Inbreeding Coefficient Estimation with Dense SNP Data: Comparison of Strategies and Application to HapMap III

Steven Gazal\textsuperscript{a,c}, Mourad Sahbatou\textsuperscript{d}, Hervé Perdry\textsuperscript{b,e}, Sébastien Letort\textsuperscript{f,g}, Emmanuelle Génin\textsuperscript{f,g}, Anne-Louise Leutenegger\textsuperscript{a,h}

\textsuperscript{a} Genetic Variability and Human Diseases, Inserm, U946, \textsuperscript{b} UMR 669 at \textsuperscript{c} Université Paris Sud, and \textsuperscript{d} Fondation Jean Dausset CEPH, Paris, \textsuperscript{e} Inserm, U669, Villejuif, \textsuperscript{f} Génétique, Génomique Fonctionnelle et Biotechnologies, Inserm, U1078, and \textsuperscript{g} Centre Hospitalier Régional et Universitaire de Brest, Brest, and \textsuperscript{h} Institut Universitaire d'Hématologie, UMR 946, Université Paris-Diderot, Paris, France

Key Words
Inbreeding · Homozygosity by descent · Linkage disequilibrium · Run of homozygosity · Hidden Markov model · HapMap III

Abstract
Background/Aims: If the parents of an individual are related, it is possible for the individual to have received at 1 locus 2 identical-by-descent alleles that are copies of a single allele carried by the parents’ common ancestor. The inbreeding coefficient measures the probability of this event and increases with increasing relatedness between the parents. It is traditionally computed from the observed inbreeding loops in the genealogies and its accuracy thus depends on the depth and reliability of the genealogies. With the availability of genome-wide genetic data, it has become possible to compute a genome-based inbreeding coefficient \( f \), and different methods have been developed to estimate \( f \) and identify inbred individuals in a sample from the observed patterns of homozygosity at markers. Methods: For this paper, we performed simulations with known genealogies using different SNP panels with different levels of linkage disequilibrium (LD) to compare several estimators of \( f \), including single-point estimates, methods based on the length of runs of homozygosity (ROHs) and different methods that use hidden Markov models (HMMs). We also compared the performances of some of these estimators to identify inbred individuals in a sample using either HMM likelihood ratio tests or an adapted version of the ERSA software. Results: Single-point methods were found to have higher standard deviations than other methods. ROHs gave the best estimates provided the correct length threshold is known. HMMs on sparse data gave equivalent or better results than HMMs modeling LD. Provided LD is correctly accounted for, the inbreeding estimates were very similar using the different SNP panels. The HMM likelihood ratio tests were found to perform better at detecting inbred individuals in a sample than the adapted ERSA. All methods accurately detected inbreeding up to sec-

E. Génin and A.-L. Leutenegger contributed equally to this work.
ond-cousin offspring. We applied the best method on release 3 of the HapMap phase III project, found up to 4% of inbred individuals, and created HAP1067, an unrelated and outbred dataset of this release. **Conclusions:** We recommend using HMMs on multiple sparse maps to estimate and detect inbreeding in large samples. If the sample of individuals is too small to estimate allele frequencies, we advise to estimate them on reference panels or to use 1,500-kb ROHs. Finally, we suggest to investigators using HapMap to be careful with inbred individuals, especially in the GIH (Gujarati Indians from Houston in Texas) population.

© 2014 S. Karger AG, Basel

Introduction

Marriages between relatives can occur in large populations where they can be encouraged for social and/or economic reasons as well as in small isolated populations where the number of possible partners is reduced. They lead to inbreeding in the offspring with the possibility of having received 2 alleles identical by descent (IBD), i.e., alleles that are copies of a single allele present in one of the parents’ common ancestors. The probability for the 2 alleles at 1 locus of an inbred individual to be IBD is referred to as the inbreeding coefficient [1]. Traditionally, inbreeding coefficients were estimated from the genealogies by counting the number of meiosis in the different descending paths from the common ancestors to the inbred offspring [2]. Recently, the availability of dense maps of markers spanning the whole genome has made an estimate of an individual’s genome-based inbreeding coefficient $f$ reliable, by inferring the proportion of the individual genome that is IBD from the observed marker homozygosity.

Different methods have been developed to estimate inbreeding coefficients from the observed marker data. The simplest methods rely on single-point information [3–5]. Their accuracy strongly depends on the population marker allele frequencies that can only be estimated from available population samples. This might be problematic when one wants to estimate the inbreeding coefficient of a single individual whose population of origin is not known or absent from reference control panels.

To be less sensitive to single-marker information, other methods take advantage of the fact that those alleles inherited IBD within an individual come in long stretches of adjacent markers that are all homozygous and IBD, thus defining homozygous-by-descent (HBD) segments. By quantifying the extent of HBD segments over the genome, it is then possible to obtain an estimate of the inbreeding coefficient. HBD segments can be identified over the genome by searching for runs of homozygosity (ROHs) that exceed a given length. Indeed, not all ROHs are HBD, since some ROHs that are generally among the shortest may exist because of linkage disequilibrium (LD) [6]. By focusing on ROHs longer than a given threshold, one can be confident that these are HBD segments and estimate $f$ as the proportion of the genome they cover [7]. This estimate does not depend on the allele frequency but on the ROH length threshold, and there is no consensus in the literature regarding this threshold. Several studies have used a threshold of 1,000 kb [8–11], although it has been suggested that 1,500 kb might be a better choice in European populations where some LD regions exceeding 1,000 kb have been observed [7]. Other thresholds based on genetic distances [12] or on the number of SNPs [13, 14] have also been suggested.

