BIOINFORMATICS

Vol. 27 ISMB 2011, pages i333–i341
doi:10.1093/bioinformatics/btr243

Reconstruction of genealogical relationships with applications to Phase III of HapMap

Sofia Kyriazopoulou-Panagiotopoulou[1,*,†], Dorna Kashef Haghighi[1,†], Sarah J. Aerni[1,2,†], Andreas Sundquist[1,3], Sivan Bercovici[1] and Serafim Batzoglou[1]

[1]Department of Computer Science, Stanford University, [2]Biomedical Informatics Training Program, Stanford University School of Medicine, Stanford, CA 94305 and [3]DNAnexus, Inc., Palo Alto, CA 94301, USA

ABSTRACT

Motivation: Accurate inference of genealogical relationships between pairs of individuals is paramount in association studies, forensics and evolutionary analyses of wildlife populations. Current methods for relationship inference consider only a small set of close relationships and have limited to no power to distinguish between relationships with the same number of meioses separating the individuals under consideration (e.g. aunt–niece versus niece–aunt or first cousins versus great aunt–niece).

Results: We present CARROT (ClAssification of Relationships with ROTations), a novel framework for relationship inference that leverages linkage information to differentiate between *rotated* relationships, that is, between relationships with the same number of common ancestors and the same number of meioses separating the individuals under consideration. We demonstrate that CARROT clearly outperforms existing methods on simulated data. We also applied CARROT on four populations from Phase III of the HapMap Project and detected previously unreported pairs of third- and fourth-degree relatives.

Availability: Source code for CARROT is freely available at http://carrot.stanford.edu.

Contact: sofiakp@stanford.edu

1 INTRODUCTION

Algorithms for relationship inference can greatly benefit association and linkage studies by detecting undeclared or misspecified relatives, and have applications in forensics and wildlife population management. Existing methods (Epstein *et al*., 2000; McPeek and Sun, 2000; Sun *et al*., 2002) based on Hidden Markov Models (HMMs) only consider a small number of relationship types. RELPAIR (Epstein *et al*., 2000) examines eight types of relationships (full siblings, parent–child, avuncular, grandparent–grandchild, half siblings, first cousins, monozygotic twins, and unrelated). PREST-plus (McPeek and Sun, 2000; Sun *et al*., 2002) extends this set of alternative relationships to include the relationship types half avuncular, half first cousins and half-siblings first-cousins. Algorithms for pedigree reconstruction (Berger-Wolf *et al*., 2007; Koch *et al*., 2008; Riester *et al*., 2009, 2010) partition the individuals into sets of sibling and parent–child pairs (Blouin, 2003; Jones and Arden, 2003) and are not designed for datasets containing distantly related individuals. Additionally, the methods mentioned above cannot differentiate between *rotated* relationships, that is, between relationships with the same number of common ancestors and the same number of meioses separating the individuals under consideration (e.g. aunt–niece versus niece–aunt or first cousins versus great aunt–niece). Distinguishing between such relationships facilitates pedigree reconstruction from relatively distant relationships.

We developed CARROT (ClAssification of Relationships with ROTations), a new framework for relationship inference that leverages linkage information to differentiate between *rotated* relationships. As suggested by Skare *et al*. (2009), we grouped relationships between outbred individuals into three categories, based on whether the individuals under consideration have one or two common ancestors or whether one of them is the ancestor of the other. Building on the ideas of Stankovich *et al*. (2005) and Bercovici *et al*. (2010), we defined three sets of HMMs, one for each category of relationships. CARROT uses a novel heuristic to decide whether one of the two individuals under consideration is closer to the common ancestors than the other and benefits from haplotype phasing in order to distinguish between certain *rotated* relationships.

We demonstrated that CARROT achieves higher accuracy than maximum-likelihood approaches, such as RELPAIR and PREST-plus, using simulated data. We also applied CARROT on real data from Phase III of the HapMap Project (The International HapMap 3 Consortium, 2010). In addition to validating relationships reported by Pemberton *et al*. (2010), we identified previously unreported third- and fourth-degree relationships in this dataset.

2 METHODS

2.1 Overview of CARROT

CARROT is a framework for predicting the relationship type for a pair of individuals from their genotypes. Similarly to previous work (Epstein *et al*., 2000; McPeek and Sun, 2000; Sun *et al*., 2002), CARROT uses HMMs to compute the likelihood of the genotype data under a set of alternative relationships. Unlike existing methods, which define a different HMM for each relationship type, CARROT defines three HMM templates from which it generates HMMs for a broad range of relationships.

CARROT uses the likelihoods computed by all alternative models as features of a classifier together with additional features that quantify the overall genetic sharing between the two individuals. In this work, we use the term *Identity By Descent* or *Identical By Descent* (IBD) to refer to genomic regions inherited from the Most Recent Common Ancestors (MRCAs) of the two individuals. The additional features we consider are the percentage of IBD, the number of transitions between IBD and non-IBD regions and the *haplo-frequencies* of the two individuals in these regions. Roughly speaking, the *haplo-frequency* of an individual A in a given genomic region is the

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

fraction of the reference population that has the same haplotype as A in that region.

The use of phased data enables CARROT to assign different likelihoods to relationships that are indistinguishable by previous methods: in the case of an aunt–niece pair, for instance, segments inherited from the common ancestors can lie on either haplotype of the aunt, but must lie on the same haplotype of the niece. This haplotype model, combined with linkage information captured in the *haplo-frequencies*, gives CARROT a significant advantage in differentiating between *rotated* relationships compared with the existing methods.

2.2 HMMs and factorial HMMs

A HMM is a probabilistic model for capturing the dependencies of a sequence of observations $G_1, G_2, \ldots, G_M$ on a chain of unknown (or hidden) variables $S_1, S_2, \ldots, S_M$ taken from a set $\mathcal{S}$. An HMM makes the following conditional independence assumptions: first, given S_k, G_k is conditionally independent of all observations and hidden states, that is $P(G_k|S_1,\ldots,S_k,G_1,G_{k-1}) = P(G_k|S_k)$. Second, given S_{k-1}, S_k is conditionally independent of all previous hidden states, that is, $P(S_k|S_1,\ldots,S_{k-1}) = P(S_k|S_{k-1})$. An HMM is, therefore, defined by a set of transition probabilities $P(S_k|S_{k-1})$, a set of emission probabilities $P(G_k|S_k)$ a probability distribution over the initial states.

Often, we want to infer the value of the hidden variables from the observed variables. The posterior probability $P(S_i|G)$ can be computed using the forward–backward algorithm in time $O(M|\mathcal{S}|^2)$ (Rabiner and Juang, 1986), where $|\mathcal{S}|$ is the number of values in $\mathcal{S}$.

