Supplemental Material

Contents

S1 The Underlying Model in Detail

We want to model the situation where we have genotype data for L di-allelic SNP loci from N individuals from the same homogeneous population, so we for each locus can infer which of the N individuals' $2N$ chromosomes are IBD. To do that we use an HMM with the genotypes as the observed data and the partitioning of chromosomes into IBD sets in all loci (the IBD configuration) as the hidden (unknown) variable which we wish to infer. Let us first assume that the haplotype phase is known, i.e. that we for each genotype know which of its constituent alleles that originates from which chromosome and thus that we know the haplotypes of all $2N$ chromosome. Let us also initially assume that the data contains no genotyping errors. In that case we can let the observed data H be a matrix with a row for each of the $2N$ chromosomes and a column for each of the L loci, where $H_{cl} \in \{0, 1\}$ is the allelic type of chromosome c at locus l and thus $H_{\cdot l}$ is the vector of allelic types of all the chromosomes at locus l. Similarly, assuming that at most k IBD sets can be present in any given locus, we can let the hidden IBD configuration, X, be a matrix with a row for each chromosome c and a column for each locus l with $X_{cl} \in \Delta = \{0, 1, 2, ..., k\}$. In this matrix the lth column, $X_{\cdot l}$, represents the partitioning of the $2N$ chromosomes into IBD sets at locus l as follows: all chromosomes with the same positive value of X_{cl} share this locus IBD, and all chromosomes with value 0 are not IBD to any of the other chromosomes.

Since IBD sharing is inherited in regions and since the lengths of these are determined by recombination events we make the assumption that the IBD states along each chromosome, i.e. each row in X, can be described as a Markov process, where distance is measured in Centi-Morgan (cM) between the different loci. For simplicity we also assume that all SNPs genotyped at different loci are independent given their IBD states, i.e. that there is no linkage disequilibrium (LD). Then, the joint probability of h and x is:

$$\begin{aligned}
&P(H{=}h, X{=}x) \qquad (1)\\
&= P(X{=}x)P(H{=}h|X{=}x)\\
&= \Big(P(X_{\cdot 1}{=}x_{\cdot 1})\prod_{l=2}^{L} P(X_{\cdot l}{=}x_{\cdot l}|X_{\cdot l-1}{=}x_{\cdot l-1})\Big)\Big(\prod_{l=1}^{L} P(H_{\cdot l}{=}h_{\cdot l}|X_{\cdot l}{=}x_{\cdot l})\Big)\\
&= \Big(P(X_{\cdot 1}{=}x_{\cdot 1})\prod_{l=2}^{L} P(X_{\cdot l}{=}x_{\cdot l}|X_{\cdot l-1}{=}x_{\cdot l-1})\Big)\Big(\prod_{l=1}^{L}\sum_{a_l \in A_l} P(A_l{=}a_l|X_{\cdot l}{=}x_{\cdot l})P(H_{\cdot l}{=}h_{\cdot l}|A_l{=}a_l, X_{\cdot l}{=}x_{\cdot l})\Big)\\
&= \Big(P(X_{\cdot 1}{=}x_{\cdot 1})\prod_{l=2}^{L} P(X_{\cdot l}{=}x_{\cdot l}|X_{\cdot l-1}{=}x_{\cdot l-1})\Big)\Big(\prod_{l=1}^{L}\sum_{A_l} P(A_l{=}a_l)P(H_{\cdot l}{=}h_{\cdot l}|A_l{=}a_l, X_{\cdot l}{=}x_{\cdot l})\Big)
\end{aligned}$$

$A_l \in \{0,1\}^k$ is here a vector of the allelic states in the k possible IBD sets and the last equality is true because A_l is independent of $X_{\cdot l}$. When calculating the emission probability of locus l, $P(H_{\cdot l}{=}h_{\cdot l}|X_{\cdot l}{=}x_{\cdot l})$, we sum over all possible values of A_l to take into account that the allelic states of the different IBD sets are not known. Each of the probabilities $P(H_{\cdot l}{=}h_{\cdot l}|A_l{=}a_l, X_{\cdot l}{=}x_{\cdot l})$ is simply calculated as the product $\prod_{c=1}^{2N} P(H_{cl}{=}h_{cl}|A_l{=}a_l, X_{cl}{=}x_{cl})$ with

$$P(H_{cl}{=}h_{cl}|A_l{=}a_l, X_{cl}{=}x_{cl}) = \begin{cases} f_{l,h_{cl}} & \text{if } x_{cl}{=}0 \\ 1, & \text{if } x_{cl} > 0 \text{ and } a_l\left[x_{cl}\right] = h_{cl} \\ 0, & \text{if } x_{cl} > 0 \text{ and } a_l\left[x_{cl}\right] \neq h_{cl} \end{cases} \tag{2}$$

where $f_{l,h_{cl}}$ is the nucleotide frequency of nucleotide h_{cl} at position l estimated from an appropriate reference population and $a_l\left[x_{cl}\right]$ is the x_{cl}th element of vector a_l and thus the allelic state in IBD set x_{cl} at position l. Similarly, $P(A_l{=}a_l)$ is simply calculated as the product of the independent frequencies of each of its constituent allelic states.

In effect, if several chromosomes are in the same IBD set, v, at locus l then the haplotypes of these chromosomes will contribute to the emission probability in l with exactly one factor in total, $f_{l,a_l[v]}$, if all the corresponding haplotypes are equal to the allelic state of IBD set v and a factor of 0 otherwise. Also, chromosomes with IBD state 0 or with IBD states that they do not share with any other chromosome contribute a factor of $f_{l,h_{cl}}$ each. For instance, if $N = 1$ and the individual's two chromosomes are in the same IBD set, the emission probability will be $f_{l,B}$ or $f_{l,b}$ if the individual is homozygous for B or b, respectively, and 0 if the individual is heterozygous. Similarly, if the two chromosomes are not in the same IBD set, the emission probability will be $f_{l,B}^2$ or $f_{l,b}^2$ for homozygous individuals and $f_{l,B}f_{l,b}$ for heterozygous individuals.

The transition probabilities of the HMM capture information regarding the length distribution of IBD regions. Because the distances between the observed nucleotide loci are large, a *continuous* time Markov model provides a tractable approach to modeling these transition probabilities. We assume that each Markov process along a chromosome jumps from a state different than 0 into state 0 at constant rate, $\rho > 0$, and from state 0 into a state different than 0 at a constant rate, $\lambda > 0$. Additionally, to keep the number of parameters at a minimum, we assume that the instantaneous rate of transition made directly between different non-zero IBD sets is 0. Under these assumptions the Markov processes are time reversible, with detailed balance equations

$$\frac{\pi_0 \lambda}{k} = \rho \pi_v \qquad \forall v \in \Delta, \quad v \neq 0$$

where π_v is the stationary probability of IBD state v. Solving Kolmogorov's Forward Equations we find the transition probabilities along each chromosome sequence as

$$P(X_{cl}{=}x_{cl} | X_{c(l-1)}{=}x_{c(l-1)}) = \begin{cases} \frac{\rho+\lambda e^{-t(\rho+\lambda)}}{\rho+\lambda} & \text{if } x_{c(l-1)} = 0 \text{ and } \ x_{cl} = 0 \\ \frac{\lambda-\lambda e^{-t(\rho+\lambda)}}{k(\rho+\lambda)} & \text{if } x_{c(l-1)} = 0 \text{ and } \ x_{cl} > 0 \\ \frac{\rho-\rho e^{-t(\rho+\lambda)}}{\rho+\lambda} & \text{if } x_{c(l-1)} > 0 \text{ and } \ x_{cl} = 0 \\ \frac{e^{-t\rho}((-1+k)\lambda+(e^{-t\lambda}-1+k)\rho)+\lambda}{k(\rho+\lambda)} & \text{if } x_{c(l-1)} > 0 \text{ and } \ x_{cl} = x_{c(l-1)} \\ \frac{e^{-t\rho}((-1)\lambda+(e^{-t\lambda}-1)\rho)+\lambda}{k(\rho+\lambda)} & \text{if } x_{c(l-1)} > 0 \text{ and } \ x_{c(l-1)} \neq x_{cl} > 0 \end{cases}$$

where t is the genetic distance (or time) between SNP locus $l-1$ and SNP locus l. Thus the stationary distribution of each chain is given by:

$$\pi_v = \begin{cases} \frac{\rho}{\rho+\lambda}, & \text{if } v = 0 \\ \frac{\lambda}{k(\rho+\lambda)}, & \text{if } v \neq 0 \end{cases}$$

The full transition probability, $P(X_{\cdot l}{=}x_{\cdot l} | X_{\cdot l-1}{=}x_{\cdot l-1})$ for a given locus l is simply calculated as the product of the transition probabilities of l for the individual chains.

It is important to note that the simplifying assumption that the instantaneous rate of transition between different non-zero IBD states is zero, only means that such transitions are not allowed to happen over infinitesimal distances along the chromosome. Since the distance between any two loci is bigger than infinitesimal, the model does indeed, despite the assumption, allow such transitions to happen between any two loci, as is evident from the fact that the transition probabilities indicated above for such transitions is non-zero when $t > 0$.

It is it also worth noting that the described model reduces to the inbreeding model by Leutenegger et al. (Leutenegger et al. 2003) when $N = 1$, $k = 1$ with a set to $\lambda + \rho$ and f set to $\frac{\lambda}{\lambda+\rho}$. This is noteworthy both because it illustrates that the new model can be viewed as an extension of the model by Leutenegger et al. and because the re-parameterization to a and f contributes some additional intuition about the parameters λ and ρ. As in the model by Leutenegger et al. a $(=\lambda + \rho)$ determines the overall instantaneous rate of change between IBD and non-IBD states and f $(= \frac{\lambda}{\lambda+\rho})$ is the stationary probability of a non-zero IBD state.

Finally it should be noted that all of the above equations are for fixed values of ρ and λ. The dependence of $P(G, X)$ and $P(X_{\cdot l}|X_{\cdot l-1})$ on ρ and λ has, for the sake of simplicity, been suppressed in the notation. The values for ρ and λ are not known in advance and will be treated as parameters to be estimated from data.

