SplittingHeirs: Inferring Haplotypes by Optimizing Resultant Dense Graphs

Sharlee Climer
Department of Computer Science and Engineering
Washington University
Saint Louis, Missouri, USA
climer@wustl.edu

Alan R. Templeton
Department of Biology
Washington University
Saint Louis, Missouri, USA
temple_a@biology.wustl.edu

Weixiong Zhang
Department of Computer Science and Engineering
Washington University
Saint Louis, Missouri, USA
zhang@cse.wustl.edu

ABSTRACT

Phasing genotype data to identify the composite haplotype pairs is a widely-studied problem due to its value for understanding genetic contributions to diseases, population genetics research, and other significant endeavors. The accuracy of the phasing is crucial as identification of haplotypes is frequently the first step of expensive and vitally important studies. We present a combinatorial approach to this problem which we call SplittingHeirs. This approach is biologically motivated as it is based on three widely accepted principles: there tend to be relatively few unique haplotypes within a population, there tend to be clusters of haplotypes that are similar to each other, and some haplotypes are relatively common. We have tested SplittingHeirs, along with several popular existing phasing methods including PHASE, HAP, EM, and Pure Parsimony, on seven sets of haplotype data for which the true phase is known. Our method yields the highest accuracy obtainable by these methods in all cases. Furthermore, SplittingHeirs is robust and had higher accuracy than any of the other approaches for the two datasets with high recombination rates. The success of SplittingHeirs validates the assumptions made by the dense graph model and highlights the benefits of finding globally optimal solutions.

1. INTRODUCTION

We might not know each other, but we have a lot in common. We have DNA sequences that are roughly 99.9% identical [24, 45, 4, 19]. Our custom-blend of the remaining one-tenth of one percent is what makes us genetically unique. This variation, combined with environmental factors, is responsible for a plethora of characteristics that distinguish us from each other. Of particular importance is the fact that genetic variation coupled with environmental factors contribute to our vulnerability to, and convalescence from, virtually every known disease [8]. Furthermore, our variation affects our responses to treatments for these diseases.

The Human Genome Project [44, 41] has mapped the 99.9% of DNA that is common, and the International HapMap Consortium [43] is currently mapping the small portion with variation. About 90% of this variation is due to single nucleotide polymorphisms (SNPs) [23].[1]

Although the HapMap project is mapping a substantially smaller number of nucleotides, they have a daunting challenge that the Human Genome Project did not have to tackle. Being diploid, we have pairs of chromosomes and, consequently, two DNA sequences at most regions of the genome. Current sequencers produce a meld of the pair of DNA strands. This is not a problem when both sequences are the same. However, if they are different it is not clear how to split the meld, or *genotype*, into its two original components, or *haplotypes*.

The importance of inferring haplotypes from genotypes, and the value of accuracy in this inference, can hardly be overstated. Because genetic variation has an impact on virtually every known disease [8], researchers studying prevention from, and treatment of, diseases will likely need information about this variation. Many complex diseases of interest are believed to be associated with a combination of several genetic factors, coupled with environmental components. Errors in haplotype inference could severely confound efforts to solve these inherently complicated studies.

Haplotype inference is also of great interest in the field of population genetics as this field is the study of genetic variation of populations of plants and animals, including humans, and how that variation changes with time and space [37]. Understanding these changes is fundamental for developing policies to reduce species extinction and environmental deterioration (e.g. [38]). Errors in haplotype inference could compromise these policies.

In 2002, the International HapMap Consortium was founded to map human genome-wide variation [43]. Data have also been produced from numerous other studies, yielding rich and accurate genotype data for many populations. The question that remains is how haplotypes should be inferred from these data.

It is possible to find the two actual separate haplotypes by a variety of laboratory techniques [1, 29]. However, these approaches are infeasible for all but very small projects. For this reason, researchers depend on computational methods

ACM-BCB 2010 Niagara Falls, NY, USA

[1]The recent work of Redon et al. [31] may result in a reduction of this estimate.