In a factorial HMM (Ghahramani and Jordan, 1997), the observation at position k depends on multiple hidden variables, $S_k^1, S_k^2, \ldots, S_k^T$, which are assumed to evolve independently, that is:

$$P(S_k = (s_k^1, s_k^2, \ldots, s_k^T) | S_{k-1} = (s_{k-1}^1, s_{k-1}^2, \ldots, s_{k-1}^T)) = \prod_{t=1}^{T} P(S_k^t = s_k^t | S_{k-1}^t = s_{k-1}^t)$$

A factorial HMM where each hidden variable S_k^t takes values from the set $\mathcal{S}$ is equivalent to an HMM with hidden variables taking values from the Cartesian product $\mathcal{S}^T$. Using the latter representation, running the forward–backward algorithm on a factorial HMM requires $O(M|\mathcal{S}|^{2T})$ time. However, by taking advantage of the independence assumptions for the hidden variables, the forward–backward algorithm can be modified to run in $O(MT|\mathcal{S}|^{T+1})$ time, which is a significant improvement when the number of hidden variables T is large.

2.3 Likelihood computation assuming linkage equilibrium

We want to infer the relationship between two individuals A and B, genotyped at a set of M unlinked SNPs. Let $H_{A0}, H_{A1}, H_{B0}, H_{B1} \in \{A, C, G, T\}^M$ be the two haplotypes of A and B, respectively, $G_A = (H_{A0}, H_{A1})$, $G_B = (H_{B0}, H_{B1})$ be their ordered, or phased, genotypes, and θ_k be the probability of recombination between SNPs k and $k+1$. Throughout this work, we assume that θ_k is the same for both sexes.

Let $\mathcal{R}$ be a set of putative relationships for individuals A and B. For any relationship $R \in \mathcal{R}$, we want to compute the likelihood of R, or the probability of the observed genotypes under the assumption that the true relationship between A and B is R, $L_R = P(G_A, G_B|R)$. As noted in Skare *et al.* (2009), assuming that A and B are not inbred, their relationship must fall into exactly one of the following categories:

(1) A and B share exactly 2 MRCAs (e.g. full siblings, first cousins);

(2) A and B share exactly 1 MRCA (e.g. half siblings, half cousins); and

(3) A is the ancestor of B or vice versa.

We call relationships R_1 and R_2 *rotated*, if R_1 and R_2 are in the same relationship category and the total number of meioses between the two individuals is the same in R_1 and R_2. Alternatively, we say that R_1 is a *rotation* of R_2.

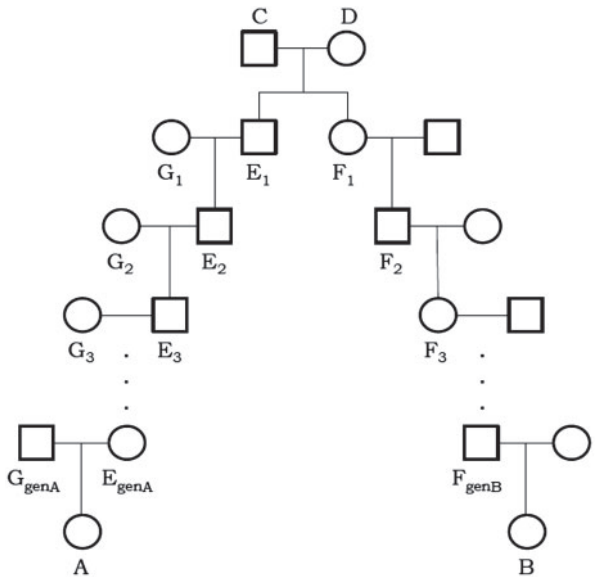

Fig. 1. Pedigree for a pair of individuals with two common ancestors: individuals A and B share two MRCAs, C and D. There are gen_A generations between the MRCAs and A (i.e. gen_A+1 meioses separating them) and gen_B generations between the MRCAs and B. The sex of the individuals is arbitrary.

We defined a set of HMMs for each of three relationship categories similarly to Stankovich *et al.* (2005) and Bercovici *et al.* (2010), who defined HMMs for cousins parameterized by the number of generations between them. Unlike these methods, the state space of our models does not increase with the number of generations of the pedigree. Below, we describe our HMMs for the first type of relationships. The models for the other two cases are derived along similar lines. Given that A and B have two MRCAs, C and D (Fig. 1), the hidden state at SNP k depends on the following binary variables:

(1) $m_C(k)$ and $m_D(k)$ indicate whether C and D, respectively, passed the same allele to their immediate descendants E_1 and F_1. For example, if both E_1 and F_1 inherited the maternal allele of C at position k, then $m_C(k) = 1$. If E_1 received the maternal allele of C, and F_1 received the paternal allele of C, then $m_C(k) = 0$.

(2) $m_{E_1}(k)$ and $m_{F_1}(k)$ indicate whether E_1 and F_1 passed to E_2 and F_2, respectively, the allele of C and not the allele of D.

(3) $d_A(k)$ takes the value 0 if A inherited the allele that E_2 got from E_1 (which came from either C or D) and the value 1 otherwise. That is, $d_A(k) = 0$, if for all $i > 2$, E_i got from E_{i-1} the allele of E_{i-2} and not the allele of G_{i-2}. If $d_A(k) = 1$, we will say that there were off-chain donations in the lineage of A at position k. $d_B(k)$ is defined in an analogous way for the lineage of B.

(4) $p_A(k)$ indicates which of the alleles of A, $H_{A0}(k)$ or $H_{A1}(k)$, comes from E_{gen_A} and is used to capture phasing errors. $p_B(k)$ is defined in an analogous way.

Each of these variables refers to a different set of meioses in the pedigree, therefore they all evolve independently from each other. We thus model the process of generating the genotypes G_A and G_B as a factorial HMM with hidden state $s(k) = (m_C(k), m_D(k), m_{E_1}(k), m_{F_1}(k), d_A(k), d_B(k), p_A(k), p_B(k))$.

Table 1. Transition probabilities for the HMMs with two MRCAs.

Variable	$Pr(0\rightarrow 0)$	$Pr(1\rightarrow 0)$
m_C	$\theta_k^2+(1-\theta_k)^2$	$2\theta_k(1-\theta_k)$
m_{E_1}	$1-\theta_k$	θ_k
d_A	$(1-\theta_k)^{\text{gen}_A-1}$	$\left(\sum_{n=1}^{\text{gen}_A-1}\binom{\text{gen}_A-1}{n}\theta_k^n(1-\theta_k)^{\text{gen}_A-1-n}\right)/\left(2^{\text{gen}_A-1}-1\right)$
p_A	$1-\omega$	ω

$P(i\rightarrow j)$ is the probability that a variable transitions from state i at SNP k to state j at SNP $k+1$, gen_A and gen_B are the generations between the MRCAs and each of A and B (Fig. 1), θ_k is the recombination probability between SNPs k and $k+1$; and ω is the probability of a phasing error. The transition probabilities for the variables m_D, m_{F_1}, p_B and d_B are derived similarly.