Incorporating Genotyping Errors and Unknown Phase

Most often the assumptions that the phase is known and that there are no genotyping errors will be unrealistic. We deal with this by making two adjustments to the model described above.

First, when the phase is unknown the allelic types of each chromosome within each individual are no longer observed. Only the genotypes of the N individuals are observed. The observed data is thus a matrix G with a column for each locus, and a single row for each individual, with the ith row in G being equal to the sum of the two rows in H that correspond to the 2 chromosomes of individual i. In this case equation 1 will still hold (with H replaced by G). But the emission probability at locus l can no longer be calculated as a product over all chromosomes. Instead we multiply over all individuals.

Second, in order to take into account the unknown phase and genotyping errors, the factor we multiply with for each individual, to get the emission probability for locus l, is calculated using an approach very similar to the approach used by Albrechtsen et al. (Albrechtsen et al. 2009): let G_{il} be the observed unphased genotype of individual i at locus l and H^1_{il}, H^2_{il}, X^1_{il} and X^2_{il} be the underlying true haplotypes and IBD states for each of individual i's two chromosomes $j \in \{1, 2\}$ at locus l. Then assuming genotyping errors occur at a rate ϵ per allele, the factor we multiply with for individual i at locus l is calculated as a sum over all four possible true haplotype pairs:

$$P(G_{il}{=}g_{il}|A_l{=}a_l, X^1_{il}{=}x^1_{il}, X^2_{il}{=}x^2_{il}, \epsilon) = \sum_{h^1_{il}, h^2_{il}} \Big(\Big(\prod_{j\in\{1,2\}} P(H^j_{il}{=}h^j_{il}|A_l{=}a_l, X^j_{il}{=}x^j_{il}) \Big) P(G_{il}{=}g_{il}|H^1_{il}{=}h^1_{il}, H^2_{il}{=}h^2_{il}, \epsilon) \Big)$$

where $P(H^j_{il}{=}h^j_{il}|A_l{=}a_l, X^j_{il}{=}x^j_{il})$ for individual i's two chromosomes $j \in \{1, 2\}$ are calculated using equation 2 and where $P(G_{il}{=}g_{il}|H^1_{il}{=}h^1_{il}, H^2_{il}{=}h^2_{il}, \epsilon)$ ispo the probability of the observed genotype g_{il} given that the true haplotypes are h^1_{il} and h^2_{il}. The values of $P(G_{il}|H^1_{il}, H^2_{il}, \epsilon)$ are summarized in table S1.

It should be noted that the unknown phase could also have been handled by inferring the most

G_{il} \ (H_{il}^1, H_{il}^2)	(0,0)	(0,1)	(1,0)	(1,1)
0	$(1-\epsilon)^2$	$\epsilon(1-\epsilon)$	$\epsilon(1-\epsilon)$	ϵ^2
1	$2\epsilon(1-\epsilon)$	$\epsilon^2+(1-\epsilon)^2$	$\epsilon^2+(1-\epsilon)^2$	$2\epsilon(1-\epsilon)$
2	ϵ^2	$\epsilon(1-\epsilon)$	$\epsilon(1-\epsilon)$	$(1-\epsilon)^2$

Table S1. The distribution of $P(G_{il}|H_{il}^1, H_{il}^2, \epsilon)$. ϵ is the genotyping error frequency on allele level.

probable phase using a program such as fastPHASE (Scheet and Stephens 2006) and assuming that this is the true phase. However, we have chosen the above solution instead, because it allows us to take the uncertainty of the phase into account.

Implementation Details

In the implementation of the above model we do not always sum over all possible values of the full k-vector, A_l, as described. k is an upper limit to the number of IBD sets in any one locus, not necessarily the actual number of IBD sets in a given locus and when only a subset of the possible k IBD sets contain at least one chromosome, we only sum over all possible values of the A_l vector that is associated to this subset of IBD sets. This approach is mathematically equivalent to a full enumeration of all values of A_l and it reduces the number of terms in the sum by a factor of two for each ”unused set”. Hence, this modification leads to much faster computation of the emission probabilities in certain situations.

S2 Inference Using MCMC

If we let $\Phi = (X, \rho, \lambda)$ we are in general interested in the posterior distribution $P(\Phi|G = g)$ for a given unphased SNP genotype data set g defined as above (including marginal posterior probabilities) and in expectations under this distribution assuming appropriate priors for λ and ρ. For instance, for the *BRCA1* data set, consisting of SNP data from five cancer patients, later analyzed, we are interested in the posterior probability that all five individuals share the *BRCA1* gene IBD. We are also interested in the posterior expectation of the number of patients that share at least one chromosome IBD in the *BRCA1* gene.

The posterior distribution is given by

$$P(\Phi|G = g) = \frac{P(G = g, \Phi)}{P(G = g)} = \frac{P(G = g, \Phi)}{\int_\Phi P(G = g, \Phi) d\Phi} \tag{3}$$

Hence the posterior expectation of any function $f(\Phi)$ is given by

$$E[f(\Phi)|G=g] = \int_{\Phi} \Big(\frac{f(\Phi)P(G=g,\Phi)}{\int_{\Phi} P(G=g,\Phi)d\Phi} \Big) d\Phi$$

Unfortunately both are computationally intractable to calculate for practically any realistic scenario.

We solve this problem by using MCMC to approximate the posterior probabilities and expectations (for a comprehensive introduction to MCMC, see Gilks et al. 1996). More specifically, we use the Metropolis Hastings algorithm. This algorithm allows us to approximate the posterior by sampling values of Φ in a manner that only requires calculations of $P(G=g, X=x|\rho,\lambda)$. After choosing an initial value for Φ, this algorithm proceeds by repeatedly proposing a new value of either X, ρ and λ and thus a new value ϕ_{prop} of Φ based only on the current value, ϕ_{cur}, of Φ. The proposed value is accepted with probability

$$\alpha = \min\left(1, \frac{P(G=g, X_{prop}=x_{prop}|\rho_{prop},\lambda_{prop})P(\rho_{prop})P(\lambda_{prop})q(\phi_{cur}|\phi_{prop})}{P(G=g, X_{cur}=x_{cur}|\rho_{cur},\lambda_{cur})P(\rho_{cur})P(\lambda_{cur})q(\phi_{prop}|\phi_{cur})}\right)$$

where $q(\phi_2|\phi_1)$ is the probability of proposing ϕ_2 when the current value of Φ is ϕ_1. The proposal distributions, q, used, are described in Supplemental Material, section S3. Note that we use uniform priors for ρ and λ, and that these priors will therefore cancel when calculating the acceptance probabilities.

The described algorithm provides samples from an ergodic Markov chain that has $P(\Phi|G)$ as its stationary distribution (Gilks et al. 1996). And for any set of values, ϕ, of Φ the posterior probability of ϕ can simply be approximated by

$$P(\phi|G=g) \simeq \frac{1}{n-b} \sum_{m=b+1}^{n} I(\phi_m \in \phi)$$

where ϕ_1,ϕ_2,...,ϕ_n are the samples generated using the MCMC-algorithm and b is a burn-in chosen to be so large that the samples are approximately independent of the starting value. For any function $f(\Phi)$, similarly $E[f(\Phi)|G=g]$ can be approximated as the ergodic average

$$E[f(\Phi)|G=g] \simeq \frac{1}{n-b} \sum_{m=b+1}^{n} f(\phi_m)$$

S3 The Moves Used in the MCMC Algorithm

The MCMC algorithm is based on a number of moves, here defined as proposal algorithms with corresponding acceptance probabilities. The moves are used to propose and evaluate changes to three different state variables: 1) the two parameters λ (the rate of change from an IBD state>0 to state 0) and ρ (the rate of change from IBD state>0 to another state) and 2) the IBD configuration matrix X.

In total there is one move for proposing changes to each of the parameters λ and ρ and six moves for proposing changes to the IBD configuration matrix X.

Notation

In the descriptions of the moves the following notation is used:

N number of individuals in total

L number of loci in total

site a specific locus on a specific chromosome.

region a number of consecutive sites on one or more chromosomes. If more chromosomes are involved the region starts and ends in the same locus on all chromosomes. In a region there can be many different IBD states.

k the maximum number of possible IBD sets in any given locus.

IBD state the number from 0 to k that a site has been assigned in the IBD configuration matrix (i.e. the value in $X[chr, locus]$)

IBD block one or more consecutive sites that all have the same IBD state.

IBD block border where two IBD blocks meet on a chromosome.

The Parameter Moves

There is one move for each of the two parameters λ and ρ. This move is based on a proposal algorithm, q, that proposes a new value uniformly in a window around the current value (if a value smaller than 0 or larger than 1 is sampled 0 or 1 is used as a mirror – if for instance the sampled value is -0.01 the proposed value will be 0.01.) Since q is a symmetric proposal distribution, i.e. for any two values p_1

and p_2 of the parameter of interest $q(p_1|p_2) = q(p_2|p_1)$, this is simply a Metropolis move. So assuming uniform priors on λ and ρ the following acceptance probability is used for λ_{prop}:

$$\alpha = \min\left(1, \frac{P(G = g, X = x_{cur}|\rho_{cur}, \lambda_{prop})}{P(G = g, X = x_{cur}|\rho_{cur}, \lambda_{cur})}\right)$$

and the following is used for ρ_{prop}:

$$\alpha = \min\left(1, \frac{P(G = g, X = x_{cur}|\rho_{prop}, \lambda_{cur})}{P(G = g, X = x_{cur}|\rho_{cur}, \lambda_{cur})}\right)$$

The X Matrix Moves

The Single Site X Move

In this move a site is picked at random and the IBD state of this site is changed to a randomly picked other state from the set $\{0, 1, ..., k\}$. The proposal algorithm behind the move more precisely consists of the following steps:

1. Pick a chromosome c uniformly

2. Pick a locus l uniformly

3. Pick a new IBD state for the chosen site uniformly

Again, since we have a move based on a symmetric proposal distribution, we have a Metropolis move. So assuming uniform priors we have the following acceptance probability

$$\alpha = \min\left(1, \frac{P(G = g, X = x_{prop}|\rho_{cur}, \lambda_{cur})}{P(G = g, X = x_{cur}|\rho_{cur}, \lambda_{cur})}\right)$$