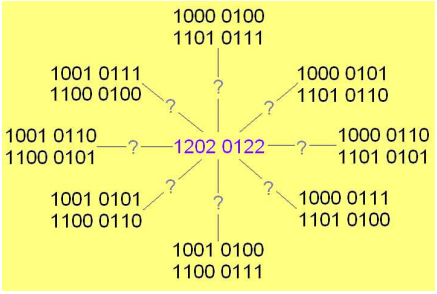

Figure 1: A genotype with four heterozygous sites is shown at the center of this figure. Surrounding it are the eight pairs of haplotypes that could resolve it.

to determine haplotype pairs from sequence data. For instance, the HapMap Consortium has used PHASE [36] to computationally infer haplotypes for their genotypes [42].

True haplotypes identified in laboratories are valuable for assessing the accuracy of computational methods. Recently, Andrés et al. [1] identified true haplotype pairs and tested the accuracy of PHASE [35, 36], fastPHASE [33], HAP [11, 18], and GERBIL [21]. The results were poor and the confidence levels computed were error prone. This finding suggests that the assumptions made by these models do not completely reflect properties that arise in nature.

1.1 The Haplotype Inference Problem

The term *haplotype* is sometimes used in reference to a "block" of SNPs where the boundaries of the block are defined by certain criteria [14]. Barnes suggested the use of a haplotype structure that has little or no recombination and a limited variation within a population [2]. In this paper, we use the term "haplotype" broadly to refer to a set of SNP nucleotide states in physical proximity to each other. This broad definition allows us to discuss sets of SNPs for which we have no prior information about recombination and/or frequency within a population.

For most SNPs, there exist only two different nucleotides in the population. These SNPs are referred to as *biallelic*. Most haplotype inference methods assume that the SNPs in the input are biallelic and we consider this simplest case. However, the general ideas presented in this paper can be extended to account for multiallelic data.

The haplotype inference problem can be abstracted into an elegant mathematical formulation. Assuming biallelic data, there are only two possible nucleotides, or *alleles*, at a given site in a haplotype. For this reason, a haplotype can be encoded into a binary string where '0' represents one allele and '1' represents the other. A genotype can be represented by a ternary string in which a '0' indicates that the corresponding alleles on both haplotypes are '0's, a '1' indicates that both are '1's, and a '2' indicates that one of each allele is present, i.e., it is a *heterozygous* site.

Let m represent the number of nucleotides in a genotype and let k represent the number of sites that are heterozygous. There exist 2^{k-1} pairs of haplotypes that can resolve a genotype with k heterozygous sites. Figure 1 depicts an example genotype.

Genotypes with $k < 2$ are unambiguous and can only be resolved by one pair of haplotypes. Given a genotype with $k \geq 2$ heterozygous sites, it is not clear what could make any one of the 2^{k-1} feasible solutions more likely than another to be the actual pair of haplotypes that exist in the individual. However, if genotypes are identified for a set of individuals from a population, there are reasons to believe that some solutions are more biologically sound than others.

The *haplotype inference* problem can be stated as follows: Given a set of n genotypes taken from a population, find the set of haplotypes that exist in the sample set and, for each genotype, find the pair of haplotypes that is most likely to exist in the given individual. This problem is also referred to as *phasing* the genotypes.

1.2 Previous Approaches

In this section, some of the previous approaches to haplotype inference are discussed and the biological assumptions upon which they are designed are summarized.

In 1990, Clark proposed the Subtraction Method for haplotype inference [5]. This method requires that at least one of the genotypes has $k < 2$, where k is the number of heterozygous sites. The resolving haplotype pairs for these genotypes are obvious and are the first to be placed in the pool of haplotypes that are inferred. A number of iterations follow in which genotypes are resolved if at least one of the haplotypes in the pool can be used to resolve the genotype. In some implementations, the algorithm is randomly restarted a large number of times and the "best" solution is selected [5], while others use a *consensus* system [29].