The transition probabilities for all the variables are shown in Table 1. We now derive the transition probabilities for d_A. Let gen_A be the number of generations between the MRCAs and A (Fig. 1). The number of meioses between E_2 and A is gen_A-1. Assume that $d_A(k)=0$, that is, the allele that E_2 inherited from the MRCAs at locus k was passed down to A. Any recombination between loci k and $k+1$ would result in $d_A(k+1)=1$, therefore $P(d_A(k+1)=0|d_A(k)=0)=(1-\theta_k)^{\text{gen}_A-1}$. If $d_A(k)=1$, then there exists at least one off-chain donation in the gen_A-1 meioses between E_2 and A. The probability that there are exactly n off-chain donations between E_2 and A is $\binom{\text{gen}_A-1}{n}(1/2)^{\text{gen}_A-1}$. Given that there are exactly n off-chain donations at SNP k, the probability that there are no off-chain donations at SNP $k+1$ is $\theta_k^n(1-\theta_k)^{\text{gen}_A-1-n}$. Therefore:

$$P(d_A(k+1)=0|d_A(k)=1)=\frac{P(d_A(k+1)=0 \text{ and } d_A(k)=1)}{P(d_A(k)=1)}$$

$$=\frac{\sum_{n=1}^{\text{gen}_A-1}\theta_k^n(1-\theta_k)^{\text{gen}_A-1-n}\binom{\text{gen}_A-1}{n}(1/2)^{\text{gen}_A-1}}{1-(1/2)^{\text{gen}_A-1}}$$

$$=\frac{1}{2^{\text{gen}_A-1}-1}\sum_{n=1}^{\text{gen}_A-1}\binom{\text{gen}_A-1}{n}\theta_k^n(1-\theta_k)^{\text{gen}_A-1-n}$$

Given $s(k)$, we can determine the IBD status at SNP k and use population allele frequencies to compute the emission probabilities (Epstein *et al.*, 2000). To account for genotyping errors, let ϵ be the probability of a genotyping error, and $f_\epsilon(x,y)$ be the probability that allele x is genotyped as y:

$$f_\epsilon(x,y)=\begin{cases}1-\epsilon & \text{if } x=y\\ \epsilon & \text{if } x\neq y\end{cases}$$

Then:

$$P(H_{A0}(k)=a,H_{B0}(k)=b|H_{A0} \text{ and } H_{B0} \text{ are IBD})=$$

$$\sum_{c\in\{A,C,G,T\}} q_c f_\epsilon(c,a) f_\epsilon(c,b)$$

where q_c is the frequency of allele c in the reference population. The rest of the emission probabilities are adjusted in a similar way.

2.4 Relationship notation

We refer to relationship types using the notation $(\text{mrcas},\text{gen}_A,\text{gen}_B)$. The variable mrcas is 2 when individuals A and B share two MRCAs and 1 otherwise. Unless A is the ancestor of B or vice versa, gen_A and gen_B are the number of generations between the MRCA(s) and A and B, respectively. If A is the ancestor of B, then gen_A is set to -1, and gen_B is the number of generations between A and B. For close relationships, we prefer to use the usual verbal description, unless space is limited. Table 2 shows the numerical notation for some common relationships.

Table 2. Numerical notation for some common relationships

Degree	Relationship	$(\text{mrcas},\text{gen}_A,\text{gen}_B)$
1	Full siblings	(2, 0, 0)
	Parent–child	(1, −1, 0)
2	Half siblings	(1, 0, 0)
	Aunt-niece	(2, 0, 1)
	Avuncular	(2, 0, 1) or (2, 1, 0)
	Grandparent–grandchild	(1, −1, 1)
3	First cousins	(2, 1, 1)
	Great grandparent–grandchild	(1, −1, 2)
	Great aunt–niece	(2, 0, 2)
	Half aunt–niece	(1, 0, 1)
4	Half first cousins	(1, 1, 1)

For all relationships with $\text{gen}_A\neq\text{gen}_B$, there is a corresponding symmetric relationship, for example (1, 0, −1) denotes a child–parent pair. Note that the term 'avuncular' does not specify a direction. The terms 'aunt' and 'niece' should be read as 'aunt/uncle' and 'niece/nephew', respectively.

2.5 Incorporating linkage information

HMMs that use unlinked markers have limited power to distinguish between relationships of the same degree (Sun *et al.*, 2002). Linkage information can help disambiguate such relationships. Assume, for instance, that we want to determine whether the relationship between individuals A and B is first cousins or great aunt–niece. An IBD block between A and B implies that they inherited overlapping genomic segments from their MRCAs (Fig. 2). If A and B are first cousins, the two scenarios of Figure 2 are equally likely. However, if A is closer to the MRCAs than B, then it is more likely that A inherited a larger segment from the common ancestor than B [scenario (a)], because we expect fewer recombinations between the MRCAs and A than between the MRCAs and B. Therefore, if we compare the haplotypes of A and B in a small window around the IBD transitions to the haplotypes of a reference population, we are more likely to find a match for the haplotype of A, than for the haplotype of B.

To quantify this intuition, assume that there is a transition in IBD status between SNPs k and $k+1$, and let $H_i(k-w+1..k+w)$ be the haplotype of individual i at positions $k-w+1,k-w+2,\ldots,k+w$, that is in a *window* of size $2w$ around k. The *haplo-frequency* of A is defined as:

$$C_A=\frac{\sum_i f_\epsilon(H_i(k-w+1..k+w),H_A(k-w+1..k+w))}{N} \quad (1)$$

where the sum is over all haplotypes in the reference population, N is the number of such haplotypes, and $f_\epsilon(H_i(k-w+1..k+w),H_A(k-w+1..k+w))$ is the probability that $H_i(k-w+1..k+w)$ is genotyped as

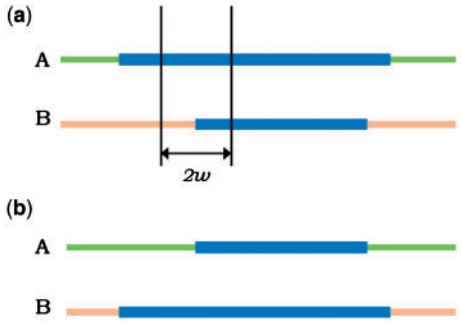

Fig. 2. Illustration of the IBD sharing between a pair of relatives: the horizontal lines represent two haplotypes of individuals A and B and the bold blue box is inherited from an MRCA. The overlapping part of the blue haplotypes is IBD in the two individuals. The first scenario is more likely if A is closer to the MRCA than B.