The Region X Move

This move consists of picking a region on a random chromosome and changing the IBD state of all sites in this region by adding the same random number to all of them. More precisely the proposal algorithm is as follows

1. Pick a chromosome c uniformly

2. Pick a region length m uniformly from 1 to a fixed maximum region length

3. Pick a region start locus l uniformly from 1 to L-m

4. Pick an integer $z \in \{1, ..., k\}$ and set the state of each site in the region to be (the current state+z) modulo $k+1$

Again, the proposal distribution is symmetrical and hence the move is a Metropolis move. So assuming uniform priors we again get the acceptance probability

$$\alpha = \min\left(1, \frac{P(G=g, X=x_{prop}|\rho_{cur}, \lambda_{cur})}{P(G=g, X=x_{cur}|\rho_{cur}, \lambda_{cur})}\right)$$

The Border X Move

The idea behind this move is to make it possible to move IBD block borders. The proposal algorithm consists of the following steps:

1. Pick a chromosome c uniformly

2. Pick a border on c uniformly. If there are no borders on c reject move.

3. Pick a side to move the border to (uniformly)

4. Pick a distance to move the border (uniformly from 1 to a fixed maximum move length)

5. Move the border in the chosen direction the chosen distance.

Note that the distance is measured in number of sites *without* borders between. And note that if the end of the X matrix is reached the rest of the distance is travelled back towards the original position of the border. In this way the proposal distribution becomes symmetrical. Thus again we have a Metropolis move. So assuming uniform priors we again get an acceptance probability of

$$\alpha = \min\left(1, \frac{P(G=g, X=x_{prop}|\rho_{cur}, \lambda_{cur})}{P(G=g, X=x_{cur}|\rho_{cur}, \lambda_{cur})}\right)$$

The Block State Change X Move

The idea behind this move is to make it possible to change IBD state on an entire IBD block on chromosome level. The proposal algorithm consists of the following steps:

1. Pick a chromosome

2. Pick an IBD block on the chromosome b uniformly and call the length of it m
3. If the IBD block is longer than a fixed maximum block length, reject the move
4. Otherwise pick a new state for the block (different from any neighboring blocks on the same chromosome and give the IBD block that state.

Since we again have a symmetric proposal distribution it must hold that (assuming uniform priors) the acceptance probability is

$$\alpha = \min\left(1, \frac{P(G = g, X = x_{prop}|\rho_{cur}, \lambda_{cur})}{P(G = g, X = x_{cur}|\rho_{cur}, \lambda_{cur})}\right)$$

The Copy X Move

This move consists of two opposite moves. The idea behind it is to make it possible to copy the IBD states of a region on one chromosome to the same region one or more other chromosomes. The proposal algorithm is as follows:

1. Pick a region length m (uniformly from 1 to a fixed maximum region length)
2. Pick a locus l (uniformly from 1 to L-m), where the region starts
3. Pick a random subset, C_n, of the $2N$ chromosomes with $|C_n| = n < 2N$.
4. Pick a move type (A or B)
    - If move type A is chosen:
        (a) Pick a chromosome c (uniformly from the $2N - n$ chromosomes not in C_n)
        (b) Copy the IBD states of the relevant region in c to all the chromosomes in C_n
    - If move type B is chosen:
        (a) If the IBD state sequence of all chromosomes in C_n are not identical to each other and at least one other chromosome in the relevant region reject the move.
        (b) Else give each site in the region in all the chromosomes in C_n a random IBD state between 0 and k.

Assuming uniform priors move A has the acceptance probability

$$\begin{aligned}
\alpha &= \min\left(1, \frac{P(G = g, X = x_{prop}|\rho_{cur}, \lambda_{cur})q(X_{cur}|X_{prop})}{P(G = g, X = x_{cur}|\rho_{cur}, \lambda_{cur})q(X_{prop}|X_{cur})}\right) \\
&= \min\left(1, \frac{P(G = g, X = x_{prop}|\rho_{cur}, \lambda_{cur})\frac{1}{(k+1)^{mn}}}{P(G = g, X = x_{cur}|\rho_{cur}, \lambda_{cur})\frac{y}{2N-n}}\right) \\
&= \min\left(1, \frac{P(G = g, X = x_{prop}|\rho_{cur}, \lambda_{cur})(2N-n)}{P(G = g, X = x_{cur}|\rho_{cur}, \lambda_{cur})\mathrm{(k+1)}^{mn}\mathrm{y}}\right)
\end{aligned}$$

where y is the number of chromosomes not in C_n that are identical to c in the region (including c itself). And move B has the acceptance probability

$$\begin{aligned}
\alpha &= \min\left(1, \frac{P(G = g, X = x_{prop}|\rho_{cur}, \lambda_{cur})q(X_{cur}|X_{prop})}{P(G = g, X = x_{cur}|\rho_{cur}, \lambda_{cur})q(X_{prop}|X_{cur})}\right) \\
&= \min\left(1, \frac{P(G = g, X = x_{prop}|\rho_{cur}, \lambda_{cur}))\mathrm{(k+1)}^{mn}\mathrm{y}}{P(G = g, X = x_{cur}|\rho_{cur}, \lambda_{cur})(2N-n)}\right)
\end{aligned}$$

The IBD Expansion/Reduction X Move

Also this move consists of two opposite moves. In the first move an IBD block is expanded to include one more chromosome. In the second move an IBD block is reduced to include one chromosome less. More precisely the proposal algorithm consists of the following steps:

1. Pick a locus l (uniformly from 1 to L)

2. Pick a move type (A or B)

    - If move type A is chosen:

        (a) Pick an IBD state s present in the lth column of the IBD matrix X and let C_n be the set of chromosomes with state s in this column and let $n = |C_n|$

        (b) Find left and right limits of the biggest IBD block that contains all of the chromosomes in C_n and locus l.

        (c) Pick a chromosome c not among the n chromosomes in C_n

        (d) If c does not have the same state in all sites between the left and right limits found in (b) the move is rejected. Otherwise c is given state s in all sites between left and right limits

- If move type B is chosen:

  (a) Pick a chromosome c uniformly

  (b) Find all $n-1$ chromosomes with same IBD state as c at locus l and let C_n be a set consisting of these chromosomes and c.

  (c) If there are no such chromosomes (i.e. if $C_n = \{c\}$) the move is rejected. Otherwise find the left and right limits of the biggest IBD block containing all the chromosomes in C_n and locus l. If the removal of c from the set would lead to another left or right limit the move is rejected. If not a new IBD state is picked uniformly and given to all sites of c in the region between the left and right limits.

Assuming uniform prior get that for move A the acceptance probability is

$$\begin{aligned}\alpha &= \min\left(1, \frac{P(G=g, X=x_{prop}|\rho_{cur}, \lambda_{cur})q(X_{cur}|X_{prop})}{P(G=g, X=x_{cur}|\rho_{cur}, \lambda_{cur})q(X_{prop}|X_{cur})}\right) \\ &= \min\left(1, \frac{P(G=g, X=x_{prop}|\rho_{cur}, \lambda_{cur})\frac{1}{2N}\frac{1}{k}}{P(G=g, X=x_{cur}|\rho_{cur}, \lambda_{cur})\frac{n}{2N}\frac{1}{2N-n}}\right) \\ &= \min\left(1, \frac{P(G=g, X=x_{prop}|\rho_{cur}, \lambda_{cur})(2N-n)}{P(G=g, X=x_{cur}|\rho_{cur}, \lambda_{cur})kn}\right)\end{aligned}$$

And for move B we get that the acceptance probability is

$$\begin{aligned}\alpha &= \min\left(1, \frac{P(G=g, X=x_{prop}|\rho_{cur}, \lambda_{cur})q(X_{cur}|X_{prop})}{P(G=g, X=x_{cur}|\rho_{cur}, \lambda_{cur})q(X_{prop}|X_{cur})}\right) \\ &= \min\left(1, \frac{P(G=g, X=x_{prop}|\rho_{cur}, \lambda_{cur})kn}{P(G=g, X=x_{cur}|\rho_{cur}, \lambda_{cur})(2N-n)}\right)\end{aligned}$$

Note that sometimes an expansion move has the exact same effect as a reduction move and vice versa. This is not corrected for in the proposal probabilities because whenever this is the case there is an equivalent opposite move that takes you back.

S4 Program Settings Used

Settings used in the example run

We used default settings for all programs, except that for Relate we turned off the LD correction since the data set does not contain LD. For the MCMC we set $k = 2$.

Settings used in the power study

MCMC We ran the MCMC method with $\epsilon = 0.005$ and $k = 1$. The ϵ value used is somewhat conservative given that the data were simulated using an error rate of approximately 0.003 to mimic realistic SNP chip data (see section about Simulated Test Data for details about the simulated data).

Relate We ran Relate under the assumption that pairs of individuals cannot share two chromosomes IBD. We ran it with default parameters except that we turned off the LD correction since the simulated data does not contain LD.

PLINK The method by Purcell et al. is based on an HMM, which infers the probability that the two individuals being analyzed share 0, 1 or 2 chromosomes IBD (Purcell et al. 2007). In the article describing the method, a region is inferred to be IBD based when the probability of sharing at least one chromosome IBD is above a certain threshold (Purcell et al. 2007). However, in the implementation available in PLINK, inference are based not only on these probabilities, but also on thresholds for the lengths of the regions measured in both kilobases and in number of SNPs. The default settings require IBD regions to be longer than 1 Mb and 100 SNPs. We are here considering shorter regions of lengths < 1 Mb, which contain significantly fewer SNPs than 100. When applying PLINK with default setting on all of the power study's 1500 positive data sets with short (< 1 Mb) IBD regions only 5 of these regions were inferred, corresponding to 0.3%. In order to give PLINK the best possible chance to find the short regions, we, therefore, chose not to use any length restrictions, but instead to use the method as it is described in Purcell et al. (Purcell et al. 2007). To be able to do so we had to force the program to output IBD probabilities for all pairs analyzed. Except for this, PLINK was run with default settings.

