Published in final edited form as:

Genet Epidemiol. 2013 September ; 37(6): 635–641. doi:10.1002/gepi.21737.

Characterization and Correction of Error in Genome-Wide IBD Estimation for Samples with Population Structure

Jean Morrison

Department of Biostatistics, University of Washington, Seattle, Washington

Abstract

The proportion of the genome that is shared identical by descent (IBD) between pairs of individuals is often estimated in studies involving genome-wide SNP data. These estimates can be used to check pedigrees, estimate heritability, and adjust association analyses. We focus on the method of moments technique as implemented in PLINK [Purcell et al., 2007] and other software that estimates the proportions of the genome at which two individuals share 0, 1, or 2 alleles IBD. This technique is based on the assumption that the study sample is drawn from a single, homogeneous, randomly mating population. This assumption is violated if pedigree founders are drawn from multiple populations or include admixed individuals. In the presence of population structure, the method of moments estimator has an inflated variance and can be biased because it relies on sample-based allele frequency estimates. In the case of the PLINK estimator, which truncates genome-wide sharing estimates at zero and one to generate biologically interpretable results, the bias is most often towards over-estimation of relatedness between ancestrally similar individuals. Using simulated pedigrees, we are able to demonstrate and quantify the behavior of the PLINK method of moments estimator under different population structure conditions. We also propose a simple method based on SNP pruning for improving genome-wide IBD estimates when the assumption of a single, homogeneous population is violated.

Keywords

identity by descent; population structure; admixture; relatedness

Introduction: Bias in the Method of Moments Estimator

The method of moments IBD estimation technique implemented in PLINK [Purcell et al., 2007] and other software assumes individuals in the sample are drawn from a single, randomly mating, ancestrally homogeneous, out-bred population. The noninbreeding assumption limits to three the number of possible IBD states between two individuals at any marker—they may share 0, 1, or both alleles identical by descent from a common ancestor. Given any pair of individuals, we wish to determine the proportion of markers in each of the three states—that is if Z is the IBD state of a randomly chosen marker, we wish to find $P[Z = k]$ for $k = 0, 1, 2$. Once these proportions are estimated, they can be compared against

expected proportions for common relationships to verify pedigree information, used in heritability calculations, or used in association testing. Method of moments estimators for $P[Z = k]$ can be derived by using Hardy–Weinberg equations to calculate the conditional probabilities of observable identity by state (IBS) patterns given IBD states and applying Bayes formula [Purcell et al., 2007]. To derive the theoretical bias resulting from violating the assumptions of the method, we will restrict our attention to $\hat{\pi}_0$, this estimator for $P[Z = 0]$ given by

$$\hat{\pi}_0 \equiv P[\widehat{Z=0}] = \frac{N[I=0]}{P[I=0|Z=0]} = \frac{N[I=0]}{\sum_{i \in \mathcal{L}} 2\hat{p}_i^2 \hat{q}_i^2} \quad (1)$$

Here $\mathcal{L}$ is the set of markers and $\hat{p}_i$ and $\hat{q}_i$ are their major and minor sample allele frequencies. These are being used to approximate population allele frequencies. The numerator, $N[I = 0]$ is the number of markers observed to have IBS status 0 between the two individuals—that is the number of loci at which the two individuals are homozygous for different alleles. In the PLINK software, each of the $2\hat{p}_i^2 \hat{q}_i^2$ terms is scaled by a constant to correct for finite sample size.