$H_A(k-w+1..k+w)$:

$$f_\epsilon(H_i(k-w+1..k+w), H_A(k-w+1..k+w)) = \prod_{j=k-w+1}^{k+w} f_\epsilon(H_i(j), H_A(j))$$

The positions of transitions in IBD status are estimated from the posterior probabilities of IBD of the HMMs using two cutoffs, c_{ibd} and $c_{\text{non-ibd}}$. Positions with posterior probability for IBD larger than c_{ibd} are called IBD, while positions with posterior probability lower than $c_{\text{non-ibd}}$ are called non-IBD. If position k is IBD, position ℓ is non-IBD, and positions $k+1 \ldots \ell-1$ have IBD probabilities between c_{ibd} and $c_{\text{non-ibd}}$, then we consider $k-w+1..\ell+w-1$ as the transition window.

2.6 Classifying relationships

Existing methods for relationship inference (Epstein *et al*., 2000; McPeek and Sun, 2000; Sun *et al*., 2002) use HMMs to compute the likelihood of the observed genotypes under a set of alternative relationships and then select the relationship that maximizes that likelihood. As explained in Section 2.5 and demonstrated in the Section 3, the likelihoods assuming unlinked markers have limited power to distinguish between relationships of the same degree. To overcome this problem, we used simulated data to train a multiclass discriminant analysis model with the following features:

(1) the likelihoods of all the alternative models;

(2) the total number of transitions between IBD and non-IBD;

(3) the estimate of the IBD sharing, $\left(\sum_k 0.5p_1(k)+p_2(k)\right)/M$, where $p_1(k)$ and $p_2(k)$ are the posterior probabilities of sharing one and two alleles IBD at SNP k, respectively, and the sum is taken over all M SNPs; and

(4) the percentage r of transitions between IBD and non-IBD in which $C_A < C_B$, where C_A and C_B are the *haplo-frequencies* of A and B, respectively [Equation 1]. As explained in Section 2.5, when A and B are equally distant from their common ancestors, r is expected to be 0.5. A value of r much larger than 0.5 indicates that C_A tends to be smaller than C_B, so A is more distant from the common ancestors than B. Similarly, a value of r much smaller than 0.5 indicates that B is more distant from the common ancestors than A.

2.7 Simulations

Phased genotypes of 165 individuals from UT, USA, with ancestry from northern and western Europe (CEU) and 88 Tuscans (TSI) were obtained from release 2 of Phase III of the HapMap Project. Some individuals in the CEU population were connected in parent–child pairs. We removed the offspring of all these pairs and combined the remaining 113 CEU individuals with the TSI individuals. After removing SNPs with minor allele frequency smaller than 1% in the combined CEU-TSI population, we obtained a set of 1 133 686 SNPs, which we call the *linked* set of SNPs. SNPs in the *linked* set that were in linkage disequilibrium (LD) were removed using the variance inflation factor method implemented in PLINK, version 1.07, (Purcell *et al*., 2007) to obtain an *unlinked* set of 79 681 SNPs. The HMMs were run on the *unlinked* set, while the *haplo-frequencies* were computed from the *linked* set.

The 201 individuals were divided randomly into two subgroups, T_{sim}, containing 57 CEU and 44 TSI individuals and T_{ref}, containing 56 CEU and 44 TSI individuals. T_{sim} was used to simulate pedigrees, while T_{ref} was used as a reference population, to compute allele and haplotype frequencies. Only autosomes were simulated. Following Haldane's model of recombination (Haldane, 1919), we set the recombination probability between SNPs k and $k+1$ to $\left(1-e^{-2g_k}\right)/2$, where g_k is the genetic distance between the SNPs. Genetic maps were obtained from the HapMap Project. The genotyping error was set to 0.9%. The window w for the computation of the *haplo-frequencies* was set to 20, and the cutoffs, c_{ibd} and $c_{\text{non-ibd}}$, were set to 0.9 and 0.2, respectively. These values were chosen based on the accuracy of capturing the IBD transitions on a set of simulated individuals distinct from the training and testing sets.

2.8 Comparison with RELPAIR and PREST-plus

We compared our method to RELPAIR, version 2.0.1 (Epstein *et al*., 2000) and PREST-plus, version 4.09 (McPeek and Sun, 2000; Sun *et al*., 2002) on simulated individuals created as described in Section 2.7. PREST-plus was run on the same 79 681 unlinked SNPs that was used to run CARROT's HMMs. RELPAIR cannot handle more than 9999 markers, so, similarly to Pemberton *et al*. (2010), we ran it on 5 random subsets of the unlinked SNPs of size 9961 SNPs each. Each subset was created by dividing the 79 681 SNPs into non-overlapping windows of size 8, and then randomly selecting one SNP from each window. The prediction of RELPAIR for a given pair of individuals was set to the relationship predicted in the majority of the five runs.

2.9 Inference of relationships in Phase III of HapMap

Phased genotypes of 83 individuals of African ancestry from the southwestern USA (ASW), 165 CEU individuals, 171 individuals of Mexican ancestry from Los Angeles, CA, USA (MXL) and 167 Yoruba individuals from Ibadan, Nigeria (YRI) were obtained from release 2 of Phase III of HapMap. We inferred relationships in each population separately, since, according to Pemberton *et al*. (2010) there is no evidence for population-labeling errors. For each population, we removed SNPs with minor allele frequency smaller than 1%, SNPs in LD, as well as SNPs that failed the hypothesis test of Hardy–Weinberg equilibrium [$P < 10^{-4}$, computed by PLINK, version 1.07 (Purcell *et al*., 2007)], and created *linked* and *unlinked* sets as described in Section 2.7.

To reduce computation time, we first obtained a rough estimate of the IBD sharing by running the HMM for full siblings for all pairs in each population. Pairs with predicted IBD percentage less than 3% were removed from further consideration. This threshold was selected to exclude most of the fifth-degree relatives, since the accuracy of our method is small in such cases. It is possible that some of the fourth-degree relatives were also excluded during this process. We note, however, that the goal of this analysis was not to exhaustively identify all the relationships in the populations studied, but rather to demonstrate the applicability of our method on a real dataset.

For the pairs that passed the above cutoff, we ran our HMMs for all relationships of degree up to 5, using previously reported unrelated individuals of the corresponding population to compute allele and haplotype frequencies. The phasing error parameter ω, was set to 0.1%. To infer the relationships of these pairs, we trained our classifiers on 100 pairs

of individuals for each relationship of degree up to 5 and 100 unrelated individuals, simulated as described in Section 2.7.

