Genetic Epidemiology 33:266–274 (2009)

Relatedness Mapping and Tracts of Relatedness for Genome-Wide Data in the Presence of Linkage Disequilibrium

Anders Albrechtsen,[1]* Thorfinn Sand Korneliussen,[2] Ida Moltke,[3] Thomas van Overseem Hansen,[4] Finn Cilius Nielsen,[4] and Rasmus Nielsen[5,6]

[1]*Department of Biostatistics, Copenhagen University, Copenhagen, Denmark*
[2]*Computer Science, Copenhagen University, Copenhagen, Denmark*
[3]*Bioinformatics Centre, Copenhagen University, Copenhagen, Denmark*
[4]*Department of Clinical Biochemistry, Rigshopitalet, Copenhagen, Denmark*
[5]*Department of Integrative Biology, University of California Berkeley, Berkeley, California*
[6]*Department of Statistics, University of California Berkeley, Berkeley, California*

Estimates of relatedness have several applications such as the identification of relatives or in identifying disease related genes through identity by descent (IBD) mapping. Here we present a new method for identifying IBD tracts among individuals from genome-wide single nucleotide polymorphisms data. We use a continuous time Markov model where the hidden states are the number of alleles shared IBD between pairs of individuals at a given position. In contrast to previous methods, our method accurately accounts for linkage disequilibrium using pairwise haplotype probabilities.

The method provides a map of the local relatedness along the genome. We illustrate the potential of the method for mapping disease genes on a real data set, and show that the method has the potential to map causative disease mutations using only a handful of affected individuals. The new IBD mapping method provides considerable improvement in mapping power in natural populations compared to standard association mapping methods. *Genet. Epidemiol.* 33:266–274, 2009. © 2008 Wiley-Liss, Inc.

Key words: identity by descent; relatedness; hidden Markov model; linkage; association; complex disease; genome-wide analysis; SNP

Contract grant sponsor: The Danish Research Consul; Contract grant sponsor: The Center for Phamacogenomics; Contract grant sponsor: The Neye Foundation.
*Correspondence to: Anders Albrechtsen, Department of Biostatistics, University of Copenhagen, Oester Farimagsgade 5, Entr. B, P.O.B. 2099, DK-1014 Copenhagen, Denmark. E-mail: albrecht@binf.ku.dk
Received 9 September 2008; Revised 9 September 2008; Accepted 10 September 2008
Published online 21 November 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20378

INTRODUCTION

Linkage mapping using multiple families has been a powerful tool for identifying rare and highly penetrant genetic variants. Most linkage studies require a large number of families with affected individuals to map the disease causing variant, and even so, the causative variant may only be mapped to a larger genomic region [Hirschhorn and Daly, 2005]. However, in recent years genome-wide association mapping using unrelated individuals and high throughput single nucleotide polymorphism (SNP)-chips has been successful in identifying variants with a relative low penetrance [Kingsmore et al., 2008]. Association mapping can be more powerful for identifying variants with a lower penetrance but the causative variant needs to be in high linkage disequilibrium (LD) with one or more markers genotyped in the study. To obtain appreciable mapping power to detect common alleles with low penetrance in genome-wide association studies, thousands of individuals are typically needed [Hirschhorn and Daly, 2005]. The challenges involved in classical approaches of linkage and association mapping, and related methods, have prompted a renewed interest in alternative methods for mapping genes in natural populations, including admixture mapping [Smith and O'Brien, 2005] and identity by descent (IBD) mapping [Cheung and Nelson, 1998; Service et al., 1999].

In a finite homogeneous population all individuals will be related to some extent. Individuals sharing a founder mutation will have higher degree of relatedness, or IBD, in the region around the founder mutation. In case-control samples, increased relatedness in a region only among the affected individuals is evidence in favor of the location of a causative mutation in that region. Identification of relatedness is, therefore, of great interest in medical genetics, in addition to forensic genetics [Evett and Weir, 1998], and molecular ecology [Thompson, 1975; Queller and Goodnight, 1989; Ritland, 1996; Lynch and Ritland, 1999].

As demonstrated by Purcell et al. [2007] a dense set of uncorrelated markers can be used to infer tracts of relatedness even if the individuals are not closely related. Their method is very well suited for situations where there is an abundance of SNPs in allowing many SNPs to remain

when SNPs that are correlated or contain missing data are pruned away.

The goal of this article is to describe a method for relatedness tract estimation that can be applied without the need for eliminating any SNPs while accounting for genotyping errors, missing data, and LD without pruning away SNPs.

Alleles are called identical by state if the same allele is observed between chromosomes. These alleles are also called IBD if the alleles are direct copies from the same ancestral allele. Mathematically IBD means that the probability of observing an allele in one individual is not independent from observing an allele in another individual. It is this property that is exploited in linkage analysis. Pairwise relatedness is measured in probabilities of alleles from the two individuals being IBD. Using Thompson's maximum likelihood method [Thompson, 1975], we can estimate the relatedness for pairs of individuals using three Jacquard coefficients [Jacquard, 1974], ω_0, ω_1, and ω_2, where $\sum_{i=0}^{2} \omega_i = 1$. Here and throughout the article we assume that the subjects are not inbred that is we assume that alleles within an individual are not IBD. ω_0 can be understood as the probability of an allele at random loci in the genome is unrelated, ω_1 the probability of having one allele IBD, and ω_2 the probability that both alleles are IBD. For known pedigrees the expectation of these coefficients are known e.g. for two full siblings the expected IBD sharing is $\omega_0 = 0.25$, $\omega_1 = 0.5$, and $\omega_2 = 0.25$. When the relationship is unknown these coefficients can be estimated using polymorphic genetic data, e.g., SNPs [Milligan, 2003]. In this article we present a method for estimating the probability of IBD between pairs of individuals at a specific locus using a dense set of SNPs possibly in LD. We will also demonstrate that these estimates can form the basis for a potentially very powerful method for mapping disease genes for very distant relatives with unknown pedigrees.