BEAGLE We ran BEAGLE with default parameters with the only exception that as recommended in the manual for small data sets, we used 20 iterations instead of 4, in the phasing part of the inference, in order to achieve better accuracy. As recommended in the manual we ran BEAGLE 10 times on each data set with different seed values and used the maximum IBD probability for each locus.

GERMLINE For each pair of sequences, GERMLINE partitions the genome into segments of a given length and identifies all segments where the two individuals are identical by state (IBS) in all loci. These IBS segment are then expanded to both sides until a segment with more than x mismatches is met. Finally, all segments that after expansion are longer than a fixed length are recorded as IBD regions. The following three parameters have to be supplied by the user: the minimum size of an IBD region (measured either in Mb or cM), the number of SNPs in each initial segment and an upper bound for the number of mismatches allowed in each segment for them to be interpreted as genotyping errors. The default settings are 5 cM, 128 SNPs and 2 errors. However, in most of the scenarios the IBD regions are shorter than 5 cM. Moreover, they almost all contain fewer SNPs than 128. In fact, we are interested in capturing regions with less than 30 SNPs. We therefore did not use the default settings, since using these would have lead to a TP rate of 0 for all the scenarios with short IBD regions. To give GERMLINE the best chances of finding the short regions we instead used an initial segment size of only 1 and allowed for 0 errors. In this way an IBD region should always be found unless it is shorter than the minimum length or it contains genotyping errors.

Settings used for real data

MCMC We ran the MCMC method with the exact same settings as in the power study with two exceptions. First, we used a genotyping error parameter ,ϵ of 0.01 instead of 0.005. This is a conservative estimate based on the BRLMM white paper (Inc 2006) , where the genotyping error rate for the HapMap individuals was estimated as 0.006 and 0.008 for homozygote and heterozygote genotype calls, respectively. Second, when applied to the *BRCA1* region alone we raised k to 3 instead of 1 in order to allow the MCMC method to detect if there are more than one IBD sharing set in the *BRCA1* region.

Relate We ran Relate with the exact same settings as in the power study with at single exception; we turned on the LD correction. And as recommended in the manual for data sets the size we have, we used a back parameter of 25 for the correction, meaning that for each SNP locus it corrects for LD with the previous 25 SNP positions.

PLINK We used the exact same settings as in the power study.

BEAGLE We used the exact same settings as in the power study.

GERMLINE We used default values, since it unclear which minimum length for IBD regions to use, if we were to use the same settings as in the power study in order to ensure a relatively low rate of false positives.

S5 Verification of the Method

To ensure that the MCMC algorithm works correctly, i.e. that it converges to a stationary distribution and that this stationary distribution is in fact the correct posterior distribution, we applied the method to a very small simulated data set, consisting of only two individuals and three SNPs. For a data set of this size it is possible (for any given values of λ and ρ) to calculate exact posterior probabilities. Hence running the algorithm on such a data set enabled us both to test if the method did converge to a stationary distribution *and* if this distribution was the right distribution. We focused on two possible IBD patterns: We investigated the posterior probability that individual 1 is inbred at the first locus and the posterior probability that all four chromosomes are unrelated at the third locus. We ran two chains each with a unique starting value and assessed convergence using the Gelman-Rubin convergence diagnostic: "the potential scale reduction factor". As can be seen in figure S1 the potential scale reduction factor stabilizes at a value close to 1 after a few thousand iterations. This suggests that the chains did converge to a stationary distribution.

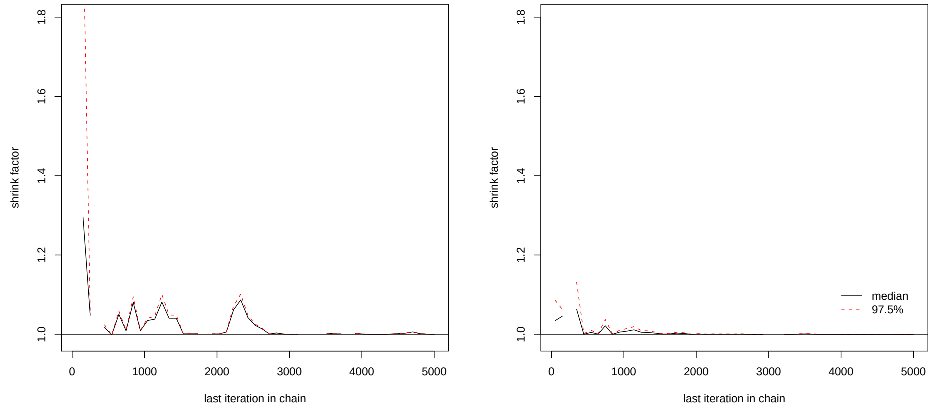

Figure S1. Gelman-Rubin plots for the two IBD patterns. Left: The Gelman-Rubin plot for IBD pattern 1. Right: The Gelman-Rubin plot for IBD pattern 2.

	Real posterior	1st run	2nd run
Individual 1 is inbred at locus 1	0.012	0.012	0.011
All four chr unrelated at locus 3	0.052	0.055	0.053

Table S2. The real posterior probabilities and the estimates obtained from the two runs of MCMC.

We then calculated the posterior probability of the two IBD patterns analytically by exact enumeration and compared these to the values obtained using MCMC. The obtained value are very close to the correct ones, see table S2, suggesting that the algorithm indeed did converge to the right distribution.

S6 Convergence Assessment of all MCMC Runs

We used the following approach for monitoring convergence for all runs described in this paper: two chains were run, each with a unique starting value and afterwards the Gelman-Rubin diagnostic was used to assess convergence for all parameters of interest. The only exception were the runs made for the power study, for which we, for computational reasons, only ran two chains for a couple of the data sets from each scenario, as the results should not be affected much by a small fraction of non-converging chains. All the Gelman-Rubin plots can be seen below (figures S2 to S6). Inspection of these plots and the fact that the final point estimates of the potential scale reduction factor in all cases were at most 1.01, suggests that convergence has indeed been achieved in all runs.

Convergence plot for the example run

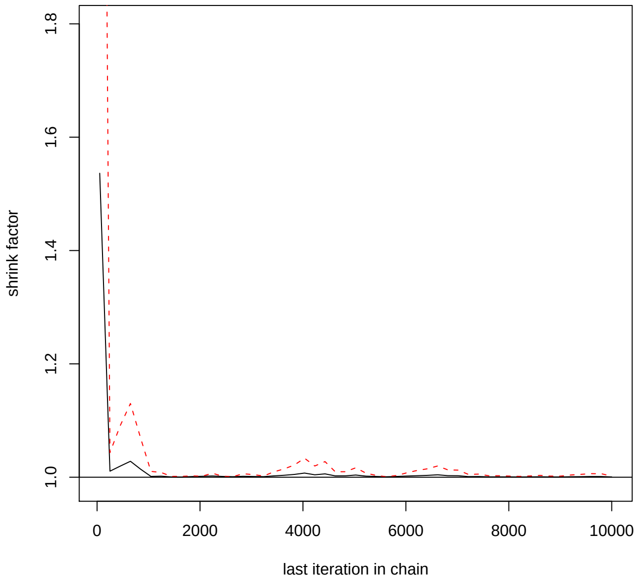

Figure S2. Gelman-Rubin plots on the number of zeros in the X matrix. This statistic was used since the MAP estimates that the inference is based is only a summary statistic in the sense that it summarizes all samples in one overall statistic and not a statistic for each sample. Hence it cannot be used for convergence monitoring. So instead we used the number of zeros in X as this at least to some extend captures the state of the entire X matrix.

Convergence plots for the power study

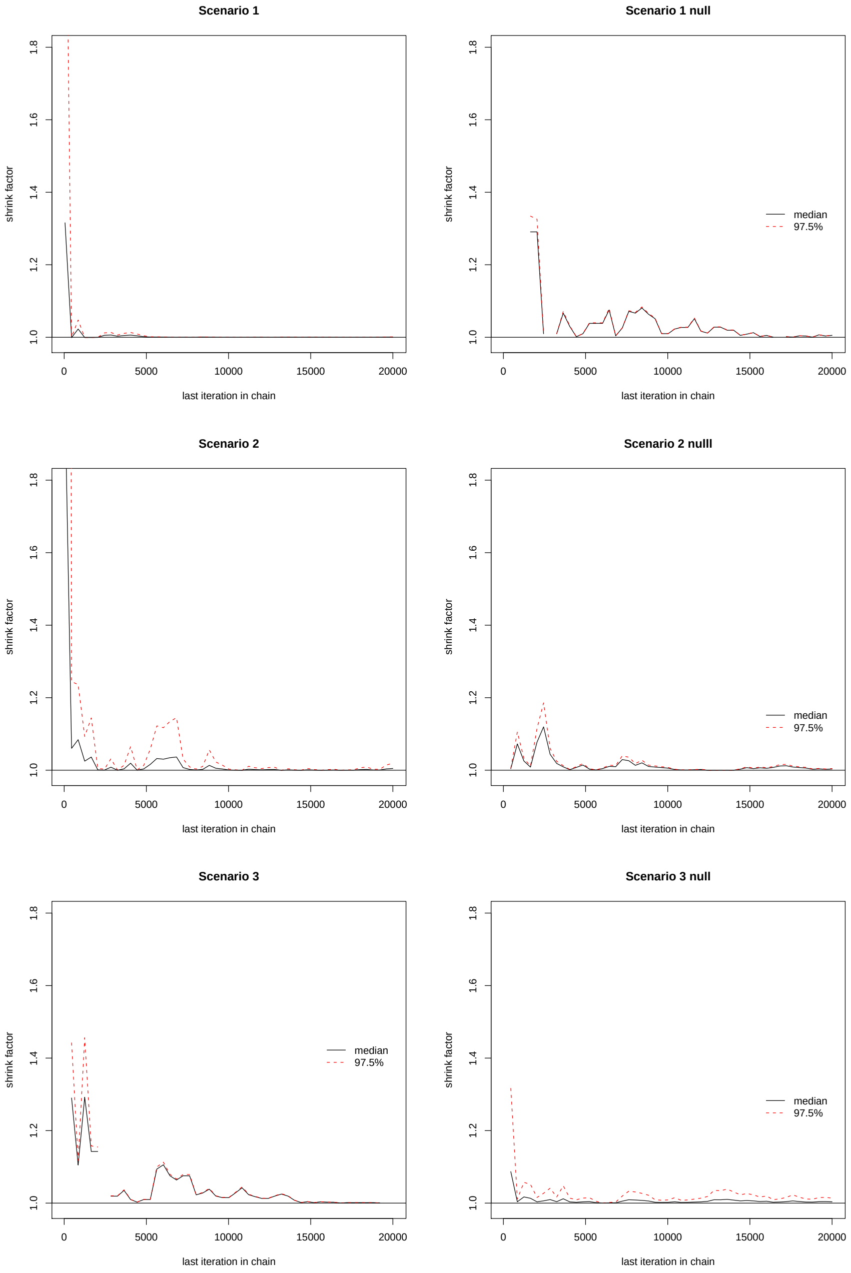

Figure S3. Gelman-Rubin plots on the statistics used to for the power study (the first 3 scenarios and their corresponding null scenarios). For a single sample from each scenario the IBD status used for the power study in a single locus (101 if part of IBD region otherwise locus 51).