We verified our predictions for the first- and second-degree relationships by comparing them with the relationships reported in HapMap and the predictions of Pemberton *et al*. (2010). To verify our predictions for relationships of higher degree, we detected sets of three or more individuals that were all predicted to be related to each other and examined whether our predictions for these individuals were mutually consistent. We were very conservative in our verification process, reporting newly discovered third- and fourth-degree relationships as verified only if they had strong support from relationships of smaller degrees.

Pedigrees were drawn using HaploPainter, version 1.043 (Thiele and Nürnberg, 2005).

3 RESULTS

3.1 Results on simulated data

Comparison with maximum-likelihood methods: we compared CARROT with two existing methods for relationship inference, RELPAIR (Epstein *et al*., 2000) and PREST-plus (McPeek and Sun, 2000; Sun *et al*., 2002). Both these methods use HMMs similar to those defined in Section 2.3 to compute the likelihoods of a set of alternative relationships and select the relationship within the set with the maximum likelihood. RELPAIR only considers the relationships such as full siblings, parent–child, avuncular, grandparent–grandchild, half siblings, first cousins, monozygotic twins and unrelated. PREST-plus considers the relationships examined by RELPAIR as well as the relationships half avuncular, half first cousins and half-siblings-first-cousins [see Sun *et al*. (2002) for the definition of that relationship]. Neither RELPAIR nor PREST-plus differentiate between *rotated* relationships, so for example they do not distinguish between a grandparent–grandchild pair and a grandchild–grandparent pair.

We first compared CARROT, RELPAIR and PREST-plus on the set of relationships that can be handled by RELPAIR, except twins, which are not considered by CARROT. We simulated 100 pairs of phased individuals for each of the relationships full siblings, parent–child, aunt/uncle–niece/nephew, grandparent–grandchild, half siblings, first cousins and unrelated. We ran RELPAIR and PREST-plus on all pairs using the seven aforementioned alternative relationships. The performance of CARROT was assessed using 10-fold cross-validation: The simulated pairs were divided into 10 subsets each containing 10 pairs from each relationship. CARROT was trained on 9 of the subsets, that is, on 90 pairs from each relationship, and tested on the remaining subset. This process was repeated 10 times and the accuracy was averaged over the 10 runs.

Figure 3 shows the number of pairs that were predicted *incorrectly* by each method. Although all three methods achieved excellent accuracy for the first-degree relationships, CARROT clearly outperformed both RELPAIR and PREST-plus on second degree pairs, particularly on half siblings. RELPAIR and PREST-plus classified more than half (56 and 68%, respectively) of the half sibling pairs as avuncular. We note that the difference in performance between RELPAIR and PREST-plus can probably be attributed to the fact that RELPAIR was run on smaller sets of SNPs (see Section 2.8).

We also compared CARROT with PREST-plus on the set of nine relationships which includes the seven relationships above as well as half avuncular and half first cousins. As shown in Figure 4, CARROT outperformed PREST-plus for most relationships as well as on average. The only relationship for which CARROT performed worse than PREST-plus was first cousins: 13% of the first cousins pairs were classified as half avuncular by CARROT, probably because the *haplo-frequencies* of these two relationships are similar.

As mentioned above, RELPAIR and PREST-plus use HMMs to compute the likelihoods of a set of alternative relationships and select the relationship with the maximum likelihood. However,

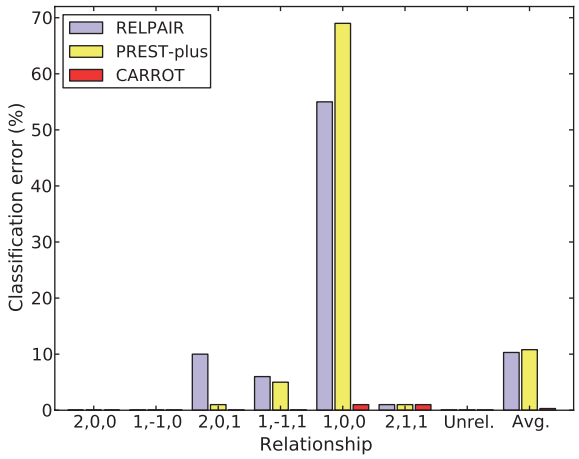

Fig. 3. Comparison between CARROT, RELPAIR and PREST-plus on a set of seven relationships: the height of the bars is the percentage of pairs that were classified *incorrectly*, so smaller bars are better. In each relationship, the leftmost bar corresponds to RELPAIR, the middle bar to PREST-plus and the rightmost bar to CARROT. Avg. is the average error over all relationships examined. (2, 0, 0): full siblings; (1, −1, 0): parent–child; (2, 0, 1): aunt-niece; (1, −1, 1): grandparent-grandchild; (1, 0, 0): half siblings; (2, 1, 1): first cousins, unrel: unrelated.

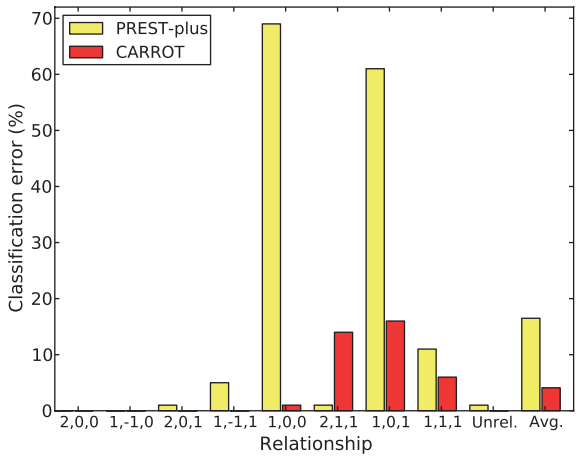

Fig. 4. Comparison between CARROT and PREST-plus on a set of nine relationships: the height of the bars is the percentage of pairs that were classified *incorrectly*, so smaller bars are better. In each relationship, the left bar corresponds to PREST-plus and the right bar to CARROT. Avg. is the average error over all relationships examined. (2, 0, 0): full siblings; (1, −1, 0): parent–child; (2, 0, 1): aunt–niece; (1, −1, 1): grandparent–grandchild; (1, 0, 0): half siblings; (2, 1, 1): first cousins; (1, 0, 1): half aunt–niece; (1, 1, 1): half first cousins, unrel: unrelated.