MATERIALS AND METHODS

Most methods for estimating relatedness ignore the dependencies in IBD state from adjacent loci, that is bound to arise from a finite number of recombination events between individuals. One way to understand this dependency is through a Markov process, where the IBD sharing state of a locus depends only on the alleles at that locus and the state of the previous locus. Since only a small fraction of loci are genotyped (and polymorphic), it is convenient to describe the process using a continuous time Markov model where the distance (or time) is measured in Centi-Morgan (cM) or approximated using the genomic position in base pairs.

As noted by [Feingold, 1993] even for simple pedigrees, IBD tract lengths do not necessarily follow Markov process. However, McPeek and Sun [2000] showed that the incorrect Markov process may be a good approximation and provides a useful tool for inference purposes. Therefore, like Boehnke and Cox [1997] and Epstein et al. [2000] we will, as an approximation, assume a Markov process. It should be pointed out that McPeek and Sun [2000] only evaluated closely related individuals while we also make a Markov assumption for distantly related individuals.

For simplicity we assume biallelic SNP markers with genotypic states in $\Phi = \{0, 1, 2\}$, where the numbers refer to the count of one of the alleles.

CONTINUOUS TIME MARKOV CHAIN

We explore a model with three states: zero, one, or two alleles shared IBD between a pair of individuals, but do not allow direct transitions between state IBD = 0 and IBD = 2 for infinitesimal distances. We have, therefore, chosen to parameterize the instantaneous rate matrix, Q as

$$Q = \begin{pmatrix} -\alpha\omega_1 & \alpha\omega_1 & 0 \\ \alpha\omega_0 & -\alpha(\omega_0 + \omega_2) & \alpha\omega_2 \\ 0 & \alpha\omega_1 & -\alpha\omega_1 \end{pmatrix} \quad (1)$$

where the rows and columns correspond to the states $IBD = 0$, $IBD = 1$, and $IBD = 2$ so that the instantaneous rate of going from $IBD = 1$ to $IBD = 0$ is $\alpha\omega_0$. The matrix is parameterized, so that the stationary distribution is given by the Jacquard coefficients and α is a parameter that decides how fast the chain changes states. The time-dependent transition probability matrix can then be found analytically using Kolmogorov's forward equations. These can be seen in Appendix A. The likelihood for the pairwise relatedness between individuals j and k assuming a first-order Markov chain is

$$\begin{aligned} P(G^{j,k}|\Omega, \alpha) = \sum_{\mathbf{x}} & \left(\prod_{i=0}^{m} P(G_i^{j,k}|X_i = \mathbf{x_i}) \right) \\ & \times \left(\prod_{i=1}^{m} P(X_i = \mathbf{x_i}|X_{i-1} = \mathbf{x_{i-1}}, \Omega, \alpha) \right) \\ & \times P(X_0 = \mathbf{x}_0|\Omega) \end{aligned} \quad (2)$$

where $\Omega = \{\omega_0, \omega_1, \omega_2\}$, X_i is the IBD state of the ith marker, m is the number of markers, $G_i^{j,k} \in \Phi^2$ are the genotypes for individual j and k at position i and $\mathbf{x}$ is all possible IBD paths through the chain. This is generalized for multiple autosomal chromosomes by assuming an infinite distance between markers from different chromosomes. $P(X_0 = \mathbf{x}_0|\Omega)$ is given by the state distribution of the Markov chain. The likelihood can recursively be calculated using the forward algorithm. The emission probabilities $P(G_i^{j,k}|X_i = x_i)$ are given in Table I and are functions of the state, the observed genotypes, and the probability of observing the alleles in the population. The probability of observing the alleles in the population are estimated as the allele frequencies in a reference population.

Estimating the parameters. In general we will treat α, ω_0, and ω_1 as free parameters while $\omega_2 = 1 - \omega_0 - \omega_1$. However, we note that if the individuals are distantly

TABLE I. Emission probabilities for pairs of individuals

$G_i^{j,k}$	$X_i = 0$	$X_i = 1$	$X_i = 2$	
AA AA	p_A^4	p_A^3	p_A^2	$\forall A$
AA aa	$2p_A^2p_a^2$	0	0	$A \neq a$
AA Aa	$4p_A^3p_a$	$2p_A^2p_a$	0	$A \neq a$
Aa Aa	$2p_A^2p_a^2$	$p_A^2p_a + p_Ap_a^2$	$2p_Ap_a$	$A \neq a$

Genotype *AA Aa* means that one of the individuals is homozygote for the allele A while the other individual is heterozygous. p_A is the frequency of the A allele in the population and by symmetry $p(G_i^{j,k}|X_i = x) = p(G_i^{k,j}|X_i = x)$.

related and no inbreeding has occurred then the probability of sharing two alleles IBD is zero and for certain pedigrees then α is a uniquely determined function of the number of meiosis and the recombination rate. The number of meiosis can be estimated from the overall IBD sharing [Purcell et al., 2007, see Appendix B for details]. For studies where the hidden Markov model (HMM) is applied to all pairs of individuals we will, for improved computational efficiency, make these two assumptions implying that only ω_0 is a free parameter. In all other cases we do not. The likelihood is maximized using the BFGS algorithm [Byrd and Nocedal, 1995].