If the assumptions hold then the probability of locus i having IBS state 0 is $P[Z=0] 2p_i^2 q_i^2$ and the sample allele frequency $\hat{p}$ is unbiased for the population allele frequency. Although the sample value $2\hat{p}^2 \hat{q}^2$ is not unbiased for the population value $2p^2 q^2$, it converges asymptotically to the population value as the number of haplotypes in the sample, N , increases. Therefore, this estimator is asymptotically unbiased.

$$\hat{\pi}_0 \xrightarrow{N \rightarrow \infty} \frac{P[Z=0] \sum_{i \in \mathcal{L}} 2p_i^2 q_i^2}{\sum_{i \in \mathcal{L}} 2p_i^2 q_i^2} = P[Z=0] \quad (2)$$

It is important to note that $\hat{\pi}_0$ can range from 0 to ∞ but in most common applications it is truncated at 1 to make estimates interpretable in terms of known relationships. This means even when the assumptions hold, the adjusted output of the PLINK software is not truly unbiased especially for unrelated individuals. This is typically acceptable since estimates are most often used in quality control steps to identify related individuals and check the accuracy of collected pedigree information.

If the sample consists of individuals from two or more different homogeneous populations, the unrestricted $\hat{\pi}_0$ estimator will be biased. If the sample contains admixed individuals, unrestricted $\hat{\pi}_0$ will have an inflated variance and may be biased. This behavior is a consequence of using sample allele frequencies estimated using a mixture of individuals from different populations. Capping $\hat{\pi}_0$ at 1 exacerbates the appearance of error in the estimator because many pairs will have excess inferred IBD, but the amount of underestimation possible is limited.

To characterize the effect of population structure on $\hat{\pi}_0$, first consider a simplified scenario in which the haplotypes of sampled individuals originate completely from one of two

populations A or B and are not recombinant between the two populations. Here, haplotype refers to a single copy of each chromosome. In this scenario, for a particular chromosome, we may have sampled an individual with two population A haplotypes, two population B haplotypes or one A and one B haplotype. This is the case for the first admixing generation and their offspring but does not hold in general for admixed populations where we would expect single haplotypes to have segments from different ancestral populations.

Let N_A be the number of sampled haplotypes with ancestry A and N_B be the number of sampled haplotypes with ancestry B . Let $N = N_A + N_B$ be the total number of haplotypes in the sample. Let $p_i^{(A)}$ be the allele frequency of marker i in population A . The expected value of the sample allele frequency $\hat{p}_i$ at any marker is a weighted sum of the two population allele frequencies. The expected value of the denominator in equation (1) converges asymptotically to a function of these weighted average allele frequencies rather than to a function of the true allele frequencies as in the single population scenario.

$$E[\hat{p}_i] = \frac{N_A}{N} p_i^{(A)} + \frac{N_B}{N} p_i^{(B)} \quad (3)$$

$$E \left[\sum_{i \in \mathcal{L}} 2\hat{p}_i^2 \hat{q}_i^2 \right] \xrightarrow{N \rightarrow \infty} \sum_{i \in \mathcal{L}} 2(E\hat{p}_i)^2 (E\hat{q}_i)^2 \quad (4)$$

The weights in equation (3) are a characteristic of the sample rather than being representative of the two individuals for whom we wish to estimate IBD sharing probabilities. Ideally the estimator would use population-specific allele frequencies matching the ancestral populations of the pair for whom estimation is being performed. For example, given two individuals known to have ancestries AA and AB , respectively, it would be preferable to use

$$\hat{\pi}_0 = \frac{N[I=0]}{\sum_{i \in \mathcal{L}} \left(p_i^{(A)2} q_i^{(A)} q_i^{(B)} + q_i^{(A)2} p_i^{(A)} p_i^{(B)} \right)} \quad (5)$$