Convergence plots for the power study (continued)

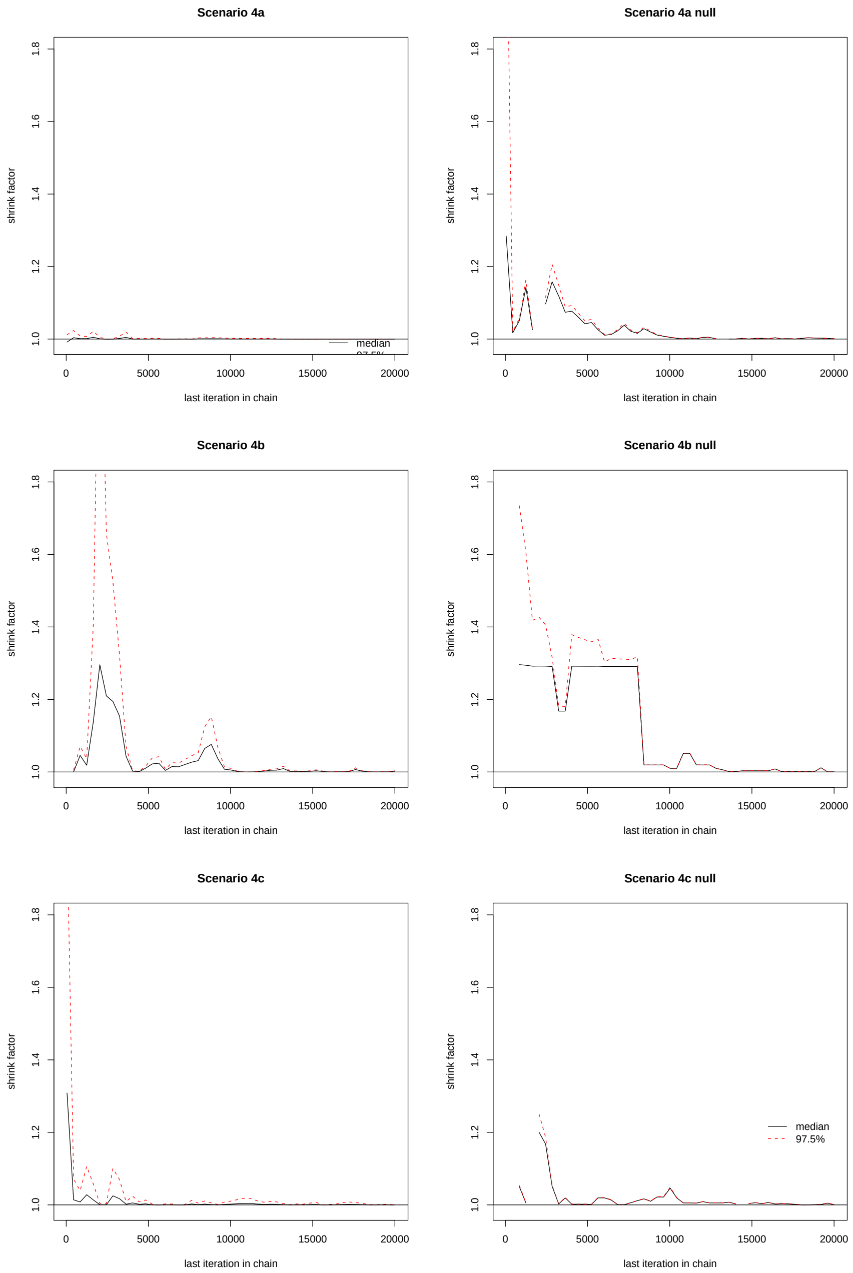

Figure S4. Gelman-Rubin plots on the statistics used to for the power study (the last 3 scenarios and their corresponding null scenarios). For a single sample from each scenario the IBD status used for the power study in a single locus (101 if part of IBD region otherwise locus 51).

Convergence plots for the founder mutation analysis

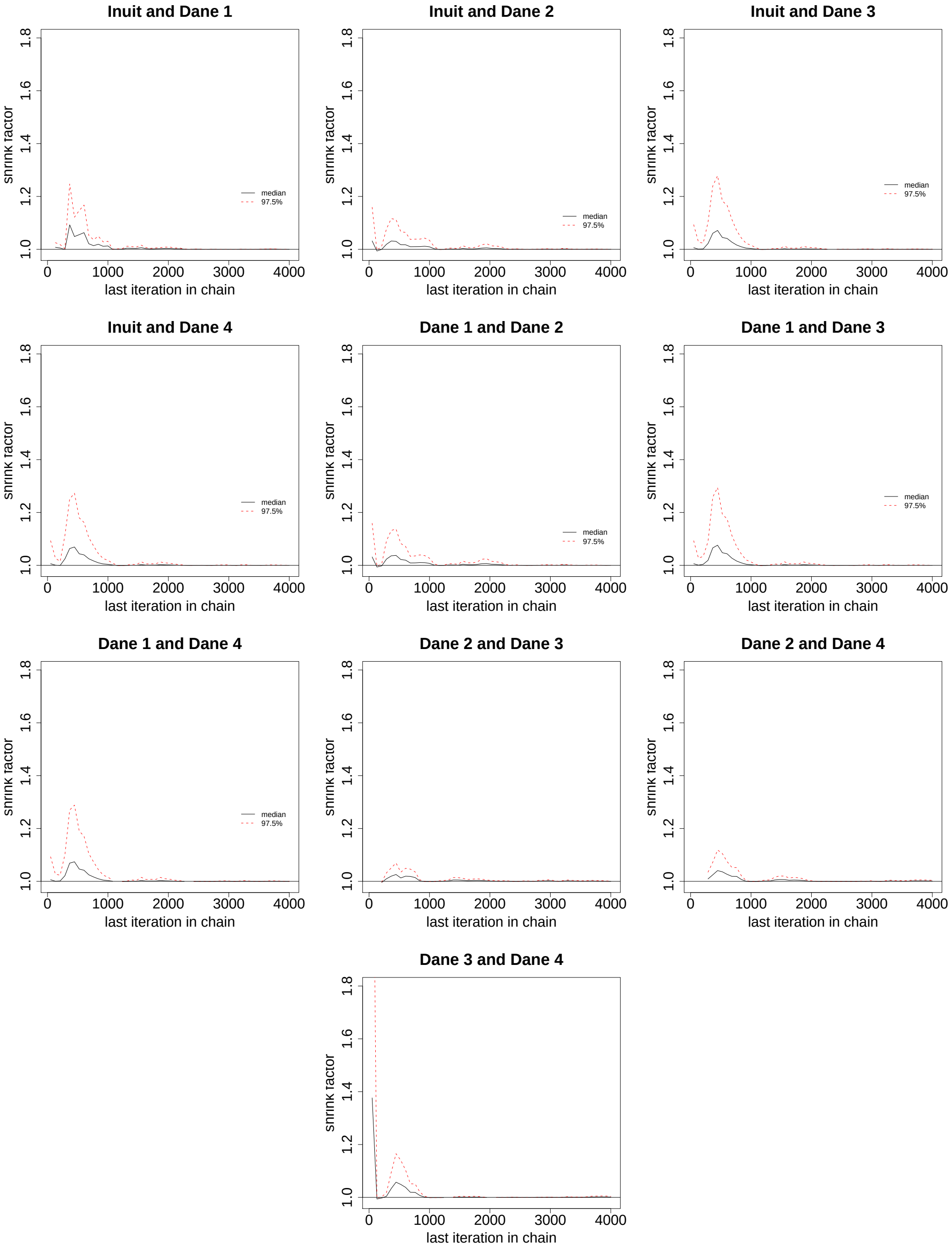

Figure S5. Gelman-Rubin plots on the statistics used to estimate pairwise relatedness for the 5 cancer patients.