LD AND GENOTYPING ERRORS

For a dense set of markers the assumption of independence of the marker data conditional on the hidden state, (2) is violated. Therefore, we explore a model that accommodates LD in a similar fashion as was explored for inbreeding tracts [Wang et al., 2006]. Instead of defining the emission probabilities for only one marker we define the emission probabilities for a marker conditional on a previous marker $P(G_i^{j,k}|G_h^{j,k}, X_i = X_h = x_i)$, where i is the current marker and h is a previous marker. The joint probabilities of observing the genotypes of the two SNPs can be seen in Table V assuming that the current SNP i is in the same state as the previous SNP h. The main difference between this and the method explored by Wang et al. [2006] is that this method can condition on any of the previous SNPs and not just the adjacent one and that we here assume the same state for the two SNPs.

In large-scale genotyping data a small fraction ε of the alleles will be genotyping errors. This can easily be included in the model, so that the conditional emission probabilities including genotyping errors become

$$P(G_i^{j,k}|G_h^{j,k}, X_i = X_h = x_i, \varepsilon) = \frac{P(G_i^{j,k}, G_h^{j,k}|X_i = X_h = x_i, \varepsilon)}{p(G_h^{j,k}|X_h = x_i, \varepsilon)} \tag{3}$$

$$= \frac{\sum_{G_i^{j',k'}, G_h^{j',k'} \in \Phi^4} P(G_i^{j',k'}, G_h^{j',k'}|X_i = X_h = x_i)(\prod_{l\in\{j,k\}} P(G_i^l|G_i^{l'}, \varepsilon))(\prod_{l\in\{j,k\}} P(G_h^l|G_h^{l'}, \varepsilon))}{\sum_{G_h^{j',k'} \in \Phi^2} P(G_h^{j',k'}|X_h = x_i)(\prod_{l\in\{j,k\}} P(G_h^l|G_h^{l'}, \varepsilon))} \tag{4}$$

where $P(G_h^j|G_h^{j'}, \varepsilon)$ is the probability of the observed genotype G_h^j for individual j at locus h given the true genotype $G_h^{j'}$ and can be seen in Table II.

When the SNP data are very dense including the previous SNP may not be enough but including longer haplotypes in the model is computationally hard. Therefore, we estimate the LD as the squared correlation coefficient, r^2, and condition on the SNP among the previous 50 SNPs with the highest amount of LD. This ad hoc procedure will work if the other SNPs are conditionally independent on the SNP with the highest amount of LD. For denser, respectively sparser, data sets more or fewer SNPs should be used.

The haplotype frequencies and the squared correlation coefficient r^2 between markers were estimated as described in Clayton and Leung [2007].

TABLE II. The probability for the observed genotype G_i^j given the true genotype $G_i^{j'}$ and an error rate of ε

$P(G_i^j\|G_i^{j'}, \varepsilon)$	$G_i^j = 0$	$G_i^j = 1$	$G_i^j = 2$
$G_{i'}^{j'} = 0$	$(1-\varepsilon)^2$	$2(1-\varepsilon)\varepsilon$	ε^2
$G_{i'}^{j'} = 1$	$(1-\varepsilon)\varepsilon$	$(1-\varepsilon)^2 + \varepsilon^2$	$(1-\varepsilon)\varepsilon$
$G_i^{j'} = 2$	ε^2	$2(1-\varepsilon)\varepsilon$	$(1-\varepsilon)^2$

IBD MAPPING

For a pair of individuals with genotypes $G_i^{j,k}$ at loci i for individuals j and k we define the function

$$h(G_i^{j,k}) = p(x_i^{j,k} > 0|G^{j,k}).$$

This is the posterior probability of sharing at least one allele IBD between individuals j and k at loci i, which can be calculated using the standard backward-forward algorithm [Rabiner and Juang, 1986]. For n individuals we get $N = n(n-1)/2$ pairwise comparisons for each of the m loci. We calculate the mean posterior probability of sharing at least one allele IBD, $E_i\omega^{j,k}$, for the pair of individuals as

$$E\omega_i^{j,k} = E\omega^{j,k} = \frac{1}{m}\sum_{i=1}^{m} h(G_i^{j,k}). \tag{5}$$

For performing linkage analysis we can either (1) test whether the affected pairs of individuals are more related in certain loci than expected by random given the overall relatedness of the individuals or (2) test whether affected pairs of individuals are more related at a locus than pairs of unaffected individuals. Since a reference population is needed in order to obtain the allele frequency in the population, we can use this as a control group if no better control group is available. Also, as the first approach depends on assumptions regarding the variances and co-variances in relatedness among individuals, the second approach will in most cases be much more robust. For example, some regions of the genome are known to have undergone recent positive selection so individuals in general might be more related in certain regions. Also, if the correction for LD is not perfect then this bias will affect both the cases and the controls.

Here we present a simple test for linkage. Let

$$U_i = \binom{n}{2}^{-1} \sum_{k<j} h(G_i^{j,k}) \quad \text{and} \quad Z_i = \frac{U_i - E(U)}{\sqrt{Var(U)}}.$$

If $G_i^1, G_i^2, \ldots, G_i^n$ are independent and identically distributed, Z_i asymptotically $N(0,1)$ distributed.

However, we will not make this assumption in the following.