However, typically both the ancestries of each pair of individuals and the desired population-specific allele frequencies are unknown. Bias in the estimator given in equation (1) results from the fact that when the sample is ancestrally heterogeneous, the sample allele frequency is not an unbiased estimator for any of the underlying population-specific allele frequencies. Using the expectations of the numerator and denominator in equation (1) when population structure is present we find

$$\hat{\pi}_0 \xrightarrow{N \rightarrow \infty} f \cdot P[Z=0] \quad (6)$$

where f , the bias factor, is a function of the sample proportions of each population, the allele frequencies in each population and the ancestries of the individuals being compared. If the

ancestries of the first individual's haplotypes are a_1 and a_2 and the ancestries of the second individual's haplotypes are a_3 and a_4 , then f will be

$$f = \frac{\sum_{i \in \mathcal{L}} \left(p_i^{(a_1)} p_i^{(a_2)} q_i^{(a_3)} q_i^{(a_4)} + q_i^{(a_1)} q_i^{(a_2)} p_i^{(a_3)} p_i^{(a_4)} \right)}{\sum_{i \in \mathcal{L}} 2E[\hat{p}_i]^2 E[\hat{q}_i]^2} \quad (7)$$

In the simplified scenario, there are six different combinations of ancestral identity for the four haplotypes, so there are six possible values of f depending on the pair. These values can be estimated assuming different sampling schemes from allele frequencies in HapMap3 populations [The International HapMap Consortium, 2007] (see Table 1). The intuitive expectation is this proportion of markers with no IBD will be underestimated when the two individuals share more than average ancestry for the sample ($f < 1$) and overestimated if the two individuals share less ancestry than average ($f > 1$). This expectation is consistent with estimates based on HapMap3 populations except for pairs of individuals each with two CEU haplotypes analyzed in a data set containing YRI haplotypes. For these pairs the estimated bias factor is slightly above 1.

For two admixed individuals the ancestral populations of the four haplotypes will vary along the chromosome so each pair of individuals will have a unique bias factor. In this case, the factor for a particular pair might be approximated as a weighted sum of the six factors from Table 1 with weights representing the proportion of the genome in each of the six ancestry states. For a sample collected from a single randomly mating admixed population, Hardy–Weinberg assumptions still hold and the average bias factor over all pairs will be close to 1. Despite this, estimates made using equation (1) remain problematic because long single ancestry blocks are broken up slowly through recombination and create persistent long-range linkage disequilibrium in the admixed population [Stephens et al., 1994]. Thus even many generations after admixture, there will be pairs of individuals with bias factors much smaller than 1. As discussed previously, bias factors greater than 1 will largely be hidden by the practice of truncating large $\hat{\pi}_0$ values to 1.

Techniques for Correcting the Estimator

We have shown how the method of moments estimators are biased when sampled individuals are drawn from distinct populations or from admixed populations because the sample allele frequencies are poor approximations of the relevant population allele frequencies for the targeted pair of individuals. This problem therefore affects any estimator that relies on sample allele frequencies, including the likelihood-based EM algorithm and the estimate of genetic covariance used by Lee et al. [Lee et al., 2011] to estimate heritability and Thornton et al. [Thornton and McPeck, 2010] to calculate association test statistics.

There are several existing approaches to obtaining accurate IBD estimates in the presence of population structure. The REAP method proposed by Thornton et al. [Thornton et al., 2012] approaches this problem by estimating a different allele frequency for each individual. This method requires prior estimation of admixture proportions for each individual. Obtaining

these estimates requires the use of reference data sets that match the ancestral populations of the sampled individuals. Although this technique works well and could be extended to use principal components rather than admixture proportions, it is currently a supervised learning approach requiring reference data. The KING method proposed by Manichaikul et al. [Manichaikul et al., 2010] also attempts to generate accurate IBD estimates in structured samples. We propose an alternative method that is simpler, requires no additional software, and which essentially uses the opposite strategy as applied in REAP. Our proposal is to eliminate markers whose allele frequencies are very different in the distinct ancestral populations (ancestrally informative markers, AIMs). This limits the marker set to only those SNPs for which the sample allele frequency is an appropriate estimate of all ancestral allele frequencies. This approach could also be applied to any other statistic that uses sample allele frequencies and could be affected by population structure.