Convergence plot for the mapping analysis

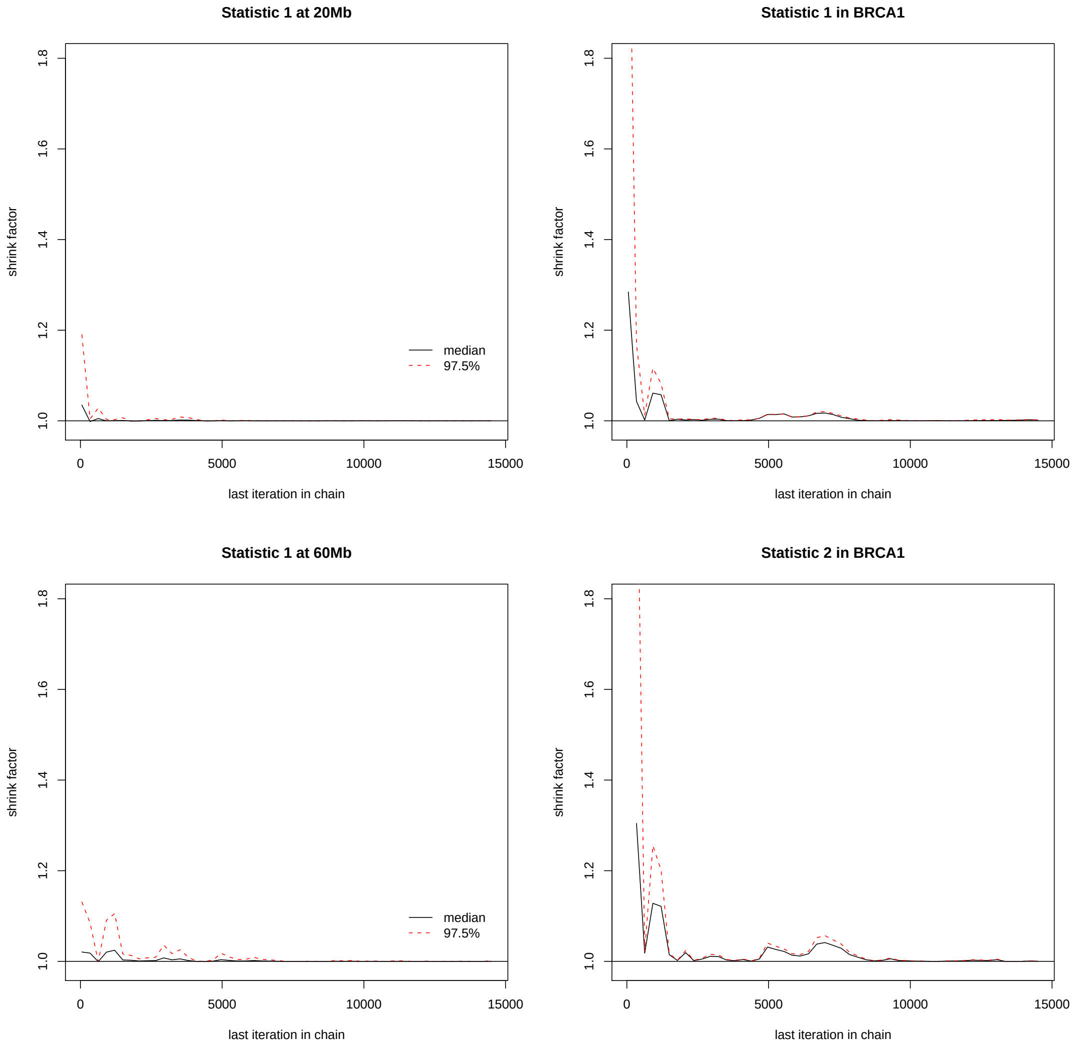

Figure S6. Gelman-Rubin plots for the two statistics used for mapping the disease causing mutation in the *BRCA1* region and two other random regions. For the first statistic (number of individuals that share at least on chromosome IBD) there is a plot for three different loci on chromosome 17; two random loci and the *BRCA1* locus. For the second statistic (all individuals IBD) there is only one plots: the plot for the *BRCA1* locus for the two random places, because not a single one of the samples in the two random regions had all 5 individuals IBD.

S7 All ROC curves for the power analysis

ROC curves for scenario 1

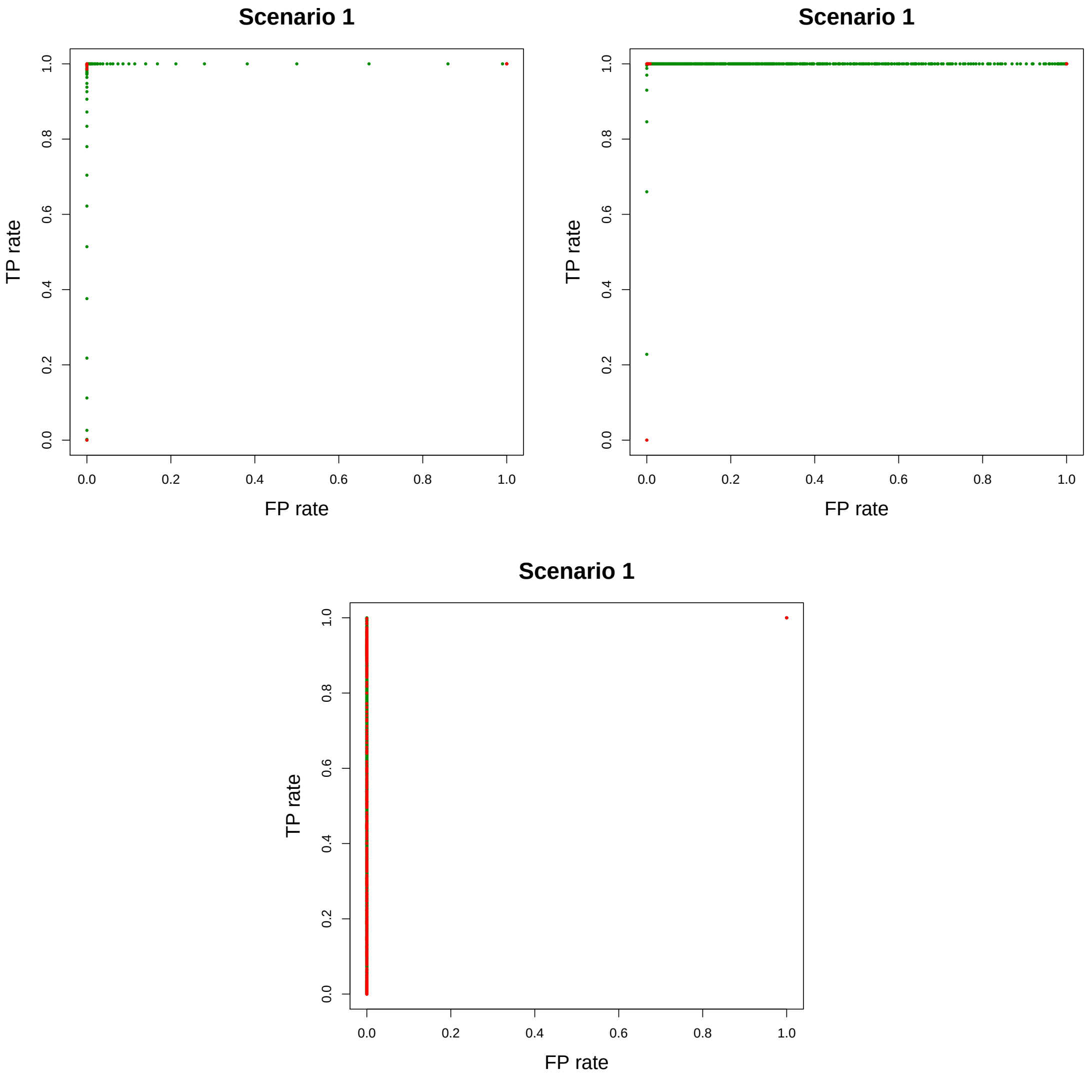

Figure S7. ROC curves for scenario 1 using three different thresholds for having inferred an IBD region: when at least half of the SNP are inferred to be IBD, when at least 95% of the SNP are inferred to be IBD and when all of the SNPs are inferred to be IBD. Since only the MCMC method and BEAGLE can infer inbreeding only ROC curves for these two methods are plotted. The green curves are from the MCMC method and the red curves are from BEAGLE.

ROC curves for scenario 2

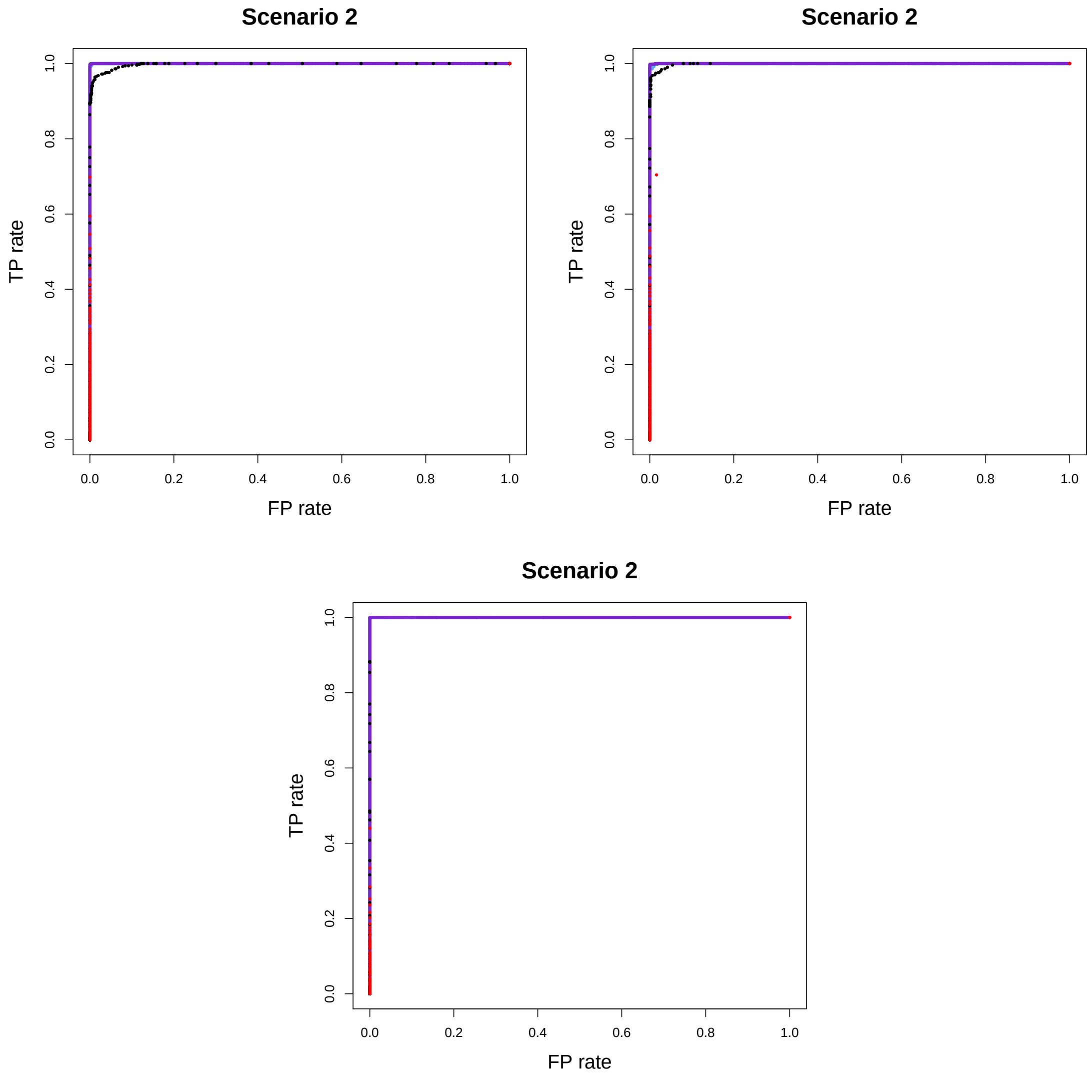

Figure S8. ROC curves for scenario 2 using three different thresholds for having inferred an IBD region: when at least half of the SNP are inferred to be IBD, when at least 95% of the SNP are inferred to be IBD and when all of the SNPs are inferred to be IBD. The green curves are the ROC curves from the MCMC method (not visible because the purple curves cover them), the purple curves are from Relate, the blue curves are from PLINK (not visible because the purple curves cover them), the red curves are from BEAGLE and the black curves are from GERMLINE.

