1 IBD Estimation Accuracy

We consider how well each method does at assigning a high probability to the true IBD state at a locus. To evaluate this we consider all loci with true IBD state $S = r$. From among these loci, and given a threshold, we plot the fraction of loci whose estimated probability of the locus being in the true state is below this threshold. That is, if we let $\hat{U}_r = P(S_k = r|G)$ be the estimated probability of a locus k being in state r given the genotype data, we plot the cumulative distribution function (CDF) $F_r(u) = P(\hat{U}_r \leq u|S_k = r)$. With complete information a perfect method would always estimate the probability of a locus being in its true state as 1.0, resulting in a horizontal line $F_r(u) = 0$ for $0 \leq u < 1.0$ and a spike $F_r(u) = 1.0$ for $u = 1.0$. Under incomplete information an exact, unbiased method would result in a smooth curve with higher amounts of information resulting in curves more closely matching the complete information case. Positively biased methods will have curves shifted to the right for positive IBD states and shifted to the left for the zero IBD state. That is, it would assign higher probability to positive IBD states and lower probability to the zero IBD state relative to an unbiased method.

Supplementary Figure S1 shows the CDF plots for IBD sharing of 0, 1 and 2 alleles for sibling pairs. We see that methods that include LD in the model show the highest accuracy at estimating the probability of being in a particular IBD state. Introducing missing data and genotyping error reduces the accuracy of all methods, with the MERLIN methods suffering most. Thinning the genotype data to one SNP per cM as a means of eliminating LD results in overall much poorer estimation accuracy.

For the large pedigree pairs, the CDF plots shown in Supplementary Figures S2 and S3 again show a significant improvement in accuracy for the methods that model LD over those that do not. We note that although the curve for both LD-1 and LD-20 when

$S = 8$ are closer to the ideal curve under complete information than the LD-RR curve, this does not necessarily indicate these methods are more accurate than LD-RR. With incomplete information an exact method would also produce a curve off the ideal complete information curve. Absent such a method, however, we cannot directly judge which of the proposed methods come closest to it. The location of the curves for LD-1 and LD-20, though, are consistent with those methods having somewhat higher bias than LD-RR.

Joint genotype probabilities given the condensed identity state

Here we give derivations for the joint genotype probabilities for two individuals given their condensed identity states, as listed in Supplementary Table S3. Define p_g and q_g to be the probabilities of individuals A and B to have genotype g, respectively, where $g \in \{0, 1, 2\}$ representing unordered genotypes $(0, 0)$, $(0, 1)$ and $(1, 1)$. We also define e_a to be the probability that a randomly sampled allele from person A to be of allelic type a so that $e_0 = p_0 + \frac{1}{2}p_1$ and $e_1 = p_2 + \frac{1}{2}p_1$, and similarly for f_0 and f_1 for person B. When the genotype frequencies in A and B are not equal, to obtain joint genotype probabilities conditional on the condensed identity state we must also consider a model for the process of how the genotypes are randomly drawn from their distributions. That is, given condensed identity state $S = r$ we can assume a model where the genotype for individual A is drawn first and then the genotype for individual B is drawn conditional on the genotype of A, or we can assume the reverse model where B's genotype is drawn first followed by A's genotype. In our computation of the joint genotype probability we consider both models and a function $M(\cdot, \cdot)$ of the two probabilities obtained under both models to arrive at a final probability.

The genotypes of the two individuals for condensed identity states 2, 4, 6 and 9 are independent and the issues described above regarding which individual's genotype is drawn first do not apply and the joint genotype probability is simply the product of the genotype probabilities. We now consider each of the remaining condensed identity states.

S=1: Both A and B must have the same homozygous genotype. Consider the case $G_A = G_B = 0$. The probability $P(G_B = 0|G_A = 0, S = 1)P(G_A = 0|S = 1) = 1 \cdot e_0$, because we draw a single allele from the distribution of A's allelic types. Similarly, $P(G_A = 0|G_B = 0, S = 1)P(G_B = 0|S = 1) = f_0$.