Another way to estimate inbreeding coefficients consists in modeling the HBD process along the genome as a Markov process and using hidden Markov models (HMMs) to account for both allele frequencies and the fact that HBD comes in stretches of adjacent markers [15]. HMMs were found to perform well on microsatellite data but are more difficult to use on dense SNP data, as the assumption of no LD between the alleles at the different markers is violated. Three approaches have been suggested to use the HMMs on SNP data. First, when the available population is large and accurate LD estimates can be obtained, a solution can consist in removing the SNPs that are in strong LD [16]. A second possibility that does not require estimating LD in the sample consists in randomly selecting one or several sparse submaps of markers that are located at given genetic distances [17, 18]. A third way to select SNPs with minimal LD consists in drawing 1 SNP in each of the LD blocks defined by the positions of the recombination hotspots estimated on HapMap data [19, 20]. Since these different approaches only select a subset of markers, some information is lost and the smallest HBD segments can be missed. To circumvent this problem, several methods have been developed recently to keep all markers and to model LD in the HMM framework. A first way to model LD consists in conditioning HMM emission probabilities on a preceding marker and replacing allele frequencies by 2-locus haplotype frequencies [18, 21, 22]. Instead of conditioning on 1 marker, it has also been proposed to condition on several previous markers through a linear model [18, 23]. Another possibility, implemented in BEAGLE, builds a tree graph of haplotypes, which automatically adapts to
the degree of LD in the genotype data from the available population [24] and integrates it in the HMMs [25, 26]. Simulation studies have shown that single-point approaches provide less reliable estimates than HMMs [16] and than ROHs [27]. ROHs were also found to outperform BEAGLE to detect HBD regions [13]. However, these different studies used different simulation scenarios and different numbers of markers and are thus difficult to compare.

For this paper, we investigated, through pedigree simulations, the accuracy of the different approaches to estimate inbreeding coefficients and to correctly identify inbred individuals in a sample of individuals genotyped on SNP chips. Our goal was to provide guidelines to investigators interested in detecting and characterizing inbred individuals in a sample using genome-wide SNP genotype data. An application to the HapMap phase III dataset [28] is presented where several inbred individuals not previously identified have been detected.

Methods

Simulations

Individual Genomes

Genome-wide SNP data were simulated for a sample of 300 individuals, including 6 first-cousin offspring (1C), 6 second-cousin offspring (2C), 18 third-cousin offspring (3C), 30 fourth-cousin offspring (4C) and 240 offspring of unrelated parents (outbred individuals; OUT). Individual genomes were simulated using gene dropping on the genealogy with the Genedrop program of MORGAN2.9 (see online suppl. fig. S1 for more details; for all online suppl. material, see www.karger.com/doi/10.1159/000358224). To have more realistic genome characteristics and LD patterns, instead of simulating founder individuals’ genotypes from allele frequencies, we used real haplotypes as founding haplotypes (see below). For each scenario, 100 replicates were simulated.

True HBD

To define true HBD in the offspring in the last generation, founder haplotype labels were used. On each replicate, the true inbreeding coefficient \( f_{true} \) of each of the 300 individuals was calculated by dividing the genome length in cM that is HBD by the total genome length obtained by adding the genetic distance between the first and the last marker on each autosome. The choice of genome length obtained by adding the genetic distance between the first and the last marker on each autosome. The choice of genome length was driven by the fact that this was the measure showing the lowest mean squared error compared to the expected inbreeding coefficient from the genealogy (data not shown).

Founder Haplotypes

First, WTCCC2 data were used. A large pool of 5,412 haplotypes was obtained from 2,706 unrelated individuals from the 1958 British Birth Cohort genotyped on the Affymetrix v6.0 chip [29, 30]. Data were phased with SHAPEIT version 2 [31]. To investigate different SNP panel densities and different population backgrounds, we then used individuals genotyped for both Affymetrix v6.0 and Illumina Human 1M SNP chips from a subset of HapMap populations [28]. Specifically, among the haplotypes distributed as part of HapMap III release 2, we selected the 226 haplotypes of the Yoruba individuals from Ibadan in Nigeria (YRI), the 232 haplotypes of the Utah residents with ancestry from Northern and Western Europe (CEU) and the 340 haplotypes combined from the Han Chinese individuals from Beijing in China (CHB) and the Japanese from Tokyo in Japan (JPT). [See online suppl. information for details on the quality control (QC) of both WTCCC and HapMap haplotype data.]

The Affymetrix v6.0 chip is later referred to as the AFFY panel (517,291 SNPs in WTCCC and 517,815 SNPs in HapMap after QC), the Illumina Human 1M chip as the ILLU panel (649,566 SNPs after QC), the union of the 2 chips as the ALL panel (987,221 SNPs after QC) and its intersection as the AFFY_ILLU panel (180,160 SNPs after QC). The lists of markers present in these different arrays were obtained from the Rutgers website (http://compgen.rutgers.edu/maps).

Genome-Based Inbreeding Coefficient Estimation Methods

All methods used to estimate \( f \) are summarized in table 1 and are described in details below. Information about the number of markers used in the different pruned or sparse maps is given in online supplementary table S1.

Single-Point Methods

Different single-point estimators based on allele frequencies have been proposed to estimate \( f \) from genetic data. Some of these methods are reviewed in Ritland [3]. In our study, we chose to compare the 4 single-point estimators that are the most commonly used in practice, as they are implemented in software to analyze genome-wide SNP data.

The first estimator we consider has been implemented in the software PLINK and is available through the option –het [4]. It is based on the genome-wide homozygous excess due to inbreeding:

\[
 f_{PLINK} = \frac{O - E}{N - E},
\]

where \( O \) is the observed number of homozygous markers of the individual, \( E \) is the expected number of homozygous markers under the Hardy-Weinberg equilibrium calculated from the allele frequencies estimated on the sample and \( N \) is the total number of markers.

The other 3 estimators are available through the –ibc option of the genome-wide complex trait analysis (GCTA) software [5]. The first estimator, GCTA1, is based on the additively recoded genotype variance (which is equivalent to the diagonal of the covariance matrix used for principal component analysis). The second estimator, GCTA2, like the PLINK estimator, is based on the homozygous excess. The third estimator, GCTA3, uses the initial definition of the inbreeding coefficient proposed by Wright in 1922 and computes the correlation between uniting gametes [32]. These estimators are based on the following formulae:

\[
 f_{GCTA1} = \frac{1}{N} \sum_{h=1}^{N} \left( \frac{Y_h - 2p_h}{h_k} \right)^2 - 1,
\]
An estimate of \( f \) can also be obtained from ROHs as the ratio of the physical length of the genome that is in ROHs to the total physical length of the genome, as proposed by McQuillan et al. [7]. A first step consists in defining ROHs and filtering out ROHs due to LD in order to only keep those that are more likely HBD regions. Different kinds of filters have been proposed in the literature and, in this study, we chose to evaluate 4 different methods of filtering to select ROH.