The main biological assumption used by the Subtraction Method is that the number of unique haplotypes in a given population is relatively small. However, this approximation method doesn't always return a solution with the fewest possible unique haplotypes.

Earl Hubbell and Dan Gusfield independently derived the Pure Parsimony model to address this issue (personal communication). Gusfield published a concise definition of this model as a Mixed Integer Linear Program (MIP) that can be computed using any generic MIP solver [16]. The solution to this MIP is guaranteed to have the minimum possible number of unique haplotypes that can resolve the given set of genotypes. The Pure Parsimony MIP is NP-hard [17]. However, efficient implementations have been made and it has been used to optimally solve moderate-sized instances [16, 3].

The only biological assumption made by the Pure Parsimony model is that the number of unique haplotypes in the population is the absolute minimum possible. Some other combinatorial methods use very different assumptions.

In 2002, Gusfield proposed the *Perfect Phylogeny* inference model [15]. This model infers haplotypes that, when combined with additional haplotypes, form a phylogenetic tree. A phylogenetic tree is a mapping of the genealogical heritage of a set of haplotypes to a common ancestor. It is assumed that there is no recombination in the data. Furthermore, the *infinite sites* model [22] is assumed. This assumption presupposes that at most only one mutation has ever occurred at any given site in the genome, and has been commonly assumed in the past. However, recent work has shown that recurrent mutations are common in the human genome *e.g.* [9].

The perfect phylogeny model recognizes that some haplotypes in a population are similar to others as they may be

mutations of the others. In this respect, it emphasizes the similarities between haplotypes without direct regard for the cardinality of the haplotype set. In summary, the main biological assumptions made by the Perfect Phylogeny model are that the haplotypes within a population can be organized into a phylogenetic tree, there is no recombination, and the infinite sites model is assumed to be valid.

Many sets of genotypes have no perfect phylogeny solution. For this reason, a number of algorithms have been developed using an imperfect phylogeny, *e.g.* [11, 34, 13, 18]. For example, HAP [11, 18] finds a perfect phylogeny for a subset of the genotypes and then resolves the remaining genotypes by using haplotypes in the pool and adding new haplotypes as necessary.

A number of haplotype inference methods are based on statistical approaches, including expectation-maximization (EM) *e.g.* [40, 12, 30], and Bayesian approaches *e.g.* [25, 28, 36]. EM uses maximum likelihood to first estimate the frequency of each haplotype in the population and then to resolve each individual genotype. These methods assume that mating is random within the population. A major difference between EM and Bayesian approaches is that the latter stays within the domain of probability distributions and uses *prior* information, which incorporates beliefs about patterns that are expected for a set of haplotypes from a population. The priors that are assumed by Bayesian approaches can vary. For example, HAPLOTYPER [28] and the method developed by Lin et al. [25] both use a Dirichlet prior. This prior applies weights to favor a resolution for a genotype if one of the haplotypes is currently assumed to be in the sample. PHASE [36, 35] uses a prior that approximates the coalescent. This prior favors the same resolutions favored by the Dirichlet prior and also applies weights to favor resolutions in which both haplotypes are *similar* to haplotypes that are currently assumed to be in the sample.

The main assumptions made by EM and Bayesian approaches are that mating within the population is random and the likelihood function or posterior probability should be maximized. The coalescent-based prior also assumes that there are relatively few unique haplotypes and there are similarities between pairs of haplotypes in the population.

One potential pitfall of statistical approaches is that a single run finds one peak in the likelihood/ probability surface. To help overcome this drawback, some of these algorithms randomly restart a number of times and output the best solution found. For instance, the use of default settings for PHASE [36, 35] results in 100 runs being computed. However, unless the likelihood/probability surface is exhaustively searched, there is always the possibility of missing the "best" peak.

In this paper, we introduce an approach to inferring haplotypes which we call *SplittingHeirs*. Using haplotype data for which the true phases are known [1, 29] we compare this new approach with Pure Parsimony [16], HAP [18], EM-DeCODER [28], and PHASE [36]. SplittingHeirs has equal or higher accuracy for all seven data sets.