Table 3. Comparison between CARROT and two other approaches for relationship inference

Method/Degree	1	2	3	4	5
Max likelihood	100	98.6	62.29	39.33	26.73
CARROT likelihoods	100	100	74	43.78	29
CARROT	100	100	**84**	**57.56**	**34.45**

max likelihood selects the relationship with the maximum likelihood; *CARROT likelihoods* uses only the likelihoods as features for the classification. We simulated 100 pairs of individuals from each relationship and performed predictions only within each degree. The number reported is the percentage of pairs classified in the correct relationship.

direct comparison between these methods and CARROT can only be done for a small set of relationships. To extend this comparison to additional relationships, we implemented a classifier that uses the HMMs of Section 2.3 for the likelihood computation and then selects the relationship with the maximum likelihood. This classifier can serve as a proxy for any method that maximizes the likelihood of unlinked markers. We compared the maximum-likelihood classifier with CARROT using a set of 100 simulated pairs of phased individuals for each relationship of degree up to five, including all *rotated* relationships. We assumed that the degree of the relationship was known, so predictions were made only within each degree. To assess whether a classification-based approach is better that a maximum-likelihood approach, we also ran CARROT using only the likelihoods as features. We observed that for relationships of degree up to two, the likelihoods are sufficient to differentiate between relationships (Table 3). For higher degrees, the likelihoods become less informative and the additional features of CARROT result in a significant increase in accuracy. Additionally, we notice that CARROT performs consistently better than the maximum-likelihood approach, even when we only use the likelihoods as features. Intuitively, the classifier can capture correlations between the likelihoods of different relationships. We observed, for instance, that when the true relationship is great grandparent–grandchild, the likelihood of the relationship great aunt/niece tends to be increased, but this effect is overlooked when we use the maximum likelihood criterion.

Differentiating between rotated relationships: to evaluate the ability of CARROT to distinguish between *rotations* of relationships, we simulated 100 pairs of individuals for each of the possible relationships of degree up to five, including all possible *rotated* relationships. We first assumed that the degree of each relationship was known, and ran CARROT separately for each degree. We started by examining the ideal case of perfect phasing, so we set the probability of phasing errors, ω, to zero.

We assessed CARROT's accuracy using 10-fold cross-validation, as described in the previous section: the simulated pairs were divided into 10 subsets each containing 10 pairs from each relationship. CARROT was trained on nine of the subsets and tested on the remaining subset. This process was repeated 10 times and the accuracy was averaged over all 10 runs. We defined the prediction accuracy as the number of pairs that were classified in the correct relationship.

Table 4. Classification accuracy of CARROT on third-degree relatives

Rel.	2,0,2	2,1,1	2,2,0	1,−1,2	1,0,1	1,1,0	1,2,−1
2,0,2	**88**	–	–	5	5	2	–
2,1,1	–	**84**	1	–	7	8	–
2,2,0	–	1	**88**	–	2	4	5
1,−1,2	2	–	–	**96**	1	1	–
1,0,1	–	11	–	–	**68**	21	–
1,1,0	–	8	–	–	24	**68**	–
1,2,−1	–	–	3	–	1	1	**95**

The value at row i and column j is the percentage of pairs of relationship i that were predicted to be of relationship j. (2, 0, 2): great aunt–niece; (2, 1, 1): first cousins; (2, 2, 0): great niece–aunt; (1, −1, 2): great grandparent–grandchild, (1, 0, 1): half aunt–niece; (1, 1, 0): great niece–aunt; (1, 2, −1): great grandchild–grandparent.

When run on first- and second-degree relatives, CARROT achieved perfect performance. The results for the third- and fourth-degree relationships are summarized in Tables 4 and 5, respectively. The average accuracy over all the pairs of the corresponding degree, was 83.86 and 56.89%, respectively. For the fifth-degree relationships, the average accuracy was 34.45% (full results not shown). As expected, the accuracy of our classifiers drops as the degree of the relationship increases. However, even within the same degree, some relationships are much harder to predict correctly than others. For example, the two half-avuncular relationships, (1, 0, 1) and (1, 1, 0), are hard to differentiate from each other and from first cousins, since the difference in the distance of each of the individuals from their MRCA is not enough for the *haplo-frequencies* to distinguish them from a balanced relationship where both individuals are equally distant from the MRCAs. Similarly, although the average accuracy for the fifth-degree relationships was 34.45%, the (2, 4, 0) and (2, 0, 4) pairs were predicted correctly in 48.5% of the cases.

Predictions across degrees: since in practice the degree of the relationship is not necessarily known, we also performed cross validation on a set of 200 pairs of individuals from each of the relationships of degree up to 5, including *rotated* relationships, as well as 200 pairs of unrelated individuals. The average classification accuracy was 57.5%, varying widely for different degrees: all the first-degree pairs were classified correctly; the average accuracy for the second-degree pairs was 99.5%, for the third-degree pairs 76.57%, for the fourth-degree pairs 46% and for the fifth-degree pairs 23.36%. Finally, 90% of the unrelated individuals were classified correctly. We note that there was a small decrease in accuracy compared with the results of the previous section, because some of the pairs were classified in relationships of the incorrect degree, while this was never the case when only within-degree predictions were made. Table 6 shows the percentage of pairs that were classified in a relationship of the correct degree for each of the degrees examined. On average, the correct degree was predicted for 89.83% of the pairs.

The effect of phasing errors: the phasing error rate is defined as the proportion of successive pairs of heterozygote SNPs that are phased incorrectly with respect to each other. To examine the effect of phasing errors on the classification accuracy of CARROT, we simulated 100 pairs of individuals for each of the third-degree

Table 5. Classification accuracy of CARROT on fourth-degree relatives

Rel.	2,0,3	2,1,2	2,2,1	2,3,0	1,−1,3	1,0,2	1,1,1	1,2,0	1,3,−1
2,0,3	**76**	4	1	–	8	8	3	–	–
2,1,2	1	**42**	24	–	–	15	11	7	–
2,2,1	–	24	**43**	–	–	5	13	15	–
2,3,0	–	–	4	**73**	–	1	3	10	9
1,−1,3	4	1	–	–	**80**	13	–	2	–
1,0,2	–	13	9	–	–	**52**	14	12	–
1,1,1	–	8	19	–	–	28	**14**	31	–
1,2,0	–	5	13	–	–	13	17	**52**	–
1,3,−1	–	–	1	4	–	2	1	12	**80**

The value at row *i* and column *j* is the percentage of pairs of relationship *i* that were predicted to be of relationship *j*.