\[
f_{\text{GCTA2}} = 1 - \frac{1}{N} \sum_{k=1}^{N} \frac{Y_k \left(2 - 2Y_k\right)}{h_k},
\]

\[
f_{\text{GCTA3}} = \frac{1}{N} \sum_{k=1}^{N} \frac{Y_k^2 - \left(1 + 2p_k\right) Y_k + 2p_k^2}{h_k},
\]

where \( Y_k \) is the genotype coded as the number of reference allele copies for the \( k \)-th SNP, \( p_k \) is the frequency of this allele in the sample and \( h_k = 2p_k(1 - p_k) \), i.e. the expected heterozygosity. In \( f_{\text{GCTA2}} \), \( Y_k (2 - 2Y_k) = 1 \) if \( Y_k \) is heterozygous and = 0 if it is homozygous. Note that GCTA2 and PLINK are both based on homozygous excess but are not identical, contrary to what is written in the GCTA documentation. The former is a sum of ratios, whereas the latter is a ratio of sums.

In the simulation study, PLINK and GCTA were applied on each replicate of the 300 individuals and negative estimates of \( f \) were set to 0.

ROHs Methods

An estimate of \( f \) can also be obtained from ROHs as the ratio of the physical length of the genome that is in ROHs to the total physical length of the genome, as proposed by McQuillan et al. [7]. A first step consists in defining ROHs and filtering out ROHs due to LD in order to only keep those that are more likely HBD regions. Different kinds of filters have been proposed in the literature and, in this study, we chose to evaluate 4 different methods of filtering to select ROH.

<table>
<thead>
<tr>
<th>Type</th>
<th>Estimation of ( f )</th>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-point estimates</td>
<td>Proportion of markers in homozygous excess</td>
<td>PLINK</td>
<td>Homozygous excess</td>
</tr>
<tr>
<td></td>
<td>Average of independent single-point estimates</td>
<td>GCTA1</td>
<td>Variance of the additively recoded genotypes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GCTA2</td>
<td>Homozygous excess</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GCTA3</td>
<td>Correlation between uniting gametes</td>
</tr>
<tr>
<td>ROHs</td>
<td>Proportion of the genome in ROHs</td>
<td>ROH_1Mb</td>
<td>ROHs of 1,000 kb and 100 SNPs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ROH_1.5Mb</td>
<td>ROHs of 1,500 kb and 100 SNPs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ROH_1cM</td>
<td>ROHs of 1 cM and 100 SNPs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ROH_50SNP</td>
<td>ROHs of 50 SNPs on pruned data</td>
</tr>
<tr>
<td>HMMs on sparse maps</td>
<td>Value maximizing the likelihood of the HMM</td>
<td>FEstim_PRU</td>
<td>FEstim on pruned data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FEstim_1SUB</td>
<td>FEstim on one submap (1 SNP/0.05 cM)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FEstim_SUBS</td>
<td>Median of the estimates of FEstim on 100 submaps</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FEstim_HOT</td>
<td>FEstim on data delimited by hotspots</td>
</tr>
<tr>
<td>HMMs modeling LD</td>
<td>Value maximizing the likelihood of the HMM</td>
<td>FEstim_LD20</td>
<td>FEstim conditioning on one of the 20 previous markers</td>
</tr>
<tr>
<td></td>
<td>Average of HBD posterior probabilities weighted by chromosome genetic length</td>
<td>GIBDLD</td>
<td>HMM conditioning on the 20 previous markers</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BEAGLE</td>
<td>BEAGLE with LD model estimated on the sample</td>
</tr>
</tbody>
</table>

For inbreeding detection, HMMs on sparse maps methods used a likelihood ratio test; methods based on ROHs and GIBDLD used our adapted version of ERSA.

We first investigate PLINK’s –homozyg default option (referred to as ROH_1Mb). It detects ROHs by using a sliding window across the genome of an individual. The detection of ROHs is done in 2 steps: first, those SNPs that are susceptible to be in ROH are identified by looking at 50-SNP windows and selecting all SNPs encompassed by at least 5% of fully homozygous windows (while accepting 1 heterozygous and 5 missing markers in each window). Second, if at least 100 of these selected SNPs are consecutive and span over more than 1,000 kb with at least 1 SNP every 50 kb, they form a ROH that is then reported. The thresholds of 1,000 kb [9–11], 100 SNPs [14] or both [8] were widely used to detect ROHs. A threshold of 1,500 kb has also been advised [7,33] and was investigated here with PLINK’s –homozyg –homozyg-kb 1,500 that only changes the minimal length from 1,000 to 1,500 kb (referred to as ROH_1.5Mb).

We also investigated a length threshold of 1 cM, as proposed by Auton et al. [12] and implemented by default in the GERMLINE software [34] (referred to as ROH_1cM). As it is not possible to set a threshold in cM with PLINK, we replaced the physical positions by genetic positions in order to run PLINK with the default option and to have a length threshold of 1 cM and at least 1 SNP every 0.05 cM.

Following Howrigan et al.’s [13] guidelines, we also considered the threshold in number of markers (ignoring the length and density thresholds) on LD-pruned data (referred to as ROH_50SNP). Markers with a minor allele frequency <5% were removed, and a moderate LD pruning was performed (PLINK’s option –maf 0.05 –indep 50 5 2, which removes SNPs within a 50-marker window that have a multiple correlation coefficient \( R^2 > 0.5 \), i.e. a vari-
ance inflation factor >2). No heterozygous markers were allowed in the 50-SNP windows, and only ROHs with at least 50 SNPs were detected (PLINK’s options –homozyg –homozyg-window-het 0 –homozyg-snps 50 –homozyg-rgb 0 –homozyg-density 5,000 – homozyg-gap 5,000; the last 3 options are set to ignore the thresholds of length, density and maximum length between 2 markers).

To estimate \( f \) from the detected ROHs, we computed, for each individual, the total ROH genetic length in cM and divided it by the total genetic length of the genome obtained by adding the genetic distance between the first and the last marker on each autosome. Note that inbreeding estimates based on ROHs are usually derived from physical distances as originally proposed [7], but here we have found that estimates based on genetic distances are better (see online suppl. fig. S2). This was true even when \( f_{\text{true}} \) was defined as the proportion of HBD markers and so it does not depend on genetic distances (data not shown).

