from NgsRelate differed markedly less from the ‘true’ values (Fig. 2 and Supplementary Fig. S6), e.g. the difference in k0 ranged from 0.002 to 0.031 for NgsRelate, whereas they ranged from 0.081 to 0.31 for genotype-based ML estimator and from 0.096 to 0.25 for PLINK. In all cases k0 was overestimated, though, note that the opposite was observed for PLINK when we changed the quality filtering of the genotypes (Supplementary Data), suggesting that estimates from the
genotype-based methods depend highly on filtering choices. However, all the real data results supported the conclusion from the simulations: for low-depth NGS data NgsRelate provides more accurate estimates.

**Funding**

TSK was funded by the Danish National Research Foundation (DNRF94), and IM was funded by the Danish Council for Independent Research (DFF-YDUN).

**Conflict of Interest**: none declared.

**References**