The primary challenge of applying the marker pruning approach is identifying loci with different allele frequencies in the ancestral subgroups. In this setting, it is necessary to identify all of the AIMs on the SNP platform so they can be removed before performing IBD estimation. This means published sets of AIMs can not be used even if the ancestries of all individuals are known. These sets are usually designed to gain the most ancestral information with a minimal number of SNPs while our goal is to remove all of the ancestrally informative markers. In a large sample, AIM identification can be based only on the sample data without using reference data set and therefore without requiring knowledge of the ancestral origin of the individuals or requiring reference data sets that match the assumed ancestries. In simulations and in application to a real data set, we identified AIMs using a threshold on the association P -values between SNPs and one or more top principal components. This is equivalent to thresholding based on the absolute value of SNP loadings. Principal components-based AIM identification has many advantages but must be performed carefully since principal components analysis is also sensitive to family structure in the sample. To address this we used an iterative approach. In the first step, naive IBD proportions are calculated using method of moments. Individuals with kinship coefficient greater than 0.02 from this analysis were then removed from principal components calculations. This threshold on kinship coefficient is slightly more conservative than the one utilized in the GENEVA QC process [Laurie et al., 2010]. The resulting principal components were then used to identify and remove AIMs before recalculating a final set of IBD estimates. This iteration only needs to be performed only once since the initial IBD estimates produce a conservative set of unrelated individuals. AIM identification can be based on as many principal components as desired. In real data described below, we used the first three principal components. In simulated data, including components after the first did not lead to much improvement.

Our simulation sample size was limited by the number of admixed genomes that could be simulated from available data since construction of each admixed genome requires one full genome from each contributing ancestral group. We found we could not produce ancestrally informative principal components using small sets of admixed founders. To identify AIMs for these samples we had to perform principal components analysis including reference samples similar to the ancestral populations. This was not necessary for the simulations with

single ancestry founder genomes indicating the number of individuals needed per subgroup in a sample including nonadmixed individuals from multiple populations is much smaller than the number needed in a fully admixed sample. In practice, determining if the sample size is sufficient to generate ancestrally informative principle components for AIM identification can be tricky. One indicator of the degree to which population structure is captured by the top principal components is the proportion of variation attributable to each component. In an admixed or mixed sample, the top component might describe 5–15% of the variation in the sample depending on the ancestral groups and the proportion of variation described by each successive component will decrease rapidly. This indicates most variation due to ancestry differences is described by the first several principal components. By contrast, in the simulations for which the sample size was too small to generate ancestrally informative principal components the proportion of variance described by each component was essentially flat at around 5% across the top ten components and pairwise scatter plots showed no discernible patterns. The smaller a sample, the more an investigator will rely on preexisting knowledge of the ancestry of genotyped individuals. If there is reason to believe a sample is admixed but no structure can be detected with principal components it may be necessary to experiment with reference data sets.

Our AIM-based SNP pruning technique was developed during the quality control process for a data set containing Latin American pedigrees. This data set included 1044 samples, 960 of which self-identified as Guatemalan and Hispanic. Ten individuals self-identified as having African ancestry and 94 self-identified as European American and non-Hispanic. Estimates of IBD calculated using all the individuals resulted in many more pairs of close relatives than expected based on the collected pedigrees. We also found that many relationship pairs fell into unusual regions of the IBD plot (Fig. 1). For example, a large number of the parent-child relationships in the upper left of the plot have a substantial estimated proportion of the genome with both chromosomes shared IBD. After pruning SNPs used for estimation based on the first three principal components, these estimates are much more consistent with the collected pedigree and fall in regions consistent with common pedigree relationships.