HMMs on Sparse Maps

The HBD states of the different markers of one individual can be modeled by HMMs as initially proposed in FEstim [15]. The observed data \( Y_k \) are the genotypes at marker \( k \), and the hidden data \( X_k \) are their HBD statuses. The emission probabilities \( P(Y_k|X_k) \) depend on the allele frequencies, and the transition probabilities depend on the genetic distance between adjacent markers and the unknown parameters of the model \( \delta \) and \( \alpha \), where \( \delta \) is the probability of being HBD at a marker, and \( \alpha \) is the instantaneous rate of change per cM from non-HBD to HBD states or equivalently that HBD segments have an expected length of \( \delta^{-1} \) cM. These parameters are then estimated by maximum likelihood, and the estimation of parameter \( \delta \) can then be used as an estimator of \( f \). This model, however, assumes that marker alleles are independent conditional on HBD status, which is not true for dense SNPs where LD may be present.

To minimize LD between SNPs, different strategies have been used. We evaluated several of them by simulations. In the first strategy, referred to as FEstim PRU, a strong LD pruning is performed on each replicate. PLINK (option –indep-pairwise 50 0.01) was used to remove SNPs that have a pairwise genotypic correlation \( r^2 > 0.01 \) within a 50-marker window. The second strategy consists in randomly extracting sparse markers every 0.5 cM to create one (FEstim_1SUB) or several (100) submaps (FEstim_SUBS). The advantage of this strategy is that it does not require estimating any LD score on the data, and it can thus be used with very small samples (even on a single individual). When 100 submaps are considered, \( f \) is estimated by the median value of the estimates obtained on the different maps after removing submaps with \( a > 1 \) as previously recommended [17].

Another way to remove LD without calculating any LD score is to rely on external information such as recombination hotspots. This idea was originally proposed by Vioight [35] using the hotspots estimated on the YRI, CEU and JPT/CHB populations of HapMap phase II [19, 20]. From the 32,996 hotspots (hg17) proposed in 2005, we obtained 32,990 in hg18 using hgLiftOver. New genetic distances and recombination intensities (cM/Mb) have been calculated from hg18 HapMap phase II genetic map files. We then only selected those 14,599 hotspots that have a recombination intensity >10 cM/Mb, as it was found to provide better \( f \) estimates than when all hotspots were kept (see online suppl. fig. S3). We ran FEstim after randomly selecting 1 marker between 2 hotspots (this estimation was labeled FEstim_HOT).

Modeling LD in HMMs

Instead of removing markers to minimize LD between SNPs, it is possible to estimate \( f \) with a modified HMM that incorporates a background LD model built from a panel of genotypes. These approaches hence cannot be applied to very small samples or single individuals, unless a large adequate control sample is available. All methods described below keep the same HBD model as FEstim, relying on parameters \( \delta \) and \( \alpha \).

The first model that was proposed in the literature conditions emission probabilities by the genotype and the HBD status of the previous adjacent marker [21]. The emission probability at \( k \) becomes \( P(Y_k|Y_{k-1}, X_k, X_{k-1}) \) and depends on the 2-locus haplotype frequencies. Instead of conditioning on the previous adjacent marker, it is also possible to condition on any previous marker \( h \) [22] after making the assumption that \( X_h = X_{h-1} \) [18]. We implemented an extension of FEstim that builds on these models and considers emission probability \( P(Y_k|Y_h, X_h, X_{h-1}) \) that depends on marker \( h \) with the highest \( r^2 \) among the 20 SNPs preceding marker \( k \) (see online suppl. information for more details). The resulting estimation of \( \delta \) by maximum likelihood was labeled FEstim_LD20.

Instead of conditioning on a single marker, Han and Abney [18, 23] developed an HMM conditioning on multiple previous markers to model LD. The emission probability at \( k \) can therefore be written as \( P(Y_k|Y_h, X_h, X_{h-L:k-1}) \), where \( Y_{h-L:k-1} \) are the genotypes at the \( L \) markers preceding marker \( k \), and is calculated from a linear model with parameters estimated on the sample. This model, labeled GIBDLD, is implemented in the IBDLD software with the parameter of the Markov chain fixed to 10^6 and parameter \( \delta \) estimated by maximum likelihood independently for each chromosome (option –ibc). We ran GIBDLD with the additional option –MAF = 0 to keep all markers, and other default options that include the setting \( L = 20 \). The software then outputs the mean HBD posterior probabilities for each chromosome. We estimated \( f \) by averaging these quantities by weighting them by the genetic length of each chromosome. We observed that it gave more accurate results than averaging the HBD posterior probabilities of all markers genome-wide, as originally suggested by Han and Abney (data not shown). Note that for 1 replicate simulated with WTCCC haplotypes, 2 replicates simulated with CEU haplotypes on the ILLU panels and for 7 replicates simulated with JPT/CHB on the ALL panel, GIBDLD was not able to estimate the parameters because of the similarity of genotype data for consecutive markers. These replicates were not used to evaluate this approach.

Another approach, implemented in BEAGLE, models LD through a tree graph of haplotypes estimated on the sample. This model allows grouping the observed haplotypes into clusters, the number of which varies at each marker position according to the settings for the paternal and maternal haplotypes, in addition to the background LD model built from a panel of genotypes. These approaches hence cannot be applied to very small samples or single individuals, unless a large adequate control sample is available. All methods described below keep the same HBD model as FEstim, relying on parameters \( \delta \) and \( \alpha \).

### Inbreeding Coefficient Estimation Accuracy

Let \( f_{\text{true}} \) and \( f \) be the true and estimated value of \( f \) for individual \( i \), and \( \Delta f = f - f_{\text{true}} \) be their difference. For each estimator, the estimation's accuracy of \( f \) is quantified by the mean of \( \Delta f \) (which is also its bias, as the expected value of \( \Delta f \) is 0), its standard deviation (sd) and its root mean square error (RMSE):

\[ \text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (f_{\text{true}} - f)^2} \]
were obtained as in Leutenegger et al. [17]. For FEstim_SUBS, which uses several maps, this test was performed on each map and the median of the p values was reported.