2. MATERIALS AND METHODS

SplittingHeirs finds a set of haplotypes that resolves a set of genotypes such that the resultant dense graph is optimized. In this section, the dense graph model is defined and the assumptions made by SplittingHeirs are summarized. We also describe the known haplotype data that are used for comparisons in our study.

2.1 The Dense Graph Model

Let h equal the number of unique haplotypes in a given solution. Consider a graph with h nodes, in which each node represents a haplotype in the solution. The weight on an edge in the graph is set equal to the distance between the two haplotypes that are endpoints of the edge. Distances between haplotypes can be defined in various ways. A simple distance measure is just the number of sites in which they differ.

For the Pure Parsimony model, the optimal solution would be a graph that has the least number of nodes possible. On the other hand, if minimization of pair-wise similarities was the objective, the optimal graph would contain edges from each node to its nearest neighbor, where the sum of the weights of these edges is the minimum possible. When relying completely on simple pair-wise distances, it is possible to have $h/2$ disjoint subgraphs with arbitrarily large distances between them. In real populations, we would expect to find *clusters* of haplotypes that are similar to each other, so it is desirable to enforce similarities beyond a single nearest neighbor. This enforcement is made by increasing the density of edges in the graph.

In a *dense* graph model, the density of the graph is required to be greater than or equal to a given value, α. The density of a graph can be defined as e/h, where e is the number of edges in the graph. By considering the additional edges required to achieve the mandated density, similarities beyond single nearest neighbors are taken into consideration. We evaluate the quality of the dense graph solution using:

$$C_D = \sum_{i=1}^{e} d_i + h \times u \tag{1}$$

where d_i is the distance of edge i and u is a weight. The dense graph with the minimum cost C_D is considered optimal. This objective simultaneously minimizes the sum of the edge distances and the cardinality of the haplotypes. Increasing the value of u will tend to yield graphs with fewer unique haplotypes. Reducing the value of u will yield graphs in which the similarities of haplotypes within each cluster tend to be increased at the cost of reducing parsimony.

We have cast this model as a Mixed Integer Linear Program (MIP). The constraints of our MIP require that the selected haplotypes resolve all of the genotypes. These constraints are similar to the constraints for the Pure Parsimony MIP formulation. The key differences between our MIP and the Pure Parsimony MIP is that our objective function is Equation (1), and we add the following constraint to ensure the density of the graph:

$$e/h \geq \alpha \tag{2}$$

The choice of the value used for α is considered in the Discussion section. Like Pure Parsimony, this problem may require exponential time to compute in the worst case. However, we were able to obtain globally optimal solutions for our trials using ILOG's Cplex 8.11, which is a generic MIP solver.

Once the set of haplotypes has been determined using this model, each genotype is resolved using haplotypes from this set. On some occasions, there may be more than one pair of haplotypes that can resolve a given genotype. When this is the case, SplittingHeirs assumes that common haplotypes

are very common, and assigns the pair that contains the haplotype with the highest frequency in the set. Alternate pairs, along with their frequencies, are also provided for the user.

Biological Observation and Intuition Behind the Dense Graph Model. Many of the methods previously used for haplotype inference have favored reduction of the cardinality of unique haplotypes. Pure Parsimony [16] is an extreme case in which a set of haplotypes are found such that the number of unique haplotypes is the least possible. SplittingHeirs favors reduced cardinality, but simultaneously considers other favorable properties and does not always yield a strictly parsimonious solution.

The dense graph model is biologically motivated as it utilizes three widely accepted principles: the number of unique haplotypes within a given population is relatively small, there are clusters of haplotypes that are similar to each other, and common haplotypes are very common. PHASE incorporates the first two of these principles in its priors. However, PHASE favors *pairs* of haplotypes that are similar. It is biologically intuitive that *clusters* of haplotypes are similar, not just pairs. SplittingHeirs favors solutions with cluster-wide similarities. However, this model does not require similarities between all haplotypes and the dense graph might be composed of more than one connected component, as illustrated by the small example that will be introduced shortly.