S=3: Note that the possible ordered genotypes for B are $\{[0,0],[0,1],[1,0],[1,1]\}$ with probabilities $q_0, \frac{1}{2}q_1, \frac{1}{2}q_1, q_2$, respectively, where the square bracket notation indicates that the genotypes are ordered. If we label the two detailed identity states $S_{3(1)}$ and $S_{3(2)}$, then $P(G_B|G_A, S = 3) = \frac{1}{2}P(G_B|G_A, S_{3(1)}) + \frac{1}{2}P(G_B|G_A, S_{3(2)})$. First consider the case where A's genotype is $G_A = (0,0)$. The conditional probabilities for the ordered genotypes are $P(G_B = [0,0]|G_A = (0,0), S_{3(1)}) = P(G_B = [0,0]|G_A = (0,0), S_{3(2)}) = q_0/(q_0 + \frac{1}{2}q_1)$, $P(G_B = [0,1]|G_A = (0,0), S_{3(1)}) = \frac{1}{2}q_1/(q_0 + \frac{1}{2}q_1)$ and $P(G_B = [1,0]|G_A = (0,0), S_{3(2)}) = \frac{1}{2}q_1/(q_0 + \frac{1}{2}q_1)$, with the other probabilities for G_B being 0. The conditional probability for the unordered genotype given the condensed identity state is obtained by summing over the possible ordered genotypes, leading to $P(G_B = (0,0)|G_A = (0,0), S = 3) = q_0/(q_0 + \frac{1}{2}q_1)$ and $P(G_B = (0,1)|G_A = (0,0), S = 3) = \frac{1}{2}q_1/(q_0 + \frac{1}{2}q_1)$. The joint probability is the product of the above probability with $P(G_A = (0,0)|S = 3) = e_0$. The probabilities for $G_A = (1,1)$ are easily obtained from symmetry. If we condition first on $G_B = (0,0)$ we have $P(G_B = (0,0)|S = 3) = q_0$ along with the conditional probability $P(G_A = (0,0)|G_B = (0,0), S = 3) = 1$. In the case where $G_B = (0,1)$ we obtain $P(G_B = (0,1)|S = 3) = q_1$ and $P(G_A = (0,0)|G_B = (0,1), S = 3) = \frac{1}{2}$, because either the 0 or 1

allele in B will be IBD with the alleles in A.

S=5: This is identical to the $S = 3$ case, but with A and B switched.

S=7: Let the observed genotypes be $G_A = g$ and $G_B = h$. The joint probability, when first conditioning on G_A is $P(G_B = h|G_A = g, S = 7)P(G_A = g|S = 7) = 1_{g=h}p_g$, where 1_X is the indicator function and equals 1 when X is true and 0 when X is false. When first conditioning on G_B, the joint probability is $P(G_A = g|G_B = h, S = 7)P(G_B = h|S = 7) = 1_{g=h}q_h$.

S=8: First, condition on $G_A = (0, 0)$, then the conditional probabilities for G_B are the same as when $S = 3$ since in both cases we know that B must have a 0 allele that is IBD with an allele from A. The probability for G_A, however, is different and is $P(G_A = (0, 0)|S = 8) = p_0$. In the case where $G_A = (0, 1)$, either the 0 or the 1 allele is IBD. With probability $\frac{1}{2}$ the 0 allele is IBD and the conditional probabilities for G_B to be (0,0) or (0,1) are $q_0/(q_0+\frac{1}{2}q_1)$ or $\frac{1}{2}q_1/(q_0+\frac{1}{2}q_1)$, respectively, for the same reasons as given above. Similarly, with probability $\frac{1}{2}$ the 1 allele is IBD and the conditional probabilities for G_B to be (1,1) or (0,1) are $q_2/(q_2+\frac{1}{2}q_1)$ or $\frac{1}{2}q_1/(q_2+\frac{1}{2}q_1)$, respectively. This gives the following conditional probabilities, $P(G_B = (0, 0)|G_A = (0, 1), S = 8) = \frac{1}{2}q_0/(q_0 + \frac{1}{2}q_1)$, $P(G_B = (0, 1)|G_A = (0, 1), S = 8) = \frac{1}{4}q_1/(q_0 + \frac{1}{2}q_1) + \frac{1}{4}q_1/(q_2 + \frac{1}{2}q_1)$, $P(G_B = (1, 1)|G_A = (0, 1), S = 8) = \frac{1}{2}q_2/(q_2+\frac{1}{2}q_1)$. The remaining cases are easily determined from symmetry arguments.

Note that in the limit where the genotype probabilities in both A and B are identical and equal to the product of the allele frequencies, the genotype probabilities given in Supplementary Table S3 reduce to the genotype probabilities in Supplementary Table S1.

Table S1 Genotype probabilities given the condensed identity state

S	$Pr(G^1 = (a,b), G^2 = (c,d)\|S)^\dagger$
1	$\delta_{ab}\delta_{ac}\delta_{ad}f_a$
2	$\delta_{ab}\delta_{cd}f_af_c$
3	$\frac{1}{2}(2-\delta_{cd})\delta_{ab}(\delta_{ac}f_af_d+\delta_{ad}f_af_c)$
4	$(2-\delta_{cd})\delta_{ab}f_af_cf_d$
5	$\frac{1}{2}(2-\delta_{ab})\delta_{cd}(\delta_{ca}f_af_b+\delta_{cb}f_cf_a)$
6	$(2-\delta_{ab})\delta_{cd}f_cf_af_b$
7	$\frac{1}{2}(2-\delta_{ab})(2-\delta_{cd})(\delta_{ac}\delta_{bd}f_af_b+\delta_{ad}\delta_{bc}f_af_b)$
8	$\frac{1}{4}(2-\delta_{ab})(2-\delta_{cd})(\delta_{ac}f_af_bf_d+\delta_{ad}f_af_bf_c+\delta_{bc}f_bf_af_d+\delta_{bd}f_bf_af_c)$
9	$(2-\delta_{ab})(2-\delta_{cd})f_af_bf_cf_d$

$^\dagger\delta_{ab}$ is the Kronecker delta function: $\delta_{ab}=\begin{cases}1 & \text{if } a=b\\ 0 & \text{otherwise}\end{cases}$ and f_a is the allele frequency of allelic type a.