For all loci

$$Var(U)N^2 = \sum_{k<j} Var(h(G^{j,k})) + \sum_{j,k,e,f|j\neq e \vee k \neq f} Cov(h(G^{j,k}), h(G^{e,f})).$$

Also, we estimate the variances and co-variances as

$$Var(h(G^{j,k})) = \frac{1}{m-1}\sum_{i=1}^{m} (h(G_i^{j,k}) - E\omega_i^{j,k})^2$$

$$Cov(h(G^{j,k}), h(G^{e,f})) = \frac{1}{m-1}\sum_{i=1}^{m} (h(G_i^{j,k}) - E\omega_i^{j,k})(h(G_i^{e,f}) - E\omega_i^{e,f})$$

and

$$E(U) = \frac{1}{N}\sum_{k,j} E\omega^{j,k}.$$

We use a test for linkage using cases and controls, that can be performed by comparing pairwise relatedness within affected individuals and pairwise relatedness within unaffected.

Let $d \in \{0,1\}^n$ be at $1 \times n$ disease vector containing n_0 unaffected individuals and n_1 affected individuals with corresponding N_0 and N_1 pairwise IBD sharing comparisons. We define a test statistic,

$$Y_i = \frac{U_i^1 - U_i^0 - E(U^1) + E(U^0)}{\sqrt{((N_0-1)Var(U^0) + (N_1-1)Var(U^1))/(N_0+N_1-2)(1/N_0+1/N_1)}} \quad (6)$$

where U^0 and U^1 are the mean pairwise posterior probabilities of sharing at least one allele for the unaffected and affected individuals, respectively. U^1 and U^0 and their variances are estimated as in the previous section using the cases and controls, respectively.

Permutations test. To obtain a robust test and to correct for multiple testing, we will perform a permutation test based on the statistic in (6). Let $Y = \{Y_1, Y_2, \ldots, Y_m\}$ and let d be a binary vector indicating whether the individuals are affected or unaffected. The permutation procedure for each i using N_{sim} permutations is as follows:

1. calculate Y and save as Y^0, $s = 1$
2. permute d
3. calculate Y and save as Y^s
4. $s = s + 1$
5. if $s < N_{\text{sim}}$ go to 2

A simulated P-value for each locus can then be obtained as

$$p_{\text{sim}}^i = \frac{\sum_{s=0}^{N_{\text{sim}}} I_{Y_i^0 < Y_i^s}}{N_{\text{sim}}} \quad (7)$$

and a P-value or a significance threshold corrected for multiple testing obtained by saving the maximum statistic for each iteration $\max(Y^j)$ and the simulated P-value is then:

$$p_{\max_{\text{sim}}} = \frac{\sum_{s=0}^{N_{\text{sim}}} I_{\max(Y^0) < \max(Y^s)}}{N_{\text{sim}}}. \quad (8)$$

Missing data. When testing for linkage using a very dense set of markers we propose inferring the missing estimates of IBD from the adjacent non-missing estimates, by taking the average of the two adjacent non-missing estimates weighted by their distance. For the missing estimates located at the start or end of a chromosome, the estimate is inferred using just one adjacent region. For less dense data the above test is still valid if missing data at a locus are discarded and the number of individuals is then the number of non-missing individuals.

DATA

We applied the method to the Affymetrix 500 K chip from the Centre d'Etude du Polymorphisme Humain (CEPH) population of HapMap [The International HapMap Consortium, 2007]. The data were produced by Affymetrix and the BRLMM algorithm was used for the base calling. The data consist of 30 trios of European ancestry. The offspring were excluded from the calculation of reference LD and allele frequencies.

To show that this method can be a very powerful tool for linkage mapping we applied the method to seven Danish patients with breast and/or ovarian cancer. These individuals are heterozygous for a deletion in their *BRCA1* gene on chromosome 17 recently identified using multiplex ligation-dependent probe amplification analysis (Hansen et al., 2008). The individuals are seemingly unrelated with a co-ancestry coefficient of less than 0.02 between pairs of individuals. The individuals were genotyped using the Affymetrix sty chip for approximately 225,000 SNPs. The base calling was performed using the BRLMM algorithm using the CEPH HapMap individuals as reference sample for the base calling.

We used the CEPH individuals and the Danish breast and/or overian cancer patients as a reference population for estimating the allele frequency and pairwise LD. Later the CEPH individuals are used as controls for performing linkage analysis. For both the affected individuals and the HapMap individuals we removed SNPs with a minor allele frequency less than 0.01, SNPs that are in near perfect LD ($r^2 > 0.99$) and SNP with more than 10% missing data, because these SNPs contain little or no information. Only the autosomal chromosomes were used in the analysis.

The genotyping error was set to a conservative 0.01 based on the BRLMM white paper [Affymetrix, Inc., 2006] where the genotyping error for the HapMap individuals was estimated as 0.006 and 0.008 for homozygotes and heterozygotes base calls, respectively.

RESULTS

We first applied the method on two CEPH HapMap offspring on chromosome 9 using the rest of the CEPH

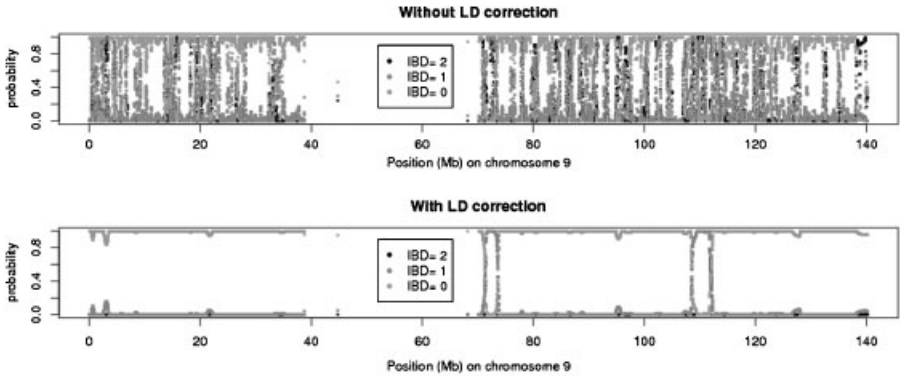

Fig. 1. Estimates of local relatedness for HapMap individuals NA10863 and NA06991 chromosome 9 using about 15,000 SNPs. Local relatedness was estimated without (top) and with (bottom) LD modified emission probabilities. A large region near the centromere (between positions 40 and 70 Mb) contains little or no SNPs. The two IBD tracts found in the bottom figure are consistent with the entire HapMap data. IBD, identity by descent; LD, linkage disequilibrium; SNPs, single nucleotide polymorphisms.