The dense graph model is also an intuitive approach when considering regions in which recombination is important. PHASE is based on a coalescent prior that basically assumes an evolutionary tree of haplotypes. However, in areas of high recombination, the tree-like structure is broken down by recombination and there is much reticulation; that is, a recombinant haplotype or clade of haplotypes will have similarities to both parental types. This reticulation results in a biological dense graph for the relationships among the haplotypes. Hence, the dense graph approach is a biologically more realistic representation of haplotype relationships when recombination is present than the coalescent evolutionary tree representation. A dense graph can effectively collapse into an evolutionary tree, but the inverse is not true. Therefore, this algorithm can deal with a broader range of realistic biological situations than programs such as PHASE.

Mixed Integer Linear Program Formulation. The following Mixed Integer Linear Program (MIP) is the model assumed by SplittingHeirs.

Assume there are n genotypes to be resolved. Let $H = \{h_1, h_2, \ldots\}$ be the set of variables representing all candidate haplotypes, where h_i is a binary variable that is equal to one if and only if haplotype i is selected for the resolving set of haplotypes. Let R_i be a set of binary variables representing the haplotype pairs that can resolve genotype i. More specifically, $R_i = \{r_{i\rho}\}$ such that ρ represents a pair of haplotypes, h_j and h_k that can resolve genotype i. $r_{i(j,k)} = 1$ if and only if genotype i is resolved by haplotypes h_j and h_k in the solution.

Each vertex in the solution graph represents a haplotype i such that $h_i = 1$. Let $X = \{x_{ij}\}$ be the set of binary variables representing edges between haplotype vertices. $x_{ij} = 1$ if and only if the edge between haplotypes i and j is in the solution graph. d_{ij} is the distance between haplotypes i and j. α is the minimum density required for the solution graph. Finally, u is a weighting factor.

The MIP formulation follows:

$$\min C_D = \sum_{i=1}^{|H|} \sum_{j=i+1}^{|H|} x_{ij} d_{ij} + \sum_{k=1}^{|H|} u h_k \tag{3}$$

subject to:

$$\sum_{R_i} r_{i\rho} = 1, \qquad 1 \le i \le n; \tag{4}$$

$$h_j \ge r_{i(j,k)}, \qquad 1 \le i \le n, (j,k) \in R_i; \tag{5}$$

$$h_k \ge r_{i(j,k)}, \qquad 1 \le i \le n, (j,k) \in R_i; \tag{6}$$

$$h_i \ge x_{ij}, \qquad 1 \le i \le |H|; \tag{7}$$

$$h_j \ge x_{ij}, \qquad 1 \le j \le |H|; \tag{8}$$

$$\sum_{i=1}^{|H|} \sum_{j=i+1}^{|H|} x_{ij} \ge \alpha \sum_{k=1}^{|H|} h_k \tag{9}$$

The objective function (3) minimizes selected distances between haplotypes and the cardinality of the haplotypes, as previously described. It has a slightly different formulation from the previous objective, but is mathematically equivalent. The version given here allows the introduction of variables that are used by the constraints. For our experiments, we used the Jukes-Cantor distance measure.

Constraints (4) require that each genotype is resolved by a pair of haplotypes. Constraints (5) and (6) require that all haplotypes that are selected to resolve genotypes are in the solution graph. Constraints (7) and (8) require that all of the edges in the graph also have their endpoints in the graph. Finally, constraint (9) requires that the density of the graph be greater than or equal to α.

SplittingHeirs has the option for allowing "absent" haplotypes. When this option is used, the graph might contain extra haplotypes that do not resolve any of the genotypes. This option is offered as the sample might not contain all of the haplotypes in the population. However, care is needed when a non-linear distance measure, such as Jukes-Cantor, is used. For Jukes-Cantor, the distance between a pair of haplotypes that differ at two sites is more than twice the distance between a pair of haplotypes that differ at only a single site. In this case, extra haplotypes might be inserted to reduce the sum of the distances. For this reason we require that each "absent" haplotype has a degree of at least three. This option was used for all of our trials.

