Inferring identity-by-descent sharing between pairs of individuals from Next Generation Sequencing Data.

Thorfinn Sand Korneliussen,[1] and Ida Moltke,[2,3,*]

[1]The Center for Geogenetics , University of Copenhagen, Copenhagen, Denmark

[2]Department of Human Genetics, University of Chicago, Chicago, Il, USA;

[3]The Bioinformatics Centre, University of Copenhagen, Copenhagen, Denmark

*Corresponding author, e-mail: ida@binf.ku.dk

October 22, 2012

Abstract

Inference of identity-by-descent (IBD) sharing between pairs of individuals has in the last few years received renewed attention. Several new inference methods have been proposed and it has successfully been shown that IBD inference results can be used for a range of purposes including identification of disease causing genes. However, all current methods base their inference on genotype data. With the emergence of vast amounts of Next Generation sequencing data this is not always ideal, especially not when only low depth sequencing data is available and thus genotypes can only be called with high uncertainly. Here we present a new method for identifying IBD tracts among pairs of individuals from Next Generation Sequencing (NGS) data. In contrast to all previous methods this new method is applicable directly to genotype likelihoods, instead of to genotype data. This allows the new method to take any uncertainty of genotypes into account. We show that this approach leads to a more accurate IBD inference in some cases and explore when(??). Finally we provide a publicly available implementation of the method so others can use it for future analyses.

Introduction

Identity-by-descent (IBD) is an old concept in genetics, first introduced by Cotterman (2012) and Malecot (1948), which has received a renewed attention the past few years (Purcell et al 2008, Browning and Browning 2008....). IBD basically means sequence identity due to common ancestry, and two alleles are said to be shared IBD if they are identical because the have been inherited from a commen ancestor. Different studies use slightly different definition the main difference being how far back in time common ancestry is considered. We here follow the definition used in xxxx ... [specify what that means]. With this definition IBD has been shown to be useful in disease mapping [3, 4], for detection selection [5], for phasing data [**?**].

All current methods for IBD inference are developed for genotype data, e.g. for SNP chip data. However, when you have Next Generation Sequencing (NGS) data, called genotypes are not always the best type of data to base statistical methods on. Especially when the depth is low it can be an advantage to take the uncertainty of the genotypes into account by basing the statistical methods directly on the genotype likelihoods, instead of first using these likelihoods for genotype calling and then running the inference methods based on these called genotypes (cf. e.g. [**?**]). Here we present a method for IBD inference that is based directly on genotype likelihoods. And we show that this approach is advantageous when??....

[Intro should be longer but will writ the rest later...]

Methods and Materials

In any one genetic locus two non-inbred individuals will share either 0,1 or 2 alleles IBD. Based on NGS data we want to be able to infer the proportions in which two given individuals are in each of these 3 possible IBD states. Also, we want to be able to infer their IBD sharing state in any specific locus.

We will first briefly describe how IBD information can be inferred from genotype data. More specifically we will describe a method very close to the one presented in [3] (a version without LD correction). We then adjust this method so it is applicable to genotype likelihoods instead and thus describe our a new method.

Basic model

Assume we know the genotypes in L diallelic loci of two individuals i and j and let $G^i = (G^i_1, G^i_2, ..., G^i_L)$ and $G^j = (G^j_1, G^j_2, ..., G^j_L)$ denote these genotypes. Also, let $X = (X_1, X_2, ..., X_L)$ denote the number of alleles the two individuals share and $\Theta = (k_0, k_1, k_2)$ denote the overall proportions of the genome in which they share 0,1 and 2 alleles IBD respectively. Finally, let the two alleles in every locus l be denoted A and a and the frequencies of the A alleles be denoted $f^A = (f^A_1, f^A_2, ..., f^A_L)$. Then, if we assume that the L loci are independent and that the genotypes are known with no uncertainty it must hold that