Simulations

To assess the technique of removing AIMs using principal components analysis and to verify the bias factors estimated in Table 1, we simulated pedigree genotype data using unrelated samples from the Human Genome Diversity Project (HGDP) as pedigree founders. We used Han, Yoruban, and French samples that match closely with the HapMap CHB, YRI, and CEU populations, respectively, when principal components analysis was performed on all six populations together. We limited the samples used for simulation to those with kinship coefficient less than 0.02 from method of moments analysis on single ancestry samples. These data were phased using BEAGLE software, so haplotypes could be used. Admixed haplotypes were generated by combining two HGDP haplotypes for each desired admixed haplotype using the method described by Price et al., [Price et al., 2009] which is very similar to the method used to simulate recombination over a single generation. Using the real or generated admixed haplotypes, as founder haplotypes we simulated nonfounder genotypes over a fixed pedigree structure that includes three distinct three

generation pedigrees with a total of 21 founders and 20 nonfounder descendants. This was done by simulating recombination as a homogeneous Poisson process disregarding the centromere and using the approximation 1 Mb = 1 cM. By retaining founder genome labels from simulated recombination, we were able to calculate the true IBD proportions between each pair and compare these with estimates calculated using PLINK.

To assess the accuracy of these estimates, we focused on the estimate of the total proportion of alleles estimated to be IBD. This is equivalent to twice the kinship coefficient for outbred populations and is defined as

$$\hat{\pi} = \hat{\pi}_2 + \frac{1}{2} \hat{\pi}_1 \quad (8)$$

We used sum of squared errors the metric

$$E = \sum_{\text{All pairs}} (\hat{\pi} - \pi)^2 \quad (9)$$

to compare errors in estimates made using different marker sets.

Table 2 shows the error in IBD estimates using nine different sets of founders—three in which all founders are from the same homogeneous population, two in which founders are drawn from two populations and four with admixed founders. The four admixture scenarios considered are an 80/20 African and European admixture after five and ten generations and a 50/50 Asian and African admixture after five and ten generations. There are several sources of variability in the error in IBD estimates including: pedigree structure, the arrangement of ancestry blocks in admixed founders and recombination events over the pedigree. For these simulations, pedigree structure was fixed for all sets of founders and within each population scenario the same set of founder haplotypes was used in every simulation. The resulting standard deviation estimates capture variability due only to the random recombination process within a fixed pedigree. Every iteration, the founder haplotype set was randomly distributed among the pedigree founders and the recombination pattern over the tree was regenerated. Simulation results are based on 100 iterations. To make the results comparable across different founder sets, we used the same 100 recombination patterns for each of the nine population scenarios.

For the bi-ancestral pedigrees, pruning the marker set used for estimation based on the first principal component improved estimates dramatically. Use of additional principal components resulted in much smaller improvements or no change at all. For single ancestry pedigrees there is no violation of assumptions, so reducing the SNP set used for estimation results in a very small increase in error in the estimates. For the admixed pedigrees, we found 21 founders was not an adequate number to generate ancestrally informative principal components. In these pedigrees, a much greater improvement could be achieved by pruning SNPs using principal components calculated by including the matching HapMap3 populations. With a larger sample size, this would not be necessary. Error in estimates for the bi-ancestral pedigrees and the African-European admixed pedigrees was reduced to

within the range of error observed in the single ancestry pedigree. Elevated error persisted in the Asian-African admixed pedigrees even after removing ancestrally informative SNPs, though the error was reduced. This may be because the F_{ST} for African and Asian populations is usually estimated to be higher than for European and African populations. A stricter threshold might be required for defining ancestrally informative SNPs or a more sophisticated technique might be needed.

Figures 2–4 show the behavior of $\hat{\pi}_0$, $\hat{\pi}_1$, and $\hat{\pi}_2$ for three simulation settings. In the simulation with 4 French founders and 17 Yoruban founders both the proportion of genome shared IBD 0 and IBD 2 are overestimated while the proportion shared IBD 1 is under estimated. This pattern leads to estimated relationships that are either biologically impossible or would require a high level of inbreeding in the population and is similar to the pattern we saw in the Guatemalan data set. This pattern has no natural interpretation in terms of either family or population structure and is an artefact of the particular properties of the SNP panel in these two populations. In the simulation using founders generated as admixed between French and Yoruban ancestral groups after ten generations of admixture, the estimates are not biased but the variance of estimates using the full SNP set is large compared to the variance of estimates using the reduced marker set. This means that the corrected and uncorrected estimates have the same interpretation. When all of the founders are French, there is no need to prune the SNP set and we found estimates using the full marker set and a pruned marker set are nearly identical.