Example. We use the following small, highly heterozygous example to illustrate the dense graphs for various algorithms: $g1$: 1111 0001, $g2$: 2212 0202, $g3$: 2220 2102, $g4$: 2222 2121, and $g5$: 2022 0222. The wide range of feasible solutions offered by this example helps to illuminate assumptions that are made by the underlying models upon which each method is based.

Despite the high heterozygosity of the example, three algorithms derived the same solution. Clark's Subtraction Method, Pure Parsimony, and EM (using EM-DeCODER) computed the same solution consisting of five unique haplotypes. The solution follows:

$g1$: 1111 0001 $\rightarrow$ 1111 0001 $\oplus$ 1111 0001
$g2$: 2212 0202 $\rightarrow$ 1111 0001 $\oplus$ 0010 0100
$g3$: 2220 2102 $\rightarrow$ 1100 1101 $\oplus$ 0010 0100
$g4$: 2222 2121 $\rightarrow$ 1100 1101 $\oplus$ 0011 0111
$g5$: 2022 0222 $\rightarrow$ 1001 0011 $\oplus$ 0010 0100

There is no perfect phylogeny solution for our example problem. The imperfect phylogeny program HAP computed a solution composed of six unique haplotypes as follows:

$g1$: 1111 0001 → 1111 0001 ⊕ 1111 0001
$g2$: 2212 0202 → 1010 0101 ⊕ 0111 0000
$g3$: 2220 2102 → 1010 0101 ⊕ 0100 1100
$g4$: 2222 2121 → 1010 0101 ⊕ 0101 1111
$g5$: 2022 0222 → 1010 0101 ⊕ 0001 0010

PHASE version 2.1 with default settings computed a solution comprised of nine unique haplotypes, which is the maximum number of unique haplotypes that could be used to resolve the example problem. The PHASE solution follows:

$g1$: 1111 0001 → 1111 0001 ⊕ 1111 0001
$g2$: 2212 0202 → 0011 0001 ⊕ 1110 0100
$g3$: 2220 2102 → 0100 0100 ⊕ 1010 1101
$g4$: 2222 2121 → 0000 1101 ⊕ 1111 0111
$g5$: 2022 0222 → 0000 0111 ⊕ 1011 0000

Using Splitting Heirs, we solved the example problem with $u = 0.25$, and $\alpha = 1.5$, which yielded a disconnected dense graph for a small example. The solution is comprised of eight unique haplotypes. The cost C_D for the dense graph is 26. The solution follows:

$g1$: 1111 0001 → 1111 0001 ⊕ 1111 0001
$g2$: 2212 0202 → 1111 0001 ⊕ 0010 0100
$g3$: 2220 2102 → 0000 0100 ⊕ 1110 1101
$g4$: 2222 2121 → 0000 0111 ⊕ 1111 1101
$g5$: 2022 0222 → 1011 0001 ⊕ 0000 0110

Clark's Subtraction Method, Pure Parsimony, HAP, EM-DeCODER, and PHASE are not designed to find the optimal dense graph, so those solutions are likely to correspond to suboptimal dense graphs and cannot have a lower cost C_D than that found by SplittingHeirs. Using the haplotypes that were selected by each method, we derived the dense graph with the least C_D cost. In other words, for a solution with h haplotypes, we found $\alpha \times h$ edges that had the least cost. Figure 2 illustrates these graphs. The costs, C_D, are equal to 33.25 for Clark's method, Pure Parsimony, and EM-DeCODER, 35.5 for HAP, and 38.25 for PHASE. As expected, these costs are all substantially more than the minimum possible cost of 26.

2.2 Haplotype Data

We have tested the accuracy of various haplotype inference methods using haplotype data for which the true phases were derived experimentally (i.e. the individual haplotypes were identified, not the melded pairs). The data came from two sources. The size of the sets, number of ambiguous genotypes, degree of heterozygosity, and recombination rates are listed in Table 1.