Table S2 The probability for the observed genotype O_i^p given the true genotype G_i^p with error rate ϵ.

$P(O_i^p\|G_i^p,\epsilon)$	$O_i^p=0$	$O_i^p=1$	$O_i^p=2$
$G_i^p=0$	$(1-\epsilon)^2$	$2\epsilon(1-\epsilon)$	ϵ^2
$G_i^p=1$	$\epsilon(1-\epsilon)$	$\epsilon^2+(1-\epsilon)^2$	$\epsilon(1-\epsilon)$
$G_i^p=2$	ϵ^2	$2\epsilon(1-\epsilon)$	$(1-\epsilon)^2$

Table S3 Joint genotype probabilities for each pairwise genotype given the condensed identity state

	G^1, G^2 or G^2, G^1								
S	$0,0$	$0,1$	$0,2$	$1,0$	$1,1$	$1,2$	$2,0$	$2,1$	$2,2$
1	$M(e_0, f_0)$	0	0	0	0	0	0	0	$M(e_1, f_1)$
2	$e_0 f_0$	0	$e_0 f_1$	0	0	0	$e_1 f_0$	0	$e_1 f_1$
3	$q_0 M(\frac{e_0}{f_0}, 1)$	$\frac{1}{2} q_1 M(\frac{e_0}{f_0}, 1)$	0	0	0	0	0	$\frac{1}{2} q_1 M(\frac{e_1}{f_1}, 1)$	$q_2 M(\frac{e_1}{f_1}, 1)$
4	$e_0 q_0$	$e_0 q_1$	$e_0 q_2$	0	0	0	$e_1 q_0$	$e_1 q_1$	$e_1 q_2$
5	$p_0 M(\frac{f_0}{e_0}, 1)$	0	0	$\frac{1}{2} p_1 M(\frac{f_0}{e_0}, 1)$	0	$\frac{1}{2} p_1 M(\frac{f_1}{e_1}, 1)$	0	0	$p_2 M(\frac{f_1}{e_1}, 1)$
6	$f_0 p_0$	0	$f_1 p_0$	$f_0 p_1$	0	$f_1 p_1$	$f_0 p_2$	0	$f_1 p_2$
7	$M(p_0, q_0)$	0	0	0	$M(p_1, q_1)$	0	0	0	$M(p_2, q_2)$
8	$p_0 q_0 M(\frac{1}{f_0}, \frac{1}{e_0})$	$\frac{1}{2} p_0 q_1 M(\frac{1}{f_0}, \frac{1}{e_0})$	0	$\frac{1}{2} p_1 q_0 M(\frac{1}{f_0}, \frac{1}{e_0})$	$\frac{1}{4} p_1 q_1 M(\frac{1}{f_0} + \frac{1}{f_1}, \frac{1}{e_0} + \frac{1}{e_1})$	$\frac{1}{2} p_1 q_2 M(\frac{1}{f_0}, \frac{1}{e_1})$	0	$\frac{1}{2} p_2 q_1 M(\frac{1}{f_1}, \frac{1}{e_1})$	$p_2 q_2 M(\frac{1}{f_1}, \frac{1}{e_1})$
9	$p_0 q_0$	$p_0 q_1$	$p_0 q_2$	$p_1 q_0$	$p_1 q_1$	$p_1 q_2$	$p_2 q_0$	$p_2 q_1$	$p_2 q_2$

Table S4 Bias and RMSE of estimated chromosome-wide kinship coefficients

			Sibling pairs		Large pedigree pairs	
Method	Missing rate	Error rate	Bias	RMSE	Bias	RMSE
NoLD	0	0	0.0226	0.0264	0.0639	0.0655
	0.05	0.02	0.0233	0.0271	0.0640	0.0656
NoLD-S	0	0	0.0123	0.0406	0.0169	0.0345
	0.05	0.02	0.0123	0.0501	0.0174	0.0399
LD-1	0	0	0.0100	0.0117	0.0195	0.0211
	0.05	0.02	0.0146	0.0121	0.0253	0.0276
LD-20	0	0	0.0026	0.0040	0.0051	0.0065
	0.05	0.02	0.0103	0.0121	0.0137	0.0161
LD-RR	0	0	0.0029	0.0057	0.0050	0.0069
	0.05	0.02	-0.0068	0.0107	0.0011	0.0077
MERLIN	0	0	0.0187	0.0207	-	-
	0.05	0.02	-0.0646	0.0714	-	-
MERLIN-CL	0	0	0.0010	0.0069	-	-
	0.05	0.02	-0.1178	0.1266	-	-