For the other methods, which do not rely on maximum likelihood estimates, no test has been proposed yet. A first naive approach would be to infer an individual as inbred if she/he has at least 1 HBD segment. However, even with a long length threshold, some ROHs can still be attributed to LD, and the HMMs methods would be to infer an individual as inbred if she/he has at least 1 HBD segment. However, even with a long length threshold, some ROHs can still be attributed to LD, and the HMMs methods were obtained as in Leutenegger et al. [17]. For FEstim_SUBS, which uses several maps, this test was performed on each map and the median of the p values was reported.

For the other methods, which do not rely on maximum likelihood estimates, no test has been proposed yet. A first naive approach would be to infer an individual as inbred if she/he has at least 1 HBD segment. However, even with a long length threshold, some ROHs can still be attributed to LD, and the HMMs methods were obtained as in Leutenegger et al. [17]. For FEstim_SUBS, which uses several maps, this test was performed on each map and the median of the p values was reported.

\[ \text{bias}(\Delta f) = \frac{1}{n} \sum_{i=1}^{n} \Delta f^{(i)}, \]  
\[ \text{sd}(\Delta f) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} [\Delta f^{(i)} - \text{bias}(\Delta f)]^2}, \]  
\[ \text{RMSE}(\Delta f) = \sqrt{\text{bias}(\Delta f)^2 + \text{sd}(\Delta f)^2} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} [\Delta f^{(i)}]^2}, \] 

where \( n \) is the number of estimated values.

### Inbreeding Detection in a Sample

Once inbreeding coefficient estimates are obtained, one might be interested in classifying individuals in 2 groups: inbred or not inbred.

For methods estimating \( f \) by maximizing the HMM likelihood, a likelihood ratio test contrasting the maximum likelihood and the likelihood of being outbred \( (\hat{\delta} = a = 0.001) \) along with the \( p \) value were obtained as in Leutenegger et al. [17]. For FEstim_SUBS, which uses several maps, this test was performed on each map and the median of the \( p \) values was reported.

For the other methods, which do not rely on maximum likelihood estimates, no test has been proposed yet. A first naive approach would be to infer an individual as inbred if she/he has at least 1 HBD segment. However, even with a long length threshold, some ROHs can still be attributed to LD, and the HMMs methods were obtained as in Leutenegger et al. [17]. For FEstim_SUBS, which uses several maps, this test was performed on each map and the median of the \( p \) values was reported.

\[ TPR = \frac{TP}{TP + FN} \quad \text{and} \quad FPR = \frac{FP}{FP + TN}, \]

where TP and FN are the number of inbred individuals inferred as inbred and outbred, respectively, and FP and TN are the number of outbred individuals inferred as inbred and outbred, respectively.

### Computational Time

For one sample of 300 individuals on the ALL panel, GCTA and PLINK only took 3 min each. FEstim on sparse maps also took 3 min, except FEstim_SUBS that ran for 2 h. For FEstim_LD20, our Perl algorithm to calculate haplotype probabilities between each SNP and the one among the 20 previous SNPs with the highest \( r^2 \) took 6 h, and the modified FEstim ran for 3 h. Finally, GIBDLD ran for 3 h and BEAGLE ran for 12 h. These running times were obtained on a Debian 6 server using 2 Intel Xeon E5540 processors of 2.53 GHz (2 × 4 cores, 2 × 8 threads).

### HapMap Phase III Data

Finally, we estimated the inbreeding coefficients of HapMap phase III individuals [28]. The release 3 data were used that consist in genome-wide SNP genotype data for a total of 1,397 individuals from 11 populations: 4 of African origin [African Americans from the Southwestern United States (ASW), Torubia from Ibadan in Nigeria (YRI), Luhya from Webuye in Kenya (LWK) and Maasai from Kinyawa in Kenya (MKK)], 2 of European origin [Tuscans from Italy (TSI) and Utah residents with ancestry from Northern and Western Europe (CEU)], 4 of Asian origin [Gujarati Indians from Houston in Texas (GH), Han Chinese from Beijing in China (CHB), Chinese from Metropolitan Denver in Colorado (CHD) and Japanese from Tokyo in Japan (JPT)] and 1 Mexican population [Mexican Americans from Los Angeles in California (MXL)]. This set of individuals, labeled HAPI397, is genotyped on 1,457,407 SNPs coming from 2 platforms: the Illumina Human 1M and the Affymetrix SNP 6.0. Pemberton et al. [37] found unknown relatedness among some HAPI397 individuals and defined a set of 1,117 unrelated individuals (referred to as HAP1117).

We performed a stringent QC procedure. First, following Pemberton et al. [37], we removed SNPs on sex chromosomes, SNPs departing from Hardy-Weinberg \( (p < 10^{-5}) \) in at least one HAP1117 population and SNPs monomorphic in at least one HAP1117 population. Then, we also removed SNPs that do not have a constant physical map position on the different assemblies of HapMap and dbSNP. Finally, genetic distances were calculated by interpolation of the genetic map positions from the Rutgers second-generation combined linkage physical map [38]. This genetic map has been used because it has more SNPs in common with the HapMap phase III data than the genetic map estimated on HapMap phase II. After these different QC steps, 1,024,555 SNPs were retained.

### Results

The 15 estimators described in table 1 were run on 100 replicates of a sample of 300 individuals. Each replicate was first simulated with WTCCC haplotypes to compare their accuracies. Then, the best methods were selected, and same replicates were simulated with YRI, CEU and JPT/CHB haplotypes, and with 4 different SNP panels (AFFY_ILLU, AFFY, ILLU and ALL), to observe the influence of SNP panels and population background on inbreeding estimation and detection.

### Inbreeding Coefficient Estimation Accuracy

The performances of the different estimators were first compared on the 5 different types of offspring (1C to 4C...
and OUT) of replicates simulated with WTCCC haplotypes (fig. 1). The bias, sd and RMSE of the estimators were obtained on one randomly drawn individual from each replicate (a total of 100 observations per offspring type). For some of the genealogies, offspring with no HBD segments were found. This was seen in our simulations for about 1% of 2C, 29% of 3C and 67% of 4C but for none of the 1C (see online suppl. table S2), and these proportions were concordant with the expectations based on theoretical computations. For this reason, we only considered 2C to 4C offspring who have at least 1 HBD segment ($f_{true}>0$) to study the accuracy of the estimators.