The first source of data is a set of 80 human *ApoE* haplotype pairs, each with nine SNPs, that was experimentally found by Orzack et al. [29]. These SNPs are drawn from the *apolipoprotein E* locus. The individuals were unrelated and 18 were classified as Asians, 19 as Blacks, and 43 as Caucasians. Templeton et al. found that there is no statistically significant recombination in this region [39]. Dataset A in Table 1 is composed of these 80 pairs of haplotypes.

The second source of data was experimentally collected by Andrés et al. [1]. It contains 39 pairs of human haplotypes, each with 411 sites, in a 48 kb region containing the

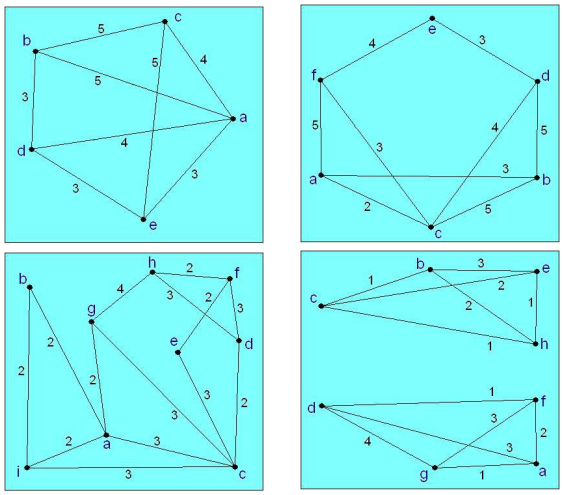

Figure 2: Dense graphs for the solutions found using (a) Clark's Subtraction Method, Pure Parsimony, and EM-DeCODER, (b) HAP, (c) PHASE, and (d) Splitting Heirs. The number of sites that are different for a pair of haplotypes are used for the edge weights.

KLK13 and *KLK14* genes. There is a substantial amount of missing data in this set. Pure Parsimony, EM-DeCODER, and the current implementation of SplittingHeirs all require complete data. Six regions of complete data from this set are used for this study and correspond to datasets B through G in Table 1. They range from 5 sites to 47 sites in length. Recombination rates for this data have been found by Maxwell et al. [27]. The 17 sites of set C have no recombination and are combined with 9 additional sites, which have a low recombination rate, to make set E. The seven datasets, as well as the input files that were used for our tests, are available by contacting the first author.

In this paper, we use a "gold standard" of experimentally derived haplotypes to evaluate the accuracy of computational methods without bias. A number of haplotype datasets have been used in previous studies, including sets from Drysdale et al. [10], Rieder et al. [32], Hinds et al. [20], and HapMap data [43]. While all of these datasets were derived from real biological genotypes, they have been phased using computational methods. Due to the lack of direct experimental validation, they are not suitable for evaluating the biological correctness of haplotype inference algorithms.

Another source of haplotype data is derived from the HapMap [43] data in the following way [26]. Roughly two-thirds of the HapMap data is composed of trios, where a trio consists of two parents and a child. The genotypes of the parents can be partially phased by examining the child's genotype. However, not all sites can be accurately phased. These cases arise most frequently at sites with high heterozygous rates throughout the population. In the previous study, these sites were treated as missing data [26]. Thus, the use of trios requires omitting data in a highly nonrandom way, excluding sites that are highly heterozygous within the population. These sites have the least pattern, and consequently increase the complexity of this combinatorial problem. Evaluating inference algorithms on data with highly heterozygous sites

Table 1: Haplotype data sets used for comparisons, including name of data set, nucleotide range in reference sequence, number of genotypes in set (n), number of sites in each genotype (m), number of genotypes that are ambiguous (number of genotypes that have at least two heterozygous sites), percentage of all sites that are heterozygous, and recombination rate.