$$\begin{aligned}
P(G^i, G^j|\Theta, f^A) &= \prod_{l=1}^{L} P(G^i_l, G^j_l|\Theta, f^A_l) \\
&= \prod_{l=1}^{L} \sum_{x_l \in \{0,1,2\}} P(X_l = x_l|\Theta, f^A_l) P(G^i_l, G^j_l|\Theta, f^A_l, X_l = x_l) \\
&= \prod_{l=1}^{L} \sum_{x_l \in \{0,1,2\}} P(X_l = x_l|\Theta) P(G^i_l, G^j_l|f^A_l, X_l = x_l) \\
&= \prod_{l=1}^{L} \sum_{x_l \in \{0,1,2\}} P(X_l = x_l|\Theta) P(G^i_l|f^A_l, X_l = x_l) P(G^j_l|f^A_l, X_l = x_l, G^i_l) \\
&= \prod_{l=1}^{L} \sum_{x_l \in \{0,1,2\}} P(X_l = x_l|\Theta) P(G^i_l|f^A_l) P(G^j_l|f^A_l, X_l = x_l, G^i_l)
\end{aligned}$$

where $P(X_l = x_l|\Theta)$ is simply:

$$P(X_l{=}x_l|\Theta) = \begin{cases} k_0 & \text{if } x_l{=}0 \\ k_1 & \text{if } x_l{=}1 \\ k_2 & \text{if } x_l{=}2 \end{cases} \tag{1}$$

and where $P(G^i_l|f^A_l)$ and $P(G^j_l|f^A_l, X_l = x_l, G^i_l)$ are given in table 1 and table 2 respectively.

Based on this simple model we can get maximum likelihood (ML) estimates of Θ using e.g. a numerical optimization algorithm such as the BFGS algorithm [**?**]. However, unless the L are very far apart the IBD states of the loci that are close to each other will not be independent. When they are not, the equation above can be seen as a composite likelihood function and the ML estimation will still be unbiased and consistent. But,

AA	Aa	aa
$(f_l^A)^2$	$2f_l^A f_l^a$	$(f_l^a)^2$

Table 1: The probability that the true genotype of individual i n locus l is G_l^igiven that the population frequency of allele A in l is $f_l^A (= 1 - f_l^a)$.

G_l^i	G_l^j	X_l=0	X_l=1	X_l=2
AA	AA	$(f_l^A)^2$	f_l^A	1
AA	Aa	$2f_l^A f_l^a$	f_l^a	0
AA	aa	$(f_l^a)^2$	0	0
Aa	AA	$(f_l^A)^2$	$\frac{1}{2}f_l^A$	0
Aa	Aa	$2f_l^A f_l^a$	$\frac{1}{2}f_l^A + \frac{1}{2}f_l^a$	1
Aa	aa	$(f_l^a)^2$	$\frac{1}{2}f_l^a$	0
aa	AA	$(f_l^A)^2$	0	0
aa	Aa	$2f_l^A f_l^a$	f_l^A	0
aa	aa	$(f_l^a)^2$	f_l^a	1

Table 2: The probability that the true genotype of individual j in locus l is G_l^j, given that the true genotype of individual i in this locus is G^i, that i and j share X_l alleles IBD and that the population frequency of allele A the locus is f_l^A $(= 1 - f^a)$.

	$O_l^k = AA$	$O_l^k = Aa$	$O_l^k = aa$
$G_l^k = AA$	$(1-\epsilon)^2$	$2(1-\epsilon)^2\epsilon$	ϵ^2
$G_l^k = Aa$	$(1-\epsilon)\epsilon$	$(1-\epsilon)^2 + \epsilon^2$	$(1-\epsilon)\epsilon$
$G_l^k = aa$	ϵ^2	$2(1-\epsilon)^2\epsilon$	$(1-\epsilon)^2$

Table 3: The probability of observing the genotype O_l^k in individual k in a locus l given that the true genotype is G_l^k and the rate of error for any one allele in any locus is ϵ.

significance testing using a simple likelihood ratio test is not possible. And perhaps more importantly not all information is used: the dependencies can be used for inference by taking into account the length distribution of the IBD tracts. This can be done using a hidden Markov model framework as described in [3]:

More specifically we assume that the IBD states can be modeled as continuous time Markov chain with instantaneous rate matrix

$$Q = \begin{pmatrix} -\alpha k_1 & \alpha k_1 & 0 \\ \alpha k_0 & -\alpha(k_0 + k_2) & \alpha k_2 \\ 0 & \alpha k_1 & -\alpha k_1 \end{pmatrix}$$

where α is the overall rate of change of the Markov chain. This basically corresponds to assuming that the length of all IBD sharing tracts are exponentially distributed. If we also assume that the genotypes are independent conditional on the IBD states, i.e. that there is no linkage disequilibrium (LD) in the data it must hold that

$$\begin{aligned}
& P(G^i, G^j|\Theta, f^A, \alpha) && (2)\\
= & \sum_x P(X = x|\Theta, f^A, \alpha) P(G^i, G^j|\Theta, f^A, \alpha, X = x) && (3)\\
= & \sum_x P(X = x|\Theta, \alpha) P(G^i, G^j|f^A, X = x) && (4)\\
= & \sum_x P(X_1 = x_1|\Theta)(\prod_{l=2}^{L} P(X_l = x_l|\Theta, \alpha))(\prod_{l=1}^{L} P(G^i_l, G^j_l|f^A_l, X_l = x_l) && (5)\\
= & \sum_x P(X_1 = x_1|\Theta)(\prod_{l=2}^{L} P(X_l = x_l|\Theta, \alpha))(\prod_{l=1}^{L} P(G^i_l|f^A_l) P(G^j_l|f^A_l, X_l = x_l, G^i_l) && (6)
\end{aligned}$$

And based on this model one can achieve ML estimates of parameters $\Theta = k_0, k_1, k_1$ and α using ML estimation. And not only that once these estimates are achieved we can make inferences about exactly where their genomes the two individuals share 0, 1 and two alleles IBD using standard HMM inference algorithms such a viterbi and the forward backwards algorithm. Note that this is the model in [3] without LD correction (see discussion for comments on LD) and without a model for genotyping errors.

When genotypes are not known with certainty

In the basic models we assumed that the genotypes were known with certainty. However this is rarely the case. Neither for SNPs chip genotype data nor for NGS data. In the case of SNP chip data this is usually accommodated for by assuming some fixed allelic error rate, ϵ, for all sites. E.g. in the model of [3] the above model is augmented by calculating the probability of the observed genotypes $O^i = (O^i_1, O^i_2, ..., O^i_L)$ and $O^i = (O^j_1, O^j_2, ..., O^j_L)$ instead of the probability of the true genotypes G^i and G^j. And this is done by for each locus summing over all possible values of the true underlying genotype pair (G^i_l,G^j_l) so equation 6 instead becomes:

$$P(O^i, O^j|\Theta, f^A, \alpha, \epsilon) =$$

$$\sum_x P(X_1 = x_1|\Theta)(\prod_{l=2}^{L} P(X_l = x_l|\Theta, \alpha))(\prod_{l=1}^{L} P(O_l^i, O_l^j|f_l^A, X_l = x_l, \epsilon)$$

$$\sum_x P(X_1 = x_1|\Theta)(\prod_{l=2}^{L} P(X_l = x_l|\Theta, \alpha))(\prod_{l=1}^{L} \sum_{(G_l^i, G_l^{j,})} P(G_l^i, G_l^j|f_l^A, X_l = x_l)P(O_l^i, O_l^j|G_l^i, G_l^j, \epsilon)$$

$$\sum_x P(X_1 = x_1|\Theta)(\prod_{l=2}^{L} P(X_l = x_l|\Theta, \alpha))(\prod_{l=1}^{L} \sum_{P(G_l^i, G_l^j)} P(G_l^i|f_l^A)P(G_l^j|f_l^A, X_l = x_l, G_l^i)P(O_l^i|G_l^i, \epsilon)P(O_l^i|G_l^i, \epsilon)$$

where $P(G_l^i|f_l^A)$ and $P(G_l^j|f_l^A, X_l = x_l, G_l^i)$ is calculated exactly as before and $P(O_l^i|G_l^i, \epsilon)$ and $P(O_l^j|G_l^j, \epsilon)$ is calculated using table 3.