Single-point estimates were found to systematically underestimate $f$. For offspring types with $f$ close or equal to 0 (4C and OUT), biases were positive only because negative estimates were set to 0. The sds and, to a lesser extent, the RMSEs were greater than those obtained with other estimates that use information on adjacent markers.

Estimates derived from ROHs had performances that varied strongly depending on the threshold used to identify ROHs. Compared to the other 14 estimators, ROH_50SNP gave the lowest RMSEs for every inbred type. Much larger positive biases were observed with a 1-Mb or 1-cM threshold than with a 1.5-Mb or a 50-SNP threshold, suggesting that some of the ROHs >1 Mb or 1 cM could be due to LD. The results were very similar whatever the genealogy.

Estimates obtained using FEstim on sparse maps had a bias close to 0, but sds starting around 0.005 for 1C and decreasing with the depth of the genealogy. The use of several submaps (FEstim_SUBS) rather than a single one (FEstim_1SUB) provided more robust (i.e. smaller RMSE) estimates. The improvement over FEstim_HOT was, however, limited and this latter strategy that is computationally much simpler might thus be of interest. Indeed, it combines the advantage of FEstim_SUBS of not requiring the computation of any LD scores on the data making it possible to use it on small samples without its drawbacks in terms of computing time.

Comparing the different methods that attempt to model LD within the HMM framework, we found that FEstim_LD20 had a much higher bias than the other methods (around 0.03). GIBDLD performed well whatever the genealogy, whereas BEAGLE underestimated the inbreeding coefficients, especially for 1C.

Based on all these results, we see that the different methods provide similar results, with 5 methods showing the smallest RMSEs whatever the genealogy: ROH_1.5Mb, ROH_50SNP, FEstim_SUBS, FEstim_HOT and GIBDLD. For further studies, we therefore selected these 5 methods.
Fig. 2. Accuracy of the selected tests to detect inbred individuals. Box plots of TPRs and FPRs for the inbreeding tests on each replicate according to offspring type. FEstim_SUBS and FEstim_HOT use an HMM likelihood ratio test. ROH_1.5Mb, ROH_50SNP and GIBDLD use our adapted version of ERSA. TPRs on 1C are not plotted because they are all equal to 1. Replicates were simulated with WTCCC haplotypes (517,291 SNPs). The different methods and their corresponding labels are described in table 1. The color code is the same as the one used in figure 1.

Fig. 3. Influence of the number of markers on inbreeding estimation and detection. Accuracy of the 5 different selected f estimators for 1C (a) and 4C (b). Biases ± 1 sd and RMSEs (numbers/1,000 on top) were calculated on one random inbred individual (ftrue >0) from each replicate (total 100) with different SNP panels. c Box plots of TPRs and FPRs for the inbreeding tests on each replicate according to offspring type and the SNP panel. FEstim_SUBS and FEstim_HOT use an HMM likelihood ratio test. ROH_1.5Mb, ROH_50SNP and GIBDLD use our adapted version of ERSA. TPRs on 1C are not plotted because they are all equal to 1. Replicates were simulated with CEU haplotypes. The different methods and their corresponding labels are described in table 1.
Inbreeding Detection in a Sample
We compared the performances of the 5 selected methods to detect inbred individuals in the sample. This was done for the methods based on FEstim (FEstim_HOT and FEstim_SUBS) by a likelihood ratio test and for the other methods by contrasting the ROH or HBD segments detected to the population background using a method adapted from ERSA. Here, we were interested in 2 measures of accuracy: TPR and FPR that quantify the probability to truly or falsely declare that an individual is inbred, respectively. For 1C, we found that TPRs were always 1, whatever the method and the SNP panel, showing that all 1C could easily be detected in a sample. For more remote inbreeding, this was no longer the case, and we observed that, as expected, TPRs decreased as the relationship became more remote (fig. 2). This is concordant with the fact that fewer and smaller HBD segments were expected in more remote consanguinity (see online suppl. tables S2 and S3). FEstim_HOT showed much higher TPRs than the other methods. Its median TPRs were around 1, 0.7 and 0.5 for 2C, 3C and 4C, respectively. However, it also had higher FPRs (the median FPR was 0.02, instead of 0 as for the other 4 tests). FEstim_SUBS had TPRs slightly higher than tests using ERSA, while having equivalent FPRs. This result is surprising as FEstim_SUBS uses sparse maps and does not detect as small HBD segments as methods using ERSA.

Impact of the SNP Panel on Inbreeding Estimation and Detection
The performances of the different estimators and tests were compared using the HapMap CEU haplotypes and 4 different SNP panels (fig. 3). The RMSEs obtained when using the AFFY panel were very close to the ones obtained previously with WTCCC haplotypes, showing that the reduction in the number of founder haplotypes when considering HapMap data does not seem to impact the results.

In general, RMSEs were rather small for the different methods on the 4 marker panels, and we did not observe an improvement of inbreeding estimation and detection with an increase in the number of markers. The results on 1C (fig. 3a) showed that ROH_50SNP and GIBDLRD biases became less negative with the increasing number of markers. As 1C have a lot of small HBD segments, this would suggest that these 2 methods detect more small HBD segments when more SNPs are used. However, this is not true when considering more remote inbreeding such as the 4C offspring, RMSEs tended to increase on the ALL panel compared to the AFFY_ILLU one, suggesting that on this former panel, LD was probably not sufficiently accounted for or removed (fig. 3b). For inbreeding detection, the increase of TPRs was slightly better (fig. 3c). Finally, ROH_1.5Mb was more accurate for the AFFY_ILLU panel, and methods using FEstim with submaps were, as expected, not sensitive to the SNP panels.

Impact of Population Background on Inbreeding Estimation and Detection
To study how results could vary depending on the population background, we also compared the methods on replicates simulated with haplotypes from YRI (low LD level), CEU (moderate LD level) and JPT/CHB (high LD level). Figure 4a showed that the estimators are not sensitive to the population for the AFFY_ILLU panel. For the ALL panel (fig. 4b), we observed small differences on simulations performed with YRI haplotypes. Bias is negative with ROH_1.5Mb when using the YRI haplotypes, while it is positive when using the CEU or JPT/CHB haplotypes. Figure 4c showed that the TPRs of the different tests are also not sensitive to the LD level of the population.