Table 6. Ability of CARROT to predict the degree of a relationship

Degree	1	2	3	4	5	Unrel
1	**100**	–	–	–	–	–
2	–	**99.5**	0.5	–	–	–
3	–	0.57	**93.57**	5.71	0.14	–
4	–	–	3.94	**82.06**	14	–
5	–	–	–	12.41	**86.64**	0.95
Unrel	–	–	–	–	10	**90**

The value in row *i* and column *j* is the percentage of pairs of degree *i* that were classified in a relationship of degree *j*.

relationships, introduced phasing errors at various rates in the range 0–1% and performed 10-fold cross-validation, as described above. We observed that, as the phasing error increased from 0% to 1%, there was an almost linear decrease in classification accuracy from 84% to 67.57%, although for phasing errors in the range 0.0–0.1%, the accuracy was almost unaffected. Current algorithms for statistical phasing of genotype data, such as BEAGLE (Browning and Browning, 2007), have phasing error rates ranging between 0.05% and 6%, depending on the number of genotyped individuals and markers. Recently, however, Fan *et al*. (2011) and Yang *et al*. (2011) proposed new experimental techniques for whole-chromosome haplotyping, which promise to achieve much smaller error rates than current methods. Given these recent developments in phasing techniques, we believe that the phasing error requirements of our algorithm are realistic.

3.2 Inference of relationships in Phase III of HapMap

Phase III of the HapMap dataset contains genotypes of 1184 individuals from 11 populations in more than a million SNPs. To facilitate phasing, many of these individuals are connected in trios (two parents and an offspring) or duos (a parent and an offspring). Recently, however, Pemberton *et al*. (2010) discovered many additional first- and second-degree relationships in the HapMap collection. We applied CARROT on four of the 11 HapMap populations: 83 individuals of African ancestry from the southwestern USA (ASW), 165 individuals from UT, USA, with ancestry from northern and western Europe (CEU), 171 individuals of Mexican ancestry from Los Angeles, CA, USA (MXL) and 167 Yoruba individuals from Ibadan, Nigeria (YRI). These populations were selected because they contained both previously reported first-degree relationships and additional relationships detected by Pemberton *et al*. (2010). Since Pemberton *et al*. (2010) did not report any relationships of degree greater than two, to verify the novel predictions of CARROT, we identified sets of three or more related individuals and examined whether our predictions for these individuals were consistent with each other and with previously reported relationships.

Table 7 summarizes our findings. For each population, we report:

(1) the number of predictions that agreed with previously known relationships;

(2) the number of identified pairs for which the relationship predicted by CARROT was inconsistent with previously known relationships but the degree of the relationship was in agreement with previously known relationships;

(3) the number of identified pairs for which the degree of the predicted relationship was inconsistent with previously known relationships; and

(4) the number of identified pairs for which we could not determine whether CARROT's prediction was correct or not based on previously known relationships.

We note that there are small discrepancies between the numbers of first- and second-degree relationships that we report here and the numbers reported by Pemberton *et al*. (2010). These discrepancies can be partially explained by the fact that Pemberton *et al*. (2010) used release 3 of Phase III of HapMap, while we used release 2. Additionally, unlike Pemberton *et al*. (2010), we did not include in our analysis individuals who failed the quality control filters during phasing.

In total, we identified 23 previously unreported third-degree pairs and 44 previously unreported fourth-degree pairs. Figure 5 shows one of the pedigrees we detected. This pedigree involves three families from the MXL population, M008, M010 and M012. All the depicted first-degree relationships as well as the aunt–nephew relationship between the individuals NA19660 and NA19664 have been reported previously. However, our algorithm correctly identified the first cousin pairs NA19664–NA19685 and NA19664–NA19662, therefore CARROT reconstructs this pedigree even in the absence of genotype data for the mother of NA19685 and NA19662. A list of all the previously unreported relative pairs detected by CARROT is given in the supplement.

Table 7. Predictions of CARROT on four HapMap populations

Pop	Degree	Reported relatives			New relatives				Total
		Correct	Degree only	Incorrect	Correct	Unverified	Degree only	Incorrect	
CEU	1	97	–	–	–	–	–	–	97
	2	2	–	–	–	–	–	–	2
	3	–	–	–	2	–	–	–	2
	4	–	–	–	1	13	1	–	15
	Total	99	–	–	3	13	1	–	116
MXL	1	56	–	–	–	–	–	–	56
	2	7	–	–	–	–	–	–	7
	3	–	–	–	2	–	1	–	3
	4	–	–	–	2	–	–	1	3
	Total	63	–	–	4	–	1	1	69
YRI	1	105	–	–	–	–	–	–	105
	2	2	1	–	–	–	–	–	3
	3	–	–	–	1	3	1	1	6
	4	–	–	–	1	3	–	–	4
	Total	107	1	–	2	6	1	1	118
ASW	1	46	–	–	–	–	–	–	46
	2	7	3	3	–	–	1	–	14
	3	–	–	–	7	4	1	–	12
	4	–	–	–	1	21	–	–	22
	Total	53	3	3	8	25	2	–	94

The results are grouped per population and relationship degree. New Relatives are relative pairs identified by CARROT that have not been previously reported. Correct predictions are predictions that agree with previously reported relationships. The column 'Degree Only' refers to cases where CARROT predicted correctly only the degree of the relationship. 'Unverified' relationships are those for which we could not determine whether CARROT's prediction was correct or not.

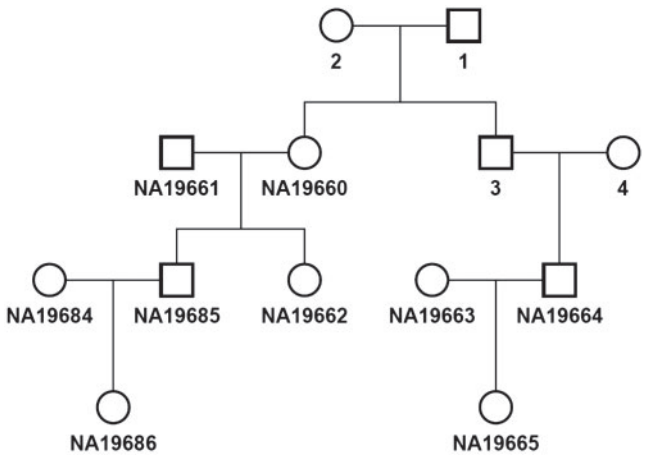

Fig. 5. Example of a pedigree identified by CARROT: the pedigree involves three families from the MXL population, M008, M010 and M012.

We note that, although for the CEU and MXL populations, the accuracy of our algorithm is consistent with the simulation results of Section 2.7, in the other two populations there was a drop in accuracy. This difference in performance might be a result of an increased phasing error in the YRI and ASW populations. Additionally, in the ASW population, the degree of relatedness is unexpectedly high, with 94 identified relationships in a sample of 83 individuals. We observed, for example, that 14 individuals from eight HapMap families (2483, 2487, 2488, 2489, 2490, 2492, 2495, and 2496) are connected to each other. With less than 50 presumably unrelated individuals [there are 53 individuals reported as unrelated in the HapMap ASW dataset, even if we ignore the relationships reported here and in Pemberton *et al*. (2010)], the estimation of allele and haplotype frequencies becomes problematic, especially in an admixed population like ASW.