However, for NGS data the error rates for called genotypes will depend highly on e.g local read depth. Which is the main motivation for our new method. We suggest to accommodate for errors in a different manner, for this type of data, where genotypes are usually called based on the likelihood of all the possible genotypes. More specifically we suggest to take genotype uncertainty into account by making inference based directly on genotype likelihoods. We can do that using a model very similar to the genotype based model just described.

Let $D^i = (D_1^i, D_2^i, ..., D_L^i)$ and $D^j = (D_1^j, D_2^j, ..., D_L^j)$ be the observed data (sequence reads) in the 2 individuals, then

$$\begin{aligned}
& P(D^i, D^j|\Theta, f^A, \alpha) \\
= & \sum_x P(X = x|\Theta, f^A, \alpha)P(D^i, D^j|\Theta, f^A, \alpha, X = x) \\
= & \sum_x P(X = x|\Theta, \alpha)P(D^i, D^j|f^A, X = x) \\
= & \sum_x P(X_1 = x_1|\Theta)(\prod_{l=2}^{L} P(X_l = x_l|\Theta, \alpha))(\prod_{l=1}^{L} P(D_l^i, D_l^j|f_l^A, X_l = x_l) \\
= & \sum_x P(X_1 = x_1|\Theta)(\prod_{l=2}^{L} P(X_l = x_l|\Theta, \alpha))(\prod_{l=1}^{L} \sum_{(G_l^i, G_l^j) \in \{0,1,2\}^2} P(D_l^i, D_l^j|G_l^i, G_l^j)P(G_l^i, G_l^j|f_l^A, X_l = x_l) \\
= & \sum_x P(X_1 = x_1|\Theta)(\prod_{l=2}^{L} P(X_l = x_l|\Theta, \alpha))(\prod_{l=1}^{L} \sum_{(G_l^i, G_l^j) \in \{0,1,2\}^2} P(D_l^i|G_i^i)P(D_l^j|G_l^j)P(G_l^i|f_l^A)P(G_l^j|f_l^A, X_l = x_l, G_l^i)
\end{aligned}$$

Note that the only difference between this model and the error augmented genotype based model is that the probabilities of the form $P(D_l^i|G_i^i)$ are not based on a fixed error rate, ϵ, that is common for all loci. Instead we use the genotype likelihoods that

genotype calling in NGS data is usually based on and thus take into account that the uncertainty of genotype varies a lot from locus to locus. [comment: In the exact same manner the full model from [3] that can accommodate for LD can be adjusted to work on NGS genotype likelihoods – do we want to do that in this paper? If not we have to argue for why we do not do it in the discussion]

Materials

In order to test our new method we used both simulated data and real data.

Simulated data

How did we simulate data...

Simulated data

What real dataset did we apply the data to?

1 Results

....

2 Discussion

LD, comparison of the two error models,when is it beneficial?

References

[1] Patterson N, Price AL, Reich D (2006) Population Structure and Eigenanalysis PLoS Genet *2*(12): e190

[2] Leutenegger, A. L. and Prum, B. and Genin, E. and Verny, C. and Lemainque, A. and Clerget-Darpoux, F. and Thompson, E. A. (2003) Estimation of the inbreeding coefficient through use of genomic data", Am. J. Hum. Genet. *73* 516–523.

[3] Albrechtsen, A., Sand Korneliussen, T., Moltke, I., van Overseem Hansen, T., Nielsen, F. C., and Nielsen, R. (2009). Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. Genet. Epidemiol. *33*, 266–274.

[4] Moltke, I., Albrechtsen, A., van Overseem Hansen, T., Nielsen, F. C., and Nielsen, R. (2011). A method for detecting IBD regions simultaneously in multiple individuals – with applications to disease genetics.. Genome Res. *21*, 1168–1180.

[5] Albrechtsen, A., Moltke, I., and Nielsen, R. (2010). Natural Selection and the Distribution of Identity-by-Descent in the Human Genome. Genetics *186*, 295–308.