Application to HapMap Phase III Release 3 Data
Based on the results of the simulations, we decided to use FEstim_SUBS to estimate the inbreeding coefficient on the HapMap data. This method has an estimation of \( f \) with a bias close to 0, high TPRs for inbreeding detection while keeping low FPRs and is not influenced by the population background. This last point is crucial to compare the inbreeding level of the 11 HapMap populations. This method was run on SNPs common in the AFFY and ILLU panels (183,574 SNPs). Allele frequencies were estimated by population on the individuals present in the unrelated set HAP1117.

The results on HAP1397 were plotted in figure 5 (for individual results, see online suppl. table S4). FEstim_SUBS inferred 58 inbred individuals (4.2% of the panel), with at least 1 inbred individual in every population. Populations with the highest number of inbred individuals were GIH (14), YRI and MKK (7 each) and MXL and TSI (6 each). The highest value of \( f \) (0.074) was obtained for an individual of the MXL population (NA19679) and was slightly higher than what is expected for a 1C offspring (1/16 = 0.0625, but within the normal range. In general, the \( f \) values were not very high (only 3 individuals have an \( f \) value >0.05).

Among the individuals detected as inbred, 1 CEU individual (NA12889) was inferred as inbred by FEstim_
Influence of the population background on inbreeding estimation and detection. Accuracy of the 5 different selected \( f \) estimators on 1C for the AFFY_ILLU panel (a, 180,160 SNPs) and the ALL panel (b, 987,221 SNPs). Biases ± 1 sd and RMSEs (numbers/1,000 on top) were calculated on one random 1C from each replicate (total 100) with different haplotype panels. C Box plots of TPRs and FPRs for the inbreeding tests on each replicate with the AFFY_ILLU panel according to offspring type and the haplotypes used for simulations. FEstim_SUBS and FEstim_HOT use an HMM likelihood ratio test. ROH_1.5Mb, ROH_50SNP and GIBDLD use our adapted version of ERSA. TPRs on 1C are not plotted because they are all equal to 1. The different methods and their corresponding labels are described in Table 1.

Inbreeding estimation and detection on HapMap III. FEstim_SUBS was used on SNPs common in the AFFY and ILLU panels (183,574 SNPs). Each point represents the \( f \) estimation for one individual. Closed circles represent the ones that are inferred as inbred with a likelihood ratio test (58 inbred individuals). Individuals are ordered in each population according to their \( f \) values.
SUBS without having any 1,500-kb ROH. This is due to a region of her/his chromosome 11 that had a low heterozygous rate. Thus, no ROH was detected, while the individual has been inferred as inbred with 59 submaps out of 100 that did not select heterozygous markers. So we decided not to consider this individual as inbred.

We propose to create an unrelated and outbred dataset of release 3 of HapMap phase III. In the HAP1117 unrelated dataset, 50 of the 57 individuals detected as inbred are present. We thus obtained a set of 1,067 unrelated and outbred individuals, which we labeled HAP1067 (see online suppl. table S4). We suggest that investigators use this panel when they need an unrelated and outbred dataset, 50 of the 57 individuals detected as inbred are present. We thus obtained a set of 1,067 unrelated and outbred individuals, which we labeled HAP1067 (see online suppl. table S4). We suggest that investigators use this panel when they need an unrelated and outbred dataset of HapMap III, for example to estimate allele frequencies or to obtain reference haplotypes.

Discussion

Based on the simulation study presented here, we recommend using FEstim on sparse maps. It is a strategy not influenced by the presence of LD and it works even for small sample sizes. It allows an estimation of individual inbreeding coefficients with a bias close to 0 and a good detection of inbred individuals within a sample using a likelihood framework. For the selection of sparse maps, 2 strategies are proposed: FEstim_SUBS, where several random submaps are generated, and FEstim_HOT, where one marker is randomly drawn within each genomic region delimited by recombination hotspots. FEstim_SUBS is more demanding in terms of computing time than FEstim_HOT but gave more robust estimations with FPRs around 0 for inbreeding detection. It could be the method of choice, if one needs to obtain very accurate estimates of individual inbreeding coefficients and needs to prevent a false detection of inbred individuals. FEstim_HOT, on the other hand, is easier to implement, and because it keeps more markers, around 14,000 SNPs, it detects a greater number of inbred individuals than FEstim_SUBS; however, there is also an increased risk of classifying some outbred individuals as inbred. Combining these 2 approaches by selecting several random markers in each genomic region delimited by hotspots (FEstim_HOT_SUBS) slightly improves the performances compared to FEstim_SUBS, especially when inbreeding is more remote (see online suppl. fig. S4).

Single-point estimates of inbreeding coefficients were found to have a negative bias and were not considered further in our comparative study. This result was at first unexpected. Indeed, in the presence of LD, one tends to expect a positive rather than a negative bias of single-point estimators. The literature is extensive about obtaining unbiased estimators of the expected heterozygosity at a given marker on which all these single-point inbreeding coefficient estimates are based. Several factors can lead to a biased estimate of the expected heterozygosity: the fact that it is estimated on a sample of size $N$ [39] or the fact that this sample contains inbred and/or related individuals [40]. As PLINK uses Nei and Roychoudhury’s correction (but not GCTA), we started our investigation of these issues with PLINK and found out that Nei and Roychoudhury’s correction was not effectively implemented in PLINK. This will be fixed in the next version of the software (S. Purcell, pers. commun.). Recomputing the PLINK estimation with Nei and Roychoudhury’s correction, we found null to positive biases but the RMSEs were not much changed (see online suppl. fig. S5). The positive biases are due to the fact that negative $f$s are set to zero. Note that although the bias disappears without this negative truncation, the RMSEs are still similar (not shown). So this does not change our conclusion that single-point estimators have the highest RMSEs.