TABLE III. The co-ancestry coefficients for pairs of breast and/or ovarian cancer patients

	Methods of moment approach						
HMM approach	1	2	3	4	5	6	7
1		0.014	0.0090	0.00062	0.0061	0.0028	0.0060
2	0.014		0.021	0.0050	0.018	0.0071	0.021
3	0.010	0.017		0.0073	0.017	0.0058	0.0091
4	0.0050	0.0024	0.018		0.0061	0	0
5	0.0085	0.0072	0.017	0.0030		0.0065	0.011
6	0.012	0.022	0.013	0.0050	0.013		0.0050
7	0.0082	0.0069	0.0062	0.0072	0.0075	0.015	

Co-ancestry coefficients were estimated using a method of moments approach and using parameters from the stationary distribution of the HMM described in this article. The 0 estimates indicates that the method of moments approach estimates a negative coefficient that is mapped to zero.

HapMap parents as the reference population. Purcell et al. [2007] showed that these two offspring were related on that chromosome using all the HapMap data. They removed all but 6,000 SNPs before performing their analysis so that most of the LD were removed. Here we used the 500K Affymetrix SNP-chip data for the same individuals and removed only the SNPs with a minor allele frequency below 1% and SNPs in near perfect LD ($r^2 > 0.99$). In Figure 1 the local IBD probabilities are shown first without any correction for LD and secondly with the correction. For these data the method adequately corrects for the local LD between SNPs and identifies two IBD regions. These regions are consistent with the entire HapMap data (not shown) in the sense that the IBD regions contain no SNPs were the two individuals are homozygous for different alleles. The second region is identical to the region presented by Purcell et al. [2007]. The run time with and without LD correction was less than a minute.

We applied the method on seven Danish breast and/or ovarian cancer patients who all had the same deletion in the *BRCA1* gene. The individuals were initially reported as unrelated. We estimated the co-ancestry coefficient using the method of moments approach described in Purcell et al. [2007] and using the HMM approach described in this article. The result can be seen in Table III.

All pairwise co-ancestry coefficients for the HMM approach were less than 0.02 indicating that none of the individuals are closely related. Therefore, we assumed that each pair could only share one or no alleles IBD and estimated the α as described in Section ''Materials and Methods.'' When estimating the local relatedness for the affected individuals we saw only a couple of tracts of relatedness across the genomes. One of the tracts in the region containing the *BRCA1* gene located on chromosome 17 was frequently observed (not shown).

To show that the method can be used for linkage mapping, we combined the local relatedness for all pairs of individuals to see if we could identify the *BRCA1* region. As described in Section ''Materials and Methods,'' we used a case-control design using the seven affected individuals as cases and the CEPH HapMap individuals as controls. The method was run on all the 2,211 pairs of individuals, which took about 12 h on a single 1.7 MHz processor. The HapMap individuals are used to show that the method can actually work even for small samples but we recommend using controls that are more appropriate i.e. from the same population. Figures 2 and 3 show a large IBD peak above the *BRCA1* region on chromosome 17. The magnitude of the peak is much larger than any other peak implying that the affected individuals are much more related in the *BRCA1* region than in any other region.

Although we do not have any information about the disease status of the HapMap individuals, we will assume that they are unaffected and thus can act as controls. This enables us to estimate both a local and a global *P*-value using a permutation procedure. Figures 2 and 3 show that the highest peak is above the *BRCA1* gene and is very significant even after correction for multiple testing. It should be noticed that because of the multiple testing problem, conventional

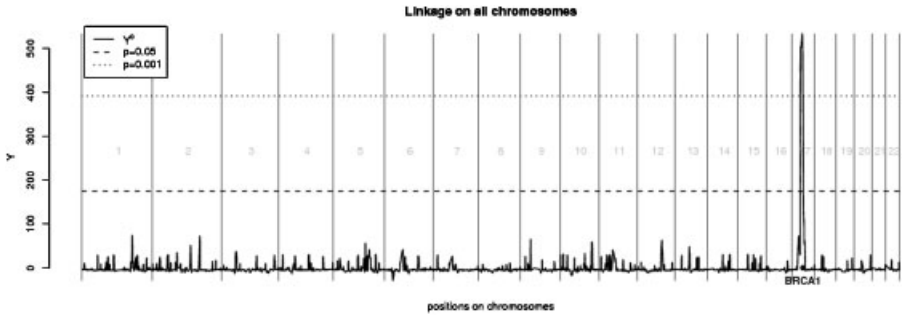

Fig. 2. Relatedness mapping for breast and/or ovarian cancer patients. The *y*-axis shows Y^0 for SNPs from all autosomal chromosomes. Each chromosome is numbered and separated by a vertical line. The dashed lines are the genome-wide *P*-value of 0.05 (large dash) and 0.001 (small dash) estimated by 10,000 permutations. The *BRCA1* gene is marked by a dot. IBD, identity by descent; LD, linkage disequilibrium; SNPs, single nucleotide polymorphisms.