To help determine whether the previously unreported pairs detected by CARROT are predicted correctly, we ran KING (Manichaikul *et al*., 2010) on all the identified third-degree relatives. KING uses identity by state to distinguish relatives of degree up to three from unrelated individuals. Six out of the seven pairs of individuals that were predicted to be third-degree relatives by CARROT and could not be otherwise verified, were also predicted to be third-degree relatives by KING. We thus concluded that these six pairs are most likely true third-degree relatives. We notice that KING correctly predicted only 13 out of the 16 verified third-degree relatives. It is therefore likely that the last unverified third-degree prediction for which there was a disagreement between CARROT and KING is also a true third-degree pair.

4 DISCUSSION

We presented CARROT, a novel framework for relationship inference that uses linkage information to infer more distant relationships than existing methods and to distinguish between *rotated* relationships, that is, relationships with the same number of common ancestors and the same number of meioses separating the individuals under consideration (e.g. aunt–niece versus niece–aunt or first cousins versus great aunt–niece). We demonstrated that CARROT achieved superior accuracy on relationships of degree up to four, clearly outperforming previous methods, such as RELPAIR

and PREST-plus on simulated data. We also applied CARROT on data from four HapMap populations, ASW, CEU, MXL and YRI, and correctly identified the vast majority of the first- and second-degree relatives recently detected by Pemberton *et al*. (2010). Additionally, CARROT detected 67 previously unreported third- and fourth-degree relative pairs.

A possible shortcoming of CARROT stems from its current reliance on phased data. However, state of the art statistical methods for phasing achieve phasing errors smaller than 1% (Browning and Browning, 2007; Howie *et al*., 2009) when large cohorts and dense SNP panels are used. Additionally, novel experimental phasing techniques have achieved haplotyping of whole chromosomes with accuracy of 99.8% (Fan *et al*., 2011; Yang *et al*., 2011).

Methods for detecting genealogical relationships and for disambiguating between *rotated* relationships can facilitate pedigree reconstruction from cohorts such as the HapMap and WTCCC (Wellcome Trust Case Control Consortium, 2007) datasets, where most individuals are not expected to be closely related. Existing algorithms for pedigree reconstruction that are based on the detection of parent–offspring and sibling pairs cannot be applied in such datasets. In studies such as WTCCC, that are based on the premise that the individuals in the dataset are unrelated, putative relatives are usually removed or reweighted. Knowledge of the exact pedigrees, however, allows a more informed selection of the individuals to be removed. More importantly, instead of discarding the information captured in related individuals, one can leverage the reconstructed relationships in a meta-analysis step. Combining the linkage analysis from multiple reconstructed partial pedigrees with the association analysis over the remaining unrelated individuals can potentially increase the power of such studies. CARROT was designed with such applications in mind and aims at bridging the gap between algorithms for pedigree reconstruction from close relatives and traditional relationship inference methods.

ACKNOWLEDGEMENTS

We would like to thank M. Schaub for helpful discussions, as well as the two anonymous reviewers for their comments.

Funding: National Institutes of Health (HG005570-01 to S.K.-P., HG005596-01 to D.K.H., HG005570-01 to S.B.); the National Science Foundation (DBI-0640211-002 to S.B.); a William R. Hewlett Stanford Graduate Fellowship (to S.J.A.); National Science Foundation Fellowship (to S.J.A.).

Conflict of Interest: none declared.

REFERENCES

Bercovici,S. *et al*. (2010) Estimating genome-wide IBD sharing from SNP data via an efficient hidden Markov model of LD with application to gene mapping. *Bioinformatics* **26**, i175–i182.

Berger-Wolf,T. *et al*. (2007) Reconstructing sibling relationships in wild populations. *Bioinformatics*, **23**, 49–56.

Blouin,M.S. (2003) DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends Ecol. Evol.*, **18**, 503–511.

Browning,S.R. and Browning,B.L. (2007) Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. *Am. J. Hum. Genet.*, **81**, 1084–1097.

Epstein,M.P. *et al*. (2000) Improved inference of relationship for pairs of individuals. *Am. J. Hum. Genet.*, **67**, 1219–1231.

Fan,H.C. *et al*. (2011) Whole-genome molecular haplotyping of single cells. *Nat. Biotechnol.*, **29**, 51–57.

Ghahramani,Z. and Jordan,M. (1997) Factorial hidden Markov models. *Mach. Learn.*, **29**, 245–273.

Haldane,J.B.S. (1919) The combination of linkage values, and the calculation of distances between the loci of linked factors. *J. Genet.*, **8**, 299–309.

Howie,B.N. *et al*. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.*, **5**, e1000529.

The International HapMap 3 Consortium (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**, 52–58.

Jones,A.G. and Arden,W.R. (2003) Methods of parentage analysis in natural populations. *Mol. Ecol.*, **12**, 2511–2523.

Koch,M. *et al*. (2008) Pedigree reconstruction in wild cichlid fish populations. *Mol. Ecol.*, **17**, 4500–4511.

Manichaikul,A. *et al*. (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics*, **26**, 2867–2873.

McPeek,M.S. and Sun,L. (2000) Statistical tests for detection of misspecified relationships by use of genome-screen data. *Am. J. Hum. Genet.*, **66**, 1076–1094.

Pemberton,T.J. *et al*. (2010) Inference of Unexpected Genetic Relatedness among Individuals in HapMap Phase III . *Am. J. Hum. Genet.*, **87**, 457–464.

Purcell,S. *et al*. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

Rabiner,L.R. and Juang,B.H. (1986) An introduction to hidden Markov models. *IEEE Acoust. Speech. Sign. Process. Mag.*, **3**, 4–16.

Riester,M. *et al*. (2009) FRANz: reconstruction of wild multi-generation pedigrees. *Bioinformatics*, **25**, 2134–2139.

Riester,M. *et al*. (2010) Reconstruction of pedigrees in clonal plant populations. *Theor. Popul. Biol.*, **78**, 109–117.

Skare,Ø. *et al*. (2009) Identification of distant family relationships. *Bioinformatics*, **25**, 2376–2382.

Stankovich,J. *et al*. (2005) Identifying nineteenth century genealogical links from genotypes. *Hum. Genet.*, **117**, 188–199.

Sun,L. *et al*. (2002) Enhanced pedigree error detection. *Hum. Hered.*, **54**, 99–110.

Thiele,H. and Nürnberg,P. (2005) HaploPainter: a tool for drawing pedigrees with complex haplotypes. *Bioinformatics*, **21**, 1730–1732.

Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature*, **447**, 661–678.

Yang,H. *et al*. (2011) Completely phased genome sequencing through chromosome sorting. *Proc. Natl Acad. Sci. USA*, **108**, 12–17.