The PLINK moment-based estimators of the three IBD coefficients for a noninbred relationship are based on comparing the observed genotypes in a pair of individuals to the expectation for two independent random variables with Hardy–Weinberg frequencies. This is the same principal used to estimate the correlation between pairs of genotypes in many techniques for estimating heritability and for correcting association tests for population or family structure. The simulation and theoretical results show that when the population is out of Hardy–Weinberg equilibrium because it contains a mixture of populations these estimates no longer have a meaningful biological interpretation because their values are now dependent on features of the marker panel and the mixture proportion of the sample. When the population is in Hardy–Weinberg equilibrium as in the case of the admixed and single ancestry pedigrees, interpretation of the parameters is valid but individual ancestries become a nuisance parameter and precision is improved by eliminating markers that are informative about population structure.

Acknowledgments

Helpful comments were made by Bruce Weir.

This work was supported in part by NIH grants R01 GM 075091 and T32 GM 81062.

Many thanks to the GENEVA dental caries consortium at the University of Pittsburgh for the Guatemala data presented in Figure 1, specifically Toby McHenry, Rebecca DeSensi, Mary L. Marazita Ph.D., John Shaffer Ph.D, and Xiaojing Wang, Ph.D.

The dental caries work was supported by NIH grants R01-DE016148, R01-DE014899, U01-DE018903.

References

- Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, Bhangale T, Boehm F, Caporaso NE, Cornelis MC, Edenberg HJ. Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet Epidemiol.* 2010; 34:591–602. others doi: 10.1002/gepi.20516. [PubMed: 20718045]
- Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet.* 2011; 88:294–305. doi: 10.1016/j.ajhg.2011.02.002. [PubMed: 21376301]
- Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. *Bioinformatics.* 2010; 26:2867–2873. doi: 10.1093/bioinformatics/btq559. [PubMed: 20926424]
- Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, Beaty TH, Mathias R, Reich D, Myers S. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics.* 2009; 5:1–18. doi: 10.1371/journal.pgen.1000519.
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MaR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007; 81:559–575. doi: 10.1086/519795. [PubMed: 17701901]
- Stephens JC, Briscoe D, O'Brien SJ. Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. *Am J Hum Genet.* 1994; 55:809–824. [PubMed: 7942858]
- The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature.* 2007; 449:851–861. doi: 10.1038/nature06258. [PubMed: 17943122]
- Thornton T, McPeck MS. ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. *Am J Hum Genet.* 2010; 86:172–84. doi: 10.1016/j.ajhg.2010.01.001. [PubMed: 20137780]
- Thornton T, Tang H, Hoffmann TJ, Ochs-balcom HM, Caan BJ, Risch N. Estimating kinship in admixed populations. *Am J Hum Genet.* 2012; 91:122–138. doi: 10.1016/j.ajhg.2012.05.024. [PubMed: 22748210]

Figure 1.

Estimated proportions of the genome shared IBD 0 (π_0) are plotted versus estimated proportion shared IBD 1 (π_1). The (+) indicators mark the expected location of the common pedigree relationships parent–offspring ($E[\pi_0 = 0]$, $E[\pi_1 = 1]$), full siblings ($E[\pi_0 = 1/4]$, $E[\pi_1 = 1/2]$), half-sibling/avuncular/grand-parental ($E[\pi_0 = 1/2]$, $E[\pi_1 = 1/2]$), first cousins ($E[\pi_0 = 3/4]$, $E[\pi_1 = 1/4]$), and second cousins ($E[\pi_0 = 7/8]$, $E[\pi_1 = 1/8]$). Panel (A) shows estimates using 595,935 autosomal SNPs. Panel (B) shows estimates using only 115,585 SNPs chosen to have P -value > 0.01 for association with any of the top three principal components.