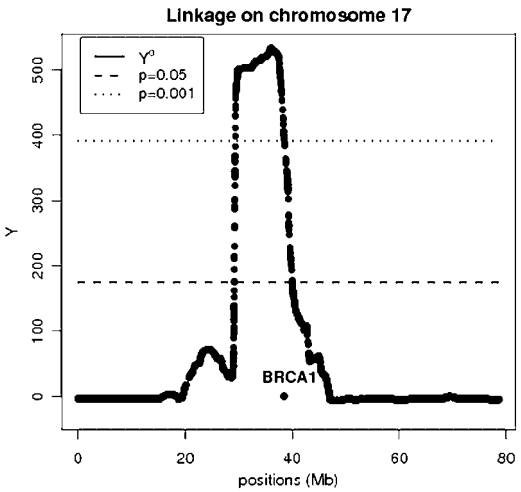

Fig. 3. Relatedness mapping for chromosome 17. The *y*-axis shows Y^0. The dashed lines are the genome-wide *P*-value of 0.05 (large dash) and 0.001 (small dash) estimated by 10,000 permutations. The *BRCA1* gene is marked by a dot.

case-control study method would have essentially zero mapping power with these sample sizes.

Many complex diseases, such as cancers, are typically multifactorial and susceptibility alleles often have incomplete penetrance. In any case-control sample, we would expect only a fraction of the affected individuals to carry any particular causative mutation, and possibly also expect some un-affected individuals to carry the mutation. Therefore, we also examined the accuracy in identifying the *BRCA1* region in the presence of affected individuals not carrying the mutation and in the presence of incomplete penetrance. In Table IV, we randomly assigned some of the 60 CEPH individuals as cases and some of the cancer patients as controls. It is clear that high power is maintained even if only a few of the affected individuals carry the mutation. For example, even with only four out of seven cases, and three out of 60 controls carrying the mutation, *BRCA1* is identified in all permutations after correcting for multiple testing.

TABLE IV. Power of the method

Controls		Affected		
w. deletion	w/o deletion	w. deletion	w/o deletion	Power[a]
0	60	7	0	1/1
0	53	7	7	1/1
0	46	7	14	1/1
2	58	5	2	21/21
2	51	5	9	21/21
2	44	5	16	11/21
3	57	4	3	35/35
3	50	4	10	22/35
3	43	4	17	8/35
4	56	3	4	20/35
4	49	3	11	0/35
4	42	3	18	0/35

[a]The fraction of times the *BRCA1* region is identified ($P < 0.05$). We assume either 7, 14, or 21 and 60, 53, or 46 controls, respectively. The number of true cases (individuals with the disease phenotype assigned to be cases) and false controls (individuals with the disease phenotype assigned to be controls) are also allowed to vary. Here all permutations of the true cases are evaluated while the true controls are randomly assigned as cases and controls. We say that the *BRCA1* region is significantly identified if the region has a genome-wide *P*-value less than 0.05 estimated using the described permutation procedure based on 1,000 permutations.

DISCUSSION

A recent method exploit the fact that if traced back far enough, all individuals are related [Purcell et al., 2007]. Using this method, it was shown that many of the HapMap individuals share many regions IBD [The International HapMap Consortium, 2007].

This method is appropriate when there is an abundance of SNPs because in that case they retain a large number of SNPs even when discarding SNPs that either have missing data or are correlated with each other.

We have here developed a method based on the same principle but with enhanced features, so no SNP data have to be discarded. This makes the method more appropriate

TABLE V. Joint probabilities for observing two genotypes at positions *i* and *h*

$G_i^{j,k}$	$G_h^{j,k}$	$X_i = 0$	$X_i = 1$	$X_i = 2$
BB BB	AA AA	p_{BA}^4	p_{BA}^3	p_{BA}^2
BB BB	AA Aa	$4p_{BA}^3 p_{Ba}$	$2p_{BA}^2 p_{Ba}$	0
BB BB	AA aa	$2p_{BA}^2 p_{Ba}^2$	0	0
BB BB	Aa Aa	$4p_{BA}^2 p_{Ba}^2$	$p_{BA}^2 p_{Ba} + p_{Ba}^2 p_{BA}$	$2p_{BA} p_{Ba}$
BB Bb	AA AA	$4p_{BA}^3 p_{bA}$	$2p_{BA}^2 p_{bA}$	0
BB Bb	AA Aa	$4p_{BA}^2 p_{bA} p_{Ba} + 4p_{BA}^3 p_{ba} + 8p_{BA}^2 p_{bA} p_{Ba}$	$2p_{BA}^2 p_{ba} + 2p_{BA} p_{Ba} p_{bA}$	0
BB Bb	AA aa	$4p_{BA}^2 p_{ba} p_{Ba} + 4p_{Ba}^2 p_{bA} p_{BA}$	0	0
BB Bb	Aa Aa	$8p_{BA}^2 p_{ba} p_{Ba} + 8p_{BA} p_{Ba}^2 p_{bA}$	$2p_{BA} p_{ba} p_{Ba} + 2p_{BA} p_{Ba} p_{bA}$	0
BB bb	AA AA	$2p_{BA}^2 p_{bA}^2$	0	0
BB bb	AA Aa	$4p_{BA} p_{Ba} p_{bA}^2 + 4p_{BA}^2 p_{bA} p_{ba}$	0	0
BB bb	AA aa	$2p_{BA}^2 p_{ba}^2 + 2p_{Ba}^2 p_{bA}^2$	0	0
BB bb	Aa Aa	$8p_{ba} p_{bA} p_{BA} p_{Ba}$	0	0
Bb Bb	AA AA	$4p_{BA}^2 p_{bA}^2$	$p_{BA} p_{bA}^2 + p_{BA}^2 p_{bA}$	$2p_{BA} p_{bA}$
Bb Bb	AA Aa	$8p_{BA}^2 p_{bA} p_{ba} + 8p_{BA} p_{bA}^2 p_{Ba}$	$2p_{BA} p_{bA} p_{ba} + 2p_{BA} p_{bA} p_{Ba}$	0
Bb Bb	AA aa	$8p_{ba} p_{bA} p_{BA} p_{Ba}$	0	0
Bb Bb	Aa Aa	$4p_{BA}^2 p_{ba}^2 + 8p_{ba} p_{bA} p_{BA} p_{Ba} + 4p_{Ba}^2 p_{bA}^2$	$p_{BA}^2 p_{ba} + p_{BA} p_{ba}^2 + p_{Ba}^2 p_{bA}$	$2p_{BA} p_{ba} + 2p_{Ba} p_{bA}$