ROC curves for scenario 3

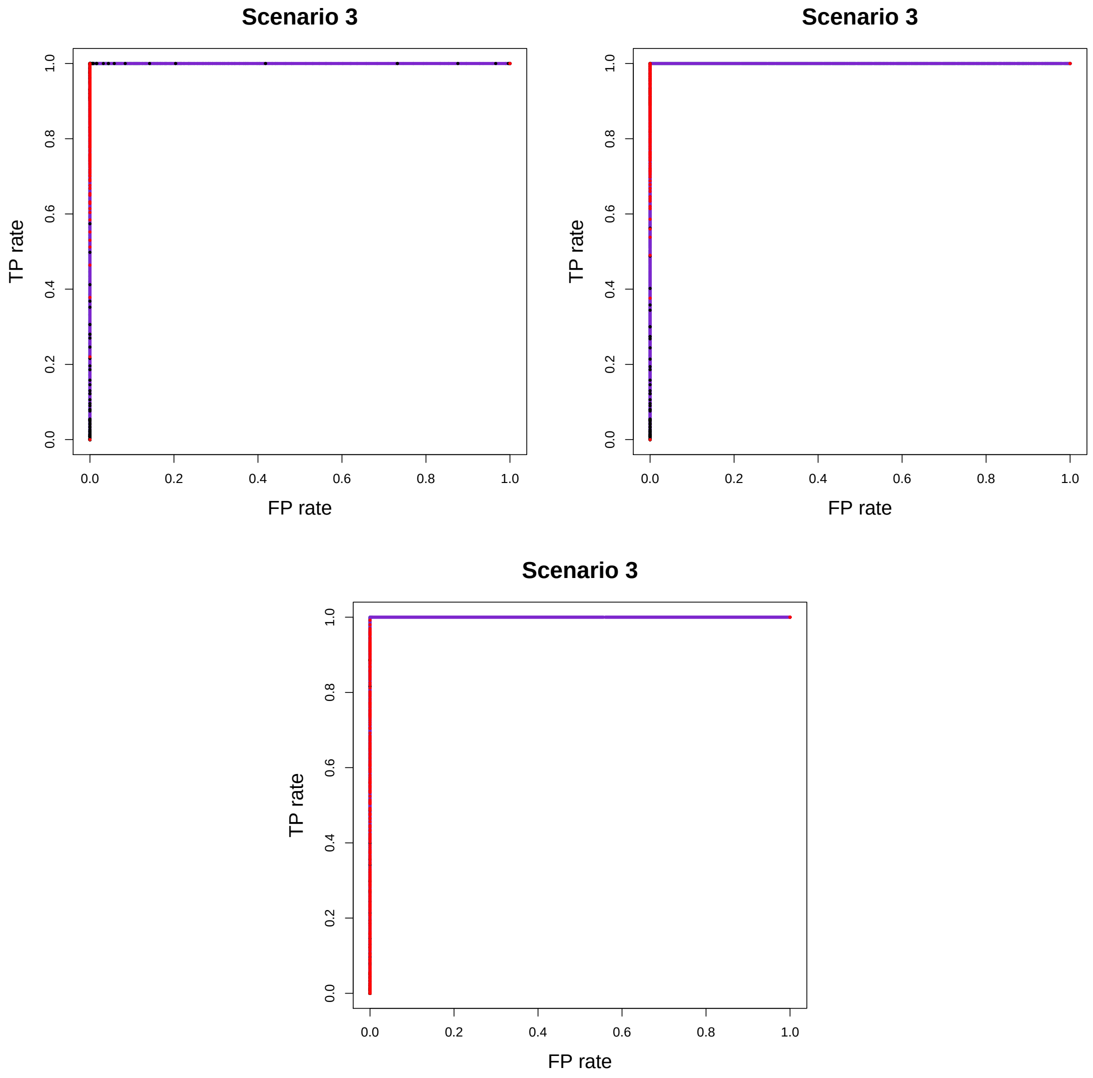

Figure S9. ROC curves for scenario 3 using three different thresholds for having inferred an IBD region: when at least half of the SNP are inferred to be IBD, when at least 95% of the SNP are inferred to be IBD and when all of the SNPs are inferred to be IBD. The green curves are the ROC curves from the MCMC method (not visible because the purple curves cover them), the purple curves are from Relate, the blue curves are from PLINK (not visible because the purple curves cover them), the red curves are from BEAGLE and the black curves are from GERMLINE.

ROC curves for scenario 4a

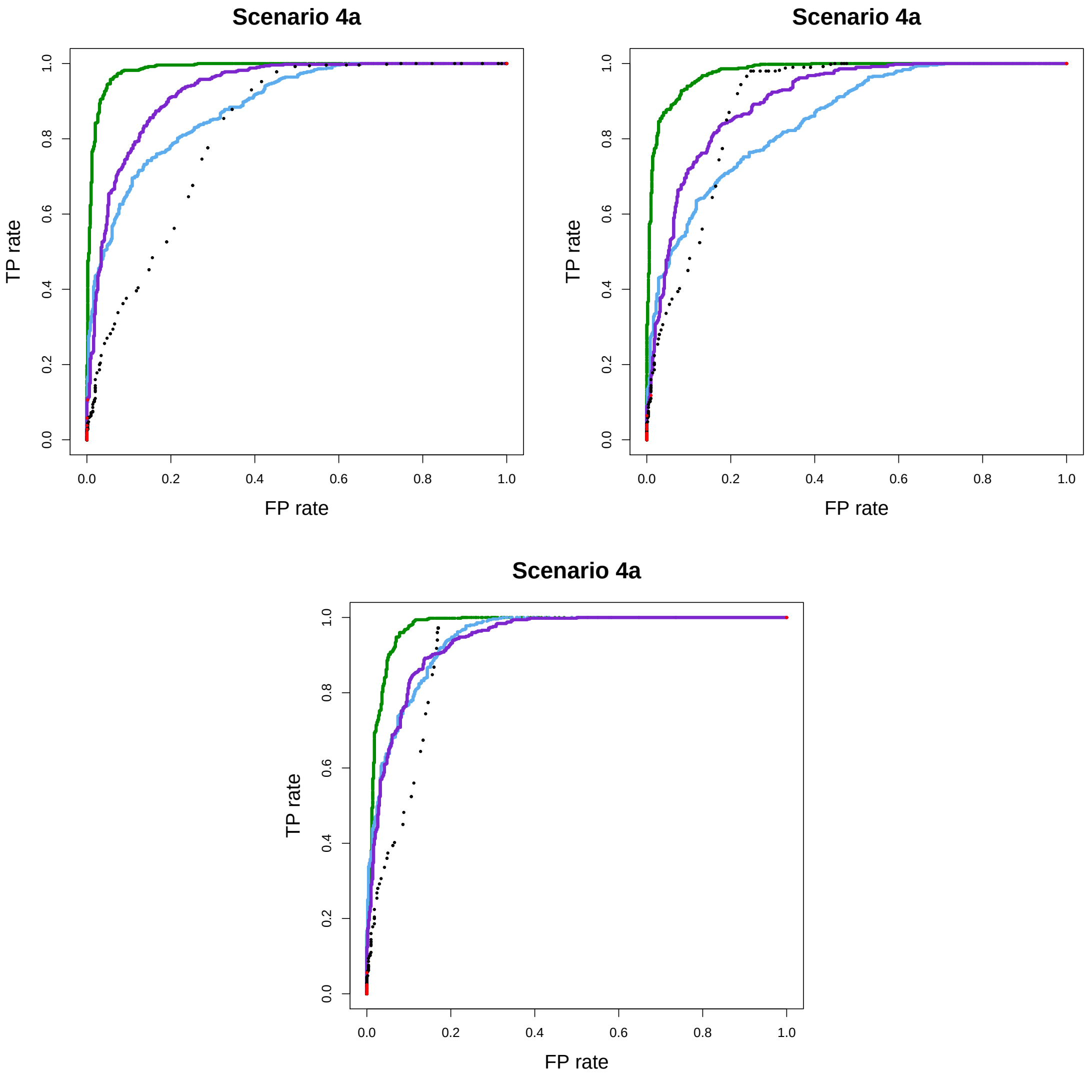

Figure S10. ROC curves for scenario 4a using three different thresholds for having inferred an IBD region: when at least half of the SNP are inferred to be IBD, when at least 95% of the SNP are inferred to be IBD and when all of the SNPs are inferred to be IBD. The green curves are the ROC curves from the MCMC method, the purple curves are from Relate, the blue curves are from PLINK, the red curves are from BEAGLE and the black curves are from GERMLINE.