Figure 2.

Simulations results for three different simulation settings. Each panel shows the true proportion of the genome shared IBD 0, 1, or 2 plotted against the estimated proportion. There is one point for each relationship in the 21 founder pedigree and results from all 100 simulations are combined. The top row of panels shows estimates made using the full marker set and the bottom row shows estimates made using a pruned marker set.

Simulations with admixed founders. Founder generation is simulated as admixed 20% French and 80% Yoruban after 10 generations.

Figure 3.

Simulations results for three different simulation settings. Each panel shows the true proportion of the genome shared IBD 0, 1, or 2 plotted against the estimated proportion. There is one point for each relationship in the 21 founder pedigree and results from all 100 simulations are combined. The top row of panels shows estimates made using the full marker set and the bottom row shows estimates made using a pruned marker set.

Figure 4.

Simulations results for three different simulation settings. Each panel shows the true proportion of the genome shared IBD 0, 1, or 2 plotted against the estimated proportion. There is one point for each relationship in the 21 founder pedigree and results from all 100 simulations are combined. The top row of panels shows estimates made using the full marker set and the bottom row shows estimates made using a pruned marker set.

Table 1

Bias factors of mar estimated using sample allele frequencies from HapMap3 Han Chinese in Beijing, China (CHB, H), Yoruba in Ibadan, Nigeria (YRI, Y) and CEPH (CEU, C) founders. Headers indicate the ancestries of the four haplotypes of the two individuals

	HH-HH	HH-HY	HH-YY	HY-HY	YY-HY	YY-YY
50% CHB - 50% YRI	0.867	0.951	1.834	0.702	0.984	0.915
67% CHB - 33% YRI	0.890	0.976	1.883	0.721	1.011	0.939

	YY-YY	YY-YC	YY-CC	YC-YC	CC-CY	CC-CC
80% YRI - 20%CEU	0.992	0.998	1.395	0.838	1.018	1.021
50% YRI - 50%CEU	0.978	0.984	1.376	0.827	1.004	1.008

Table 2

Sum of squared errors in IBD estimates using nine different founder haplotype sets with standard deviations. Admixed founders for the French/Yoruba pedigrees are admixed with an 80% Yoruban mixture proportion. Founders for the Han/Yoruba pedigrees are admixed with a 50% Yoruban mixture proportion. For each principal component, SNPs with an association P -value of < 0.01 were pruned

Pedigree Founders	All SNPs	PC1	PC2	PC3
Single Ancestry				
Han Founders	0.041(0.008)	0.046 (0.008)	0.049 (0.009)	0.049 (0.009)
Yoruban Founders	0.019 (0.004)	0.021 (0.005)	0.022 (0.005)	0.022 (0.005)
French Founders	0.024 (0.005)	0.025 (0.006)	0.025 (0.006)	0.024 (0.006)
Bi-Ancestral				
10 Han, 11 Yoruban	0.210 (0.138)	0.025 (0.097)	0.025 (0.098)	0.026 (0.099)
4 French, 17 Yoruban	0.092 (0.023)	0.014 (0.003)	0.014 (0.003)	0.014 (0.004)
Admixed				
Han/Yoruba (5)	0.445 (0.076)	0.177 (0.033)	0.179 (0.033)	0.177 (0.033)
Han/Yoruba (10)	0.426 (0.081)	0.217 (0.037)	0.216 (0.037)	0.216 (0.037)
French/Yoruba (5)	0.190 (0.032)	0.030 (0.006)	0.037 (0.007)	0.044 (0.008)
French/Yoruba (10)	0.186 (0.032)	0.040 (0.007)	0.040 (0.007)	0.040 (0.007)