Each SNP has two alleles denoted by A and a at positions *i* and B and b at position *h*. The probability for observing a haplotype *AB* is denoted as p_{AB} and the probability for observing allele A is denoted as p_A. The IBD states at the two positions are assumed to be the same. IBD, identity by descent; SNP, single nucleoside polymorphism.

for SNP chip data. In particular, it allows for LD between markers. Through the correction for LD is only based on pairwise correlation between markers, we have shown that it performs very well on real data. We have also accommodated the occurrence of genotyping errors in the model. For SNP-chip data genotyping errors are often a concern and if we had not included the errors in the model a single error could force the chain to change states. Although the IBD tracts are interesting enough by themselves, performing actual population-based IBD mapping is probably the method's largest potential.

In recent years the use of very high throughput SNP data has become common for association mapping. One challenge in population-based genome-wide association mapping is that very large samples are necessary to overcome the multiple testing problem. However, IBD mapping based on the current method may not suffer similar problems if the mapping population is sufficiently outbred. As shown in this study, it is possible to map a locus with just a few individuals. The reason for the remarkably high power is that shared IBD among multiple individuals in an outbred population is very unlikely to occur by random. However, if some cases share a rare mutation, the region around it will be IBD among these individuals, leading to high mapping power. IBD mapping based on this method in outbred populations is particularly useful when the causative mutations are rare—and if there are not too many different causative mutations. However, if affected individuals do not share the same causative mutation, the method will not have any mapping power. The reduced power in the presence of highly multifactorial traits is a challenge the method shares with association mapping methods.

The precision in the map estimates depends, as with other methods, on the number of meioses in the ancestry of the individuals sharing the causative mutation. If individuals are only distantly related, it is in theory possible to get very accurate estimates. However, the method can also be applied to closely related individuals, similar to the data used in traditional linkage studies. In that case, the precision in the mapping estimate would be similar to the precision obtained from the traditional linkage methods, as it is limited by the number of meioses in the pedigree. Nonetheless, the method has the potential to combine data from different families and to obtain mapping precision defined not only by the meioses occurring in the family tree, but also by the meioses separating the different families.

CONCLUSION

We have developed an improved method for inferring relatedness tracts from genome-wide SNP data. Using real data, we have shown that an IBD mapping method based on relatedness can have very high mapping power even when only a few affected individuals are included in the study. However, high mapping power relies on the assumption of a shared causative mutation among affected individuals.

WEB RESOURCES

Affymetrix HapMap individuals:
http://www.affymetrix.com/support/technical/sample_data/500k_hapmap_genotype_data.affx

Software for the pairwise relatedness estimation and linkage analysis is available at:
www.biostat.ku.dk/~ande/software

REFERENCES

Affymetrix, Inc. 2006. BRLMM: an Improved Genotype Calling Method for the GeneChip Human Mapping 500 K Array Set. http://www.affymetrix.com/support/technical/whitepapers/brlmm_whitepaper.pdf.

Boehnke M, Cox NJ. 1997. Accurate inference of relationships in sib-pair linkage studies. Am J Hum Genet 61:423–429.

Byrd RH, Lu P, Nocedal J. 1995. A limited memory algorithm for bound contstrained optimization, SIAM J. Statist. Sci. Comput. 16(5):1190–1208.

Cheung V, Nelson S. 1998. Genomic mismatch scanning identifies human genomic DNA shared identical by descent. Genomics 47:1–6.

Clayton D, Leung HT. 2007. An R package for analysis of whole-genome association studies. Hum Hered 64:45–51.

Epstein MP, Duren WL, Boehnke M. 2000. Improved inference of relationship for pairs of individuals. Am J Hum Genet 67:1219–1231.

Evett IW, Weir BS. 1998. Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists. USA: Sinauer Associates Inc.

Feingold E. 1993. Markov processes for modeling and analyzing a new genetic mapping method. J Appl Probab 30:766–779.

Hansen TV, Jønson L, Albrechtsen A, Andersen MK, Ejlertsen B, Nielsen FC. 2008. Large BRCA1 and BRCA2 genomic rearrangements in Danish high risk breast-ovarian cancer families. Breast Cancer Res Treat. In press.

Hirschhorn J, Daly M. 2005. Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6:95–108.

Jacquard A. 1974. The Genetic Structure of Populations. New York: Springer.