ROC curves for scenario 4b

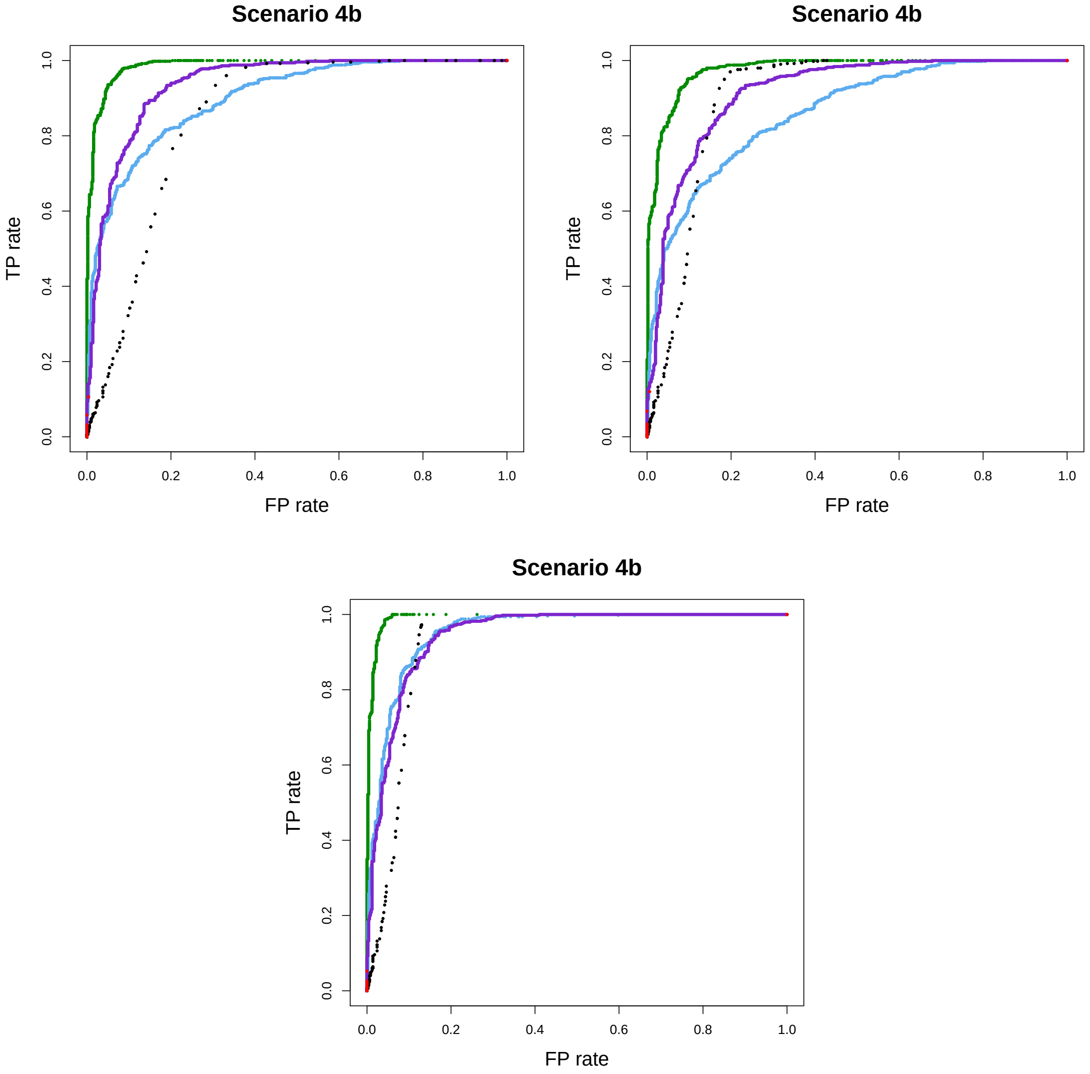

Figure S11. ROC curves for scenario 4b using three different thresholds for having inferred an IBD region: when at least half of the SNP are inferred to be IBD, when at least 95% of the SNP are inferred to be IBD and when all of the SNPs are inferred to be IBD. The green curves are the ROC curves from the MCMC method, the purple curves are from Relate, the blue curves are from PLINK, the red curves are from BEAGLE and the black curves are from GERMLINE.

ROC curves for scenario 4c

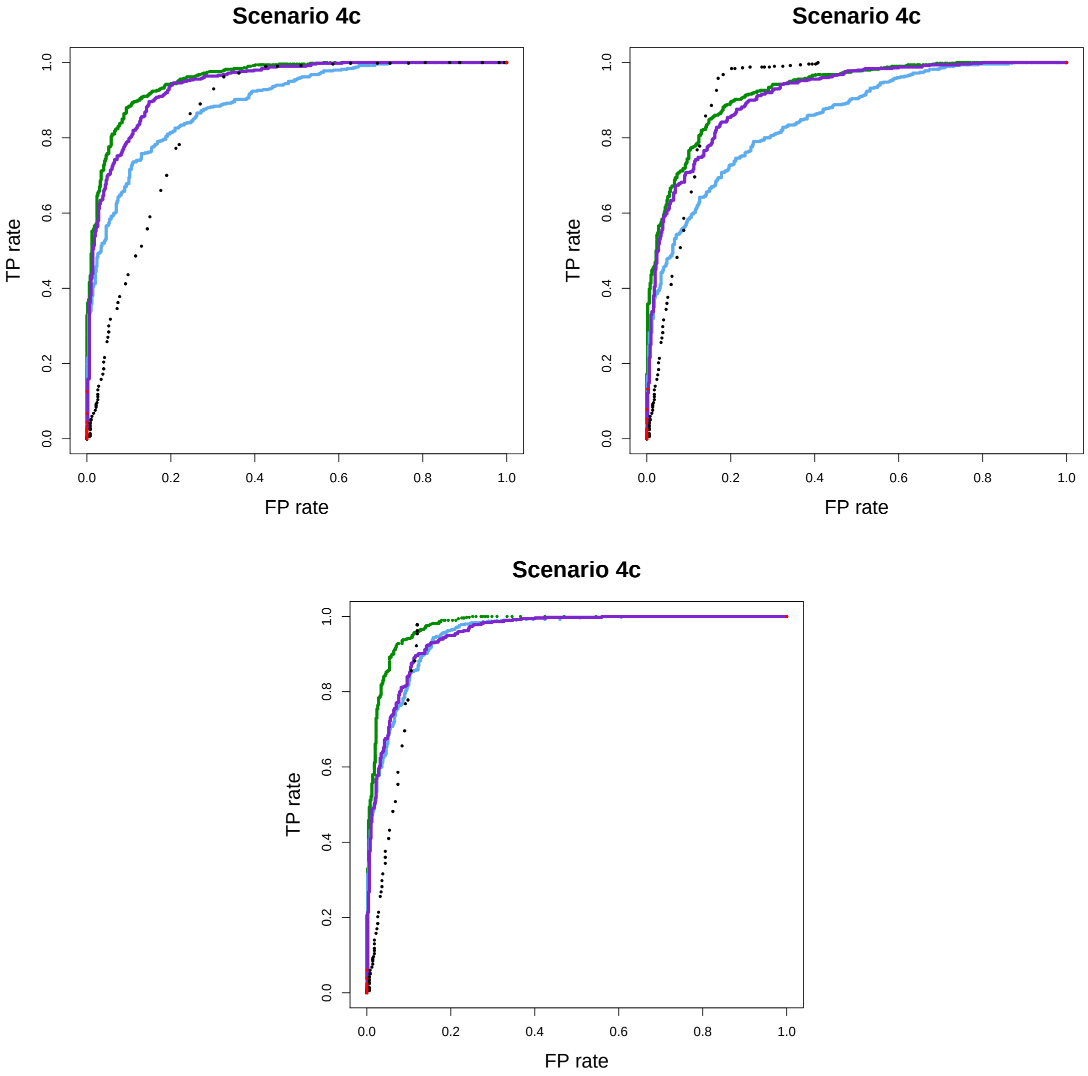

Figure S12. ROC curves for scenario 4c using three different thresholds for having inferred an IBD region: when at least half of the SNP are inferred to be IBD, when at least 95% of the SNP are inferred to be IBD and when all of the SNPs are inferred to be IBD. The green curves are the ROC curves from the MCMC method, the purple curves are from Relate, the blue curves are from PLINK, the red curves are from BEAGLE and the black curves are from GERMLINE.

S8 MAP estimates for the *BRCA1* region

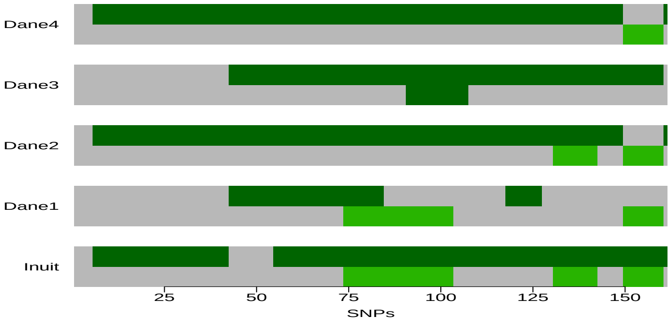

Figure S13. MCMC based estimates for 161 SNP loci in a 12 Mb region surrounding the *BRCA1* gene. For each locus the figure depicts the IBD set partitioning with the highest posterior probability, i.e. the estimated MAP IBD set partition. Here the lower chromosome of each individual has been placed in light green IBD sets (if the individual has a chromosome in this set) and the upper chromosome has been placed in the dark green (if the individual has a chromosome in this set). However it should be noted that the estimates for the different loci are made independently and that for each position the order of the sets within each individual has not been taken into account when making the MAP estimates and could hence just as well be the opposite. The *BRCA1* gene is situated close to SNP 76.

References

Albrechtsen A, Korneliussen TS, Moltke I, van Overseem Hansen T, Nielsen FC, Nielsen R 2009. Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genet Epidemiol* **33**: 266–274.

Gilks WR, Richardson S, Spiegelhalter DJ 1996. *Markov Chain Monte Carlo In Practice*. Chapman and Hall/CRC.

Affymetrix, Inc. 2006. Brlmm: an improved genotype calling method for the genechip human mapping 500k array set. http://www.affymetrix.com/support/technical/whitepapers/brlmm_whitepaper.pdf.

Leutenegger AL, Prum B, Gnin E, Verny C, Lemainque A, Clerget-Darpoux F, Thompson EA 2003. Estimation of the inbreeding coefficient through use of genomic data. *Am J Hum Genet* **73**: 516–523.

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

Scheet P Stephens M 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* **78**: 629–644.