For ROHs, the definition of the minimal length or the minimal number of SNPs for ROH detection is not a simple issue. Howrigan et al. [13] proposed guidelines that gave the lowest RMSEs for simulations with WTCCC haplotypes. However, we observed that their estimator slightly underestimates inbreeding coefficients in the AFFY_ILLU panel (fig. 3a) and slightly overestimates them for individuals of African origin (fig. 4b). Fine-tuning of PLINK options could result in an even better performance of this method, for example by giving different guidelines according to population origin. Recently, Pemberton et al. [41] proposed an approach to obtain a population-specific length threshold, and they applied it on HapMap III populations. We compared the $f$ estimations based on their length thresholds and the fixed 1,500-kb length threshold (see online suppl. fig. S6). The estimations were highly correlated for all populations except JPT and CHB. Indeed, the population-specific threshold for these 2 populations (around 1 Mb) gave too high estimations compared to the 1,500-kb threshold. This is consistent with what we observed in our simulation results with a 1-Mb threshold for JPT/CHB (see online suppl. fig. S7). So we feel that the issue of the minimal length or the minimal number of SNPs for ROH detection is still an open question. But if one wants to use ROHs for the estimation of $f$, we highly recommend that it be done using segment lengths expressed in genetic distances, as we found that it lowers the variability of the estimator significantly (see online suppl. fig. S2).
We were surprised that HMMs modeling LD methods (GIBLDLD and BEAGLE) did not give better results than a simple HMM on sparse data. This conclusion differs from that by Han and Abney [18] because they investigated a very sparse map with only 1 marker per cM. So it is important to note that the sparse map should not be too sparse. In addition, these methods improve their performances as sample size increases and cannot be applied to small sample sizes (here, we already had 300 individuals). In addition, the fact that no likelihood ratio test could be performed for inbreeding detection is limiting.

Another surprising result for HMMs modeling LD methods was the poor performance of FEstim_LD20. Although it significantly reduces the bias that would have been observed without modeling LD, it is still highly biased (see online suppl. fig. S8). Such a different magnitude of bias from GIBLDLD and BEAGLE can be explained by their different management of the HMM parameters. Indeed, FEstim_LD20 estimates them by maximum likelihood, while the others fix at least 1 of the parameters. By fixing HMM parameters in FEstim_LD20, we were able to obtain results similar to those of GIBLDLD and BEAGLE (see online suppl. fig. S8).

The tests for detecting significant inbreeding that we have evaluated here are very different in what they are testing. The likelihood method we have previously proposed based on FEstim estimates [17] tests whether the individual is outbred. The method derived from ERSA [36] used when no likelihood computation was possible tests whether the individual is more inbred than the population. A way to have these 2 approaches test more similar hypotheses is to modify ERSA to estimate its population parameters iteratively by keeping only the individuals inferred as outbred in the previous step (ERSAit). Indeed when we estimated the population parameters only on the 240 outbred individuals of the simulation study (ERSAout), the TPRs for ERSA were as good as the original ERSA approach (see online suppl. fig. S9). However, when applied to the real HapMap phase III data, ERSAit was found to perform differently depending on the population context. Individuals with the same inbreeding coefficient could be classified as inbred when they were sampled from a population with few inbred individuals and as outbred when they were sampled from a population where inbreeding was more frequent.

Some of the populations in the HapMap panel contain parent-offspring trio data. So we decided to compare these data to check our f estimates on the offspring with the relationships that had been inferred between the parents in previous studies. We considered the 158 known HapMap trios and 10 novel trios identified by Pemberton et al. [37], who did not detect any relationships between the parents of their 10 trios. We, on the other hand, identified 4 inbred offspring (NA19224, NA19763, NA19918 and NA21425). As their f values are small (max. 0.0129), and the 10 novel reported relationships are only up to 1C (expected inbreeding coefficient of 1/16 = 0.0625), it makes sense that these relationships have not been discovered previously.

Several studies proposed to detect HBD coming from 5 [27], 20 or 50 [13], 100 [25] and 200 [42] generations in the past. Here, we did not simulate offspring with deeper genealogy than 4C, but our simulation results showed that detecting inbreeding in 4C offspring is already very hard in the presence of LD (TPR <0.5 for FEstim_SUBS). In addition, one possible application of such detections is to find inbred cases in case-control datasets, in order to perform homozygosity mapping [43, 44] with heterogeneity to find a possible monogenic form of a complex disease [45]. For such a strategy, including individuals with HBD coming from very remote generations would be very informative but could also increase locus heterogeneity.

In conclusion, we feel that HMMs on a carefully designed sparse map strikes the right balance between extracting useful HBD information from the data and modeling complexity. It presents the advantage of enabling the use of the powerful likelihood framework for inference and hypothesis testing.

Acknowledgements

The authors would like to thank Lide Han, Mark Abney and Ben Voight for their helpful comments. We wish to thank Marie-Claude Babron for her proofreading and her helpful comments. S.G. is funded by the plateforme de génomique constitutionnelle (Faculté de Médecine, Université Paris-Diderot, Paris, France).

This study makes use of data generated by the Wellcome Trust Case Control Consortium and the Wellcome Trust Sanger Institute. The Affymetrix genotype data for individuals in the 1958 British Birth Cohort was generated by the Wellcome Trust Sanger Institute. A full list of the investigators who contributed to the generation of the Wellcome Trust Case Control Consortium data is available from www.wtcc.org.uk.

Disclosure Statement

The authors declare that they have no competing interest.
Web Resources

European Genotype Archive (repository of WTCCC data): http://www.ebi.ac.uk/ega/.
Wellcome Trust Case Control Consortium: http://www.wtccc.org.uk.
HapMap3_r2.

Genetic map files: http://compgen.rutgers.edu/maps.hgliftOver: http://genome.ucsc.edu/cgi-bin/hgliftOver.
MORGAN version 2.9: http://www.stat.washington.edu/thompson/Genepi/MORGAN.
PLINK version 1.07: http://pngu.mgh.harvard.edu/~purcell/plink.
GCTA version 1.0: http://www.complextraitgenomics.com/software/gcta.
Festim version 1.3: available upon request from anne-louise.leutenegger@inserm.fr.
IBDLD version 2.08: http://sourceforge.net/projects/ibdld.
BEAGLE version 3.2.0: http://faculty.washington.edu/browning/beagle/beagle.html.
ERSA version 1.0: http://jorde-lab.genetics.utah.edu/ersa.

All plots were done using R statistical software 2.13.1. (http://www.r-project.org).

References


30. Welcome Trust Case Control Consortium: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661–678.


35. Voight B: http://coruscant.itmat.upenn.edu/whamm.