Kingsmore S, Lindquist I, Mudge J, Gessler D, Beavis W. 2008. Genome-wide association studies: progress and potential for drug discovery and development. Nat Rev Drug Discov 7:221–230.

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K. 2002. A high-resolution recombination map of the human genome. Nat Genet 31:241–247.

Lynch M, Ritland K. 1999. Estimation of pairwise relatedness with molecular markers. Genetics 152:1753–1766.

McPeek MS, Sun L. 2000. Statistical tests for detection of misspecified relationships by use of genome-screen data. Am J Hum Genet 66:1076–1094.

Milligan BG. 2003. Maximum-likelihood estimation of relatedness. Genetics 163:1153–1167.

Press WH, Flannery BP, Teukolsky SA, Vetterling WT. 1992. Numerical Recipes in C: The Art of Scientific Computing. Cambridge: Cambridge University Press.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575.

Queller DC, Goodnight KF. 1989. Estimating Relatedness Using Genetic Markers Evolution 43:258–275

Rabiner LR, Juang BH. 1986. An introduction to hidden Markov models. IEEE ASSP Mag 3:4–16.

Ritland K. 1996. Estimators for pairwise relatedness and inbreeding coefficients. Genet Res 67:175–186.

Service S, Lang D, Freimer N, Sandkuijl L. 1999. Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. Am J Hum Genet 64:1728–1738.

Smith M, O'Brien S. 2005. Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. Nat Rev Genet 6:623–632.

The International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851–861.

Thompson EA. 1975. The estimation of pairwise relationships. Ann Hum Genet 39:173–188.

Wang H, Lin C, Service S, Chen Y, Freimer N, Sabatti C. 2006. Linkage disequilibrium and haplotype homozygosity in population samples genotyped at a high marker density. Hum Hered 62:175–189.

APPENDIX A: TRANSITION PROBABILITIES

First we rewrite the transition probabilities $P(X_i = \mathbf{x_i}| X_{i-1} = \mathbf{x_{i-1}}) = p(X(t+s) = \mathbf{x_i}|X(s) = \mathbf{x_{i-1}})$, where s is the position of SNP $i-1$ and t is the distance from SNP $i-1$ to SNP i. Then by solving Kolmogorov's forward equations for Equation (1) with boundary conditions $P(X(0) = \mathbf{x_i}|X(0) = \mathbf{x_i}) = 1$ for any state $\mathbf{x_i}$, we get the following transition matrix:

$$T = \begin{pmatrix} 1-(1-e^{-\alpha t})\omega_1 - T_{0,2} & (1-e^{-\alpha t})\omega_1 & T_{0,2} \\ (1-e^{-\alpha t})\omega_0 & (1-e^{-\alpha t})\omega_1 + e^{-at} & (1-e^{-\alpha t})\omega_2 \\ T_{2,0} & (1-e^{-\alpha t})\omega_1 & 1-(1-e^{-\alpha t})\omega_1 - T_{2,0} \end{pmatrix} \tag{A.1}$$

where

$$T_{0,2} = \frac{e^{-\alpha\omega_1 t}\omega_2}{\omega_1 - 1} + e^{-\alpha t}\omega_1 + \frac{e^{-\alpha t}\omega_0\omega_1}{\omega_1 - 1} + \omega_2 \tag{A.2}$$

and

$$T_{2,0} = \frac{e^{-\alpha\omega_1 t}\omega_0}{\omega_1 - 1} + e^{-\alpha t}\omega_1 + \frac{e^{-\alpha t}\omega_2\omega_1}{\omega_1 - 1} + \omega_0. \tag{A.3}$$

Here $T_{i,j}$ is the transition probability for going from state i to state j.

APPENDIX B: ESTIMATING α FROM THE OVERALL RELATEDNESS

The number of meiosis, M, can be calculated from the number of meiosis from both offspring lineages from a common ancestor and the relatedness estimates [Purcell et al., 2007]

$$M = M_1 + M_2, \quad M_i = 1 - \log(z_i)/\log(2)$$

where

$$z_1 = \frac{\omega_1 + 2\omega_2 + \sqrt{(\omega_1 + 2\omega_2)^2 - 4\omega_2}}{2} \tag{B.1}$$

$$z_2 = \frac{\omega_2}{z_1}. \tag{B.2}$$

Note that if $\omega_2 = 0$ then we simply get $z_1 = \omega_1$ and m_2 is set to zero.

A central assumption in Purcell et al. [2007] is that the probability of IBD between two individuals can be calculated as $f = \frac{1}{2}^{(M-1)}$. This expression is valid when the individuals are related through exactly one specific parental pair, implying that they are related through two paths in the pedigree, each of length M.

In general, the result from Purcell holds for any pedigree in which the individuals are only related through paths with exactly k meiosis, and all ancestors of the path are outbred. Then M must be interpreted

as the $M = k^{2-n}$, where n is the number of paths of length k. When there are multiple paths of different lengths in the pedigree, the IBD probabilities cannot be expressed simply as a function of a single integer, M.

When the number of meiosis is known then we know α, because the probability of not changing states in the Markov chain is $e^{-\alpha}$ per distance unit (e.g. Mb) which is equal to $(1-\theta)^M$, where θ is the recombination rate, implying

$$\alpha = -M \log(1-\theta) \tag{B.3}$$

For the recombination rate we use 1.3 cM per 1 Mb as estimated by Kong et al. [2002].

APPENDIX C: EMISSION PROBABILITIES WITH PAIRWISE LD

The emission probabilities are estimated from the joint probabilities for observing two SNPs as shown in Table V and the probabilities for observing a single SNP as shown in Table I.