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):
Sequence analysis

NgsRelate: a software tool for estimating pairwise relatedness from next-generation sequencing data

Thorfinn Sand Korneliussen¹ and Ida Moltke²,*

¹Center for GeoGenetics and ²Department of Biology, University of Copenhagen, 2200 Copenhagen, Denmark

*To whom correspondence should be addressed.

Received on June 23, 2015; revised on August 6, 2015; accepted on August 24, 2015

Abstract

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.

1 Introduction

Estimation of how related two individuals are from genetic data plays a key role in several research areas, including medical genetics and population genetics. For example, in medical genetics it is used for excluding closely related individuals from association studies and thereby to avoid inflated false positive rates. How related two individuals are is usually described through the concept of identity-by-descent (IBD), i.e. genetic identity due to a recent common ancestor. Historically, several summary statistics have been used, such as the kinship coefficient \( h \), however almost all of these statistics can be calculated from \( R = (k_0, k_1, k_2) \), where \( k_m \) is the fraction of genome in which the two individuals share \( m \) alleles IBD. For example \( \theta = \frac{k_1}{2} + \frac{k_2}{4} \). We will therefore here focus on \( R \).

Many estimators for \( R \) have been proposed, both method of moments (Purcell et al., 2007; Ritland, 1996) and maximum likelihood (ML) estimators (Thompson, 1975). Common to them all is that they are based on genotype data and it has been shown that they work well on single nucleotide polymorphism (SNP) chip data. However, next-generation sequencing (NGS) is becoming increasingly common and often NGS data are only of low depth, which means that genotypes can only be called with high uncertainty (O’Rawe et al., 2015). For such data it has been shown that it can be an advantage to take the uncertainty of the genotypes into account by basing statistical methods on so-called genotype likelihoods (GLs), instead of genotypes (Skotte et al., 2013). Motivated by this we developed NgsRelate; a ML method for estimating the pairwise relatedness parameter \( R \) from NGS data based on GLs. In the following, we present this method and show that for low-depth NGS data it performs markedly better than two state-of-the-art genotype-based methods.

2 Methods

To estimate \( R \) for two non-inbred individuals \( i \) and \( j \) we use the following probabilistic framework: Let \( D' = (D_1', D_2', \ldots, D_L') \) and \( D' = (D_1', D_2', \ldots, D_L') \) denote the observed NGS data for \( i \) and \( j \) at \( L \) diallelic loci and \( G' = (G_1', G_2', \ldots, G_L') \) and \( G' = (G_1', G_2', \ldots, G_L') \).
is taken into account by summing over all possible true genotypes
all but the true genotype and in that case the method reduces to the
dependence between loci is violated, since the function that is opti-
consistency and we note that this is also true if the assumption of in-
(Supplementary Data). Like all other ML estimators, this estimator
hood function with an Expectation Maximization algorithm
medium-depth data (Fig. 1). Further inspection of the results revealed
from Choi
the maximum GLs and applied the genotype-based ML method
which we applied NgsRelate to. We also called genotypes based on
called genotypes. See Supplementary Data for details. The simula-
tively high depth (16
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 or from SNP chip data. NgsRelate provides ML
estimates of R by finding the value of R that maximizes this likeli-
hood function with an Expectation Maximization algorithm (Supplementary Data). Like all other ML estimators, this estimator
is consistent and we note that this is also true if the assumption of in-
dependence between loci is violated, since the function that is optim-
ized 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 di allelic loci from 100 pairs of indi-
viduals from each of the relationships: parent–child, full siblings,
half-siblings, first cousins and unrelated individuals. To make it pos-
sible 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 4\times) over medium (8\times) to rel-
atively high depth (16\times). 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 simula-
tions showed that all three methods perform well on high-depth data,
but that the two genotype-based methods did not provide ac-
curate 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
\(k_0\) 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 4\times (Fig. 1). Even for 2\times data it is only
slightly biased (Supplementary Figs S1–S5) and for 1\times 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 (~4\times) NGS data from six genomes from the
1000 Genomes Project Consortium (2012). These individuals have

\(G_j \) denote the true unobserved genotypes at the \(L\) loci. Further, let
\(X_i \in \{0, 1, 2\}\) denote the unobserved number of alleles \(i\) and \(j\)
share IBD at locus \(l\). Finally, let the two alleles at each locus be
denoted \(A\) and \(a\) and the frequencies of the \(A\) alleles be denoted
\(f^A = (f^A_1, f^A_2, \ldots, f^A_L)\). Then, assuming the loci are independent
and that \(f^A\) is known the likelihood function for \(R\), can be written:

\[
L(R|D', D', f^A) = \prod_{l=1}^{L} \sum_{m=0}^{L} \sum_{|G_j|=0} \prod_{l=1}^{L} \sum_{m=0}^{L} \sum_{|G_j|=0} P(D_l'|D_l, G_j) P(D_l'|G_j) P(G_j|f^A) P(G_j|f^A, X_l = m, G_j)
\]

with \(P(X_l = m|R) = k_m\) and

\[
P(D_l'|D_l, G_j) = \sum_{G_j, G_j|G_j|G_j| \in [0, 1, 2]} P(D_l'|G_j) P(D_l'|G_j) P(G_j|f^A) P(G_j|f^A, X_l = m, G_j)
\]

Here \(P(D_l'|G_j)\) and \(P(D_l'|G_j)\) are GLs, which can be estima-
ted using ANGSD (Korneliussen et al., 2014) and \(P(G_j|f^A)\) and
\(P(G_j|f^A, X_l = m, G_j)\) are given in Supplementary Table S1–S2. \(f^A\)
and major and minor alleles can be precalculated from NGS data
using ANGSD or from SNP chip data. NgsRelate provides ML
estimates of \(R\) for 6 of each combination of four relationship types and five average
sequencing depths 1, 2, 4, 8 and 16 (see Supplementary Fig. S5 for results for
unrelated pairs). For each combination estimates were obtained with
NgsRelate (left), genotype-based ML (middle) and PLINK (right). RMSD will
be zero if the estimate is equal to the simulated \(R\).

Fig. 1. Root mean square deviation (RMSD) between estimated and simulated \(R\) for 100 of each combination of four relationship types and five average
sequencing depths 1, 2, 4, 8 and 16 (see Supplementary Fig. S5 for results for
unrelated pairs). For each combination estimates were obtained with
NgsRelate (left), genotype-based ML (middle) and PLINK (right). RMSD will
be zero if the estimate is equal to the simulated \(R\).

Fig. 2. RMSD between the estimated and the true \(R\) for six pairs of ~4\times gen-
omes. RMSD will be 0 if the estimate is equal to the true \(R\)
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**