Data	Nucleotide Range	n	m	# Ambiguous	% Heterozygous	Recombination
A	17874-21388	80	9	47	21.4%	none
B	667-1464	39	5	13	22.6%	none
C	32107-34389	39	17	27	20.5%	none
D	12867-13729	39	8	15	16.7%	low
E	32107-35800	39	26	28	19.0%	low
F	8043-9256	39	22	26	21.7%	high
G	33117-38365	39	47	33	14.0%	high

missing is akin to evaluating chess strategies on games in which the knights have been removed. There is no reason to believe that the champion chess algorithm would triumph for regular chess games.

Andres et al. [1] tested PHASE [35, 36], fastPHASE [33], HAP [11, 18], and GERBIL [21] on known haplotypes that were directly identified, not inferred from genotype data. They found the accuracy was poor for all of the methods tested and the confidences computed were error prone. On the other hand, Marchini et al. found the accuracy to be very high (close to the typing error rate) when using HapMap trio data [26]. This discrepancy illustrates the bias that is introduced when using trio data for testing accuracy.

3. RESULTS

In this section, we compare SplittingHeirs with several popular haplotype inference methods: Pure Parsimony [16], HAP [18], EM-DeCODER [28], and PHASE [36]. Two of these implementations use combinatorial methods and the other two use statistical approaches.

We used the default settings for all programs. For SplittingHeirs, we set $u = 0.175$ and $\alpha = 2$, which are the current default values. The "best" pairs predicted by each method were used to score the accuracy for that technique. EM-DeCODER cannot handle more than 20 SNPs, so there are no results for the three largest sets of data using this method.

To reduce computation time and space for our SplittingHeirs implementation, Taylor Maxwell suggested using a *candidate* haplotype list, instead of considering every possible haplotype pair that can resolve each genotype. These candidate haplotypes can be derived by a number of methods, such as simply removing all haplotypes that can only be used to infer a single genotype each. By reducing the input size, the computation time and space can be reduced. However, the filtering process can lead to errors if the wrong data is omitted. With SplittingHeirs, the user can decide whether or not to filter their data and to what extent. One method for filtering the data is to compute a list of haplotypes that have a probability above some threshold according to a statistical method. We used all of the feasible haplotypes for datasets A, B, and D. To reduce computation time, we used candidate lists for the remaining datasets. The haplotypes in these lists have a probability of at least 1% of appearing in the solution, according to supplementary information provided by PHASE.

Figure 3 shows the results for the four datasets, A, B, C, and D, that were solvable using all five methods: Pure Parsimony, HAP, EM-DeCODER, PHASE, and SplittingHeirs. All of these datasets have little or no recombination. When a genotype is incorrectly phased, it may contain just a single site that is incorrect, or a number of sites. The plots in Figure 3 show the number of genotypes incorrectly phased as well as the total number of sites that were incorrectly phased by each method. SplittingHeirs did better than, or as well as, all of the other solvers in every case.

Figure 4 depicts the results for the other three datasets. Two of these datasets, F and G, have high recombination rates. Splitting Heirs outperformed the other algorithms on all three datasets.

Our tests were run as single processes on Athlon 1.9 MHz dual processors with two gigabytes shared memory. None of the datasets tested required more than three minutes to compute using HAP, EM-DeCODER, or PHASE. Computation times varied dramatically for SplittingHeirs, requiring less than two minutes for sets A and B; several minutes for C and E; 25 minutes for set D;[2] and several days for the two sets with high recombination rates. These results suggest that high recombination rates may negatively impact computation times.

We are currently working on strategies to speed up our implementation. In particular, we are investigating the use of Cut-and-Solve [7] for solving the MIP. This method has been shown to speed up computation times for difficult MIPs. Furthermore, it is an *anytime* solver that can be terminated early with the best solution found thus far returned. In the meantime, SplittingHeirs is valuable for research focused on small regions, such as candidate locus studies. Improving the accuracy of phasing will strongly impact the quality of results from subsequent analyses, such as nested clade analysis and tree scanning.

Table 3 lists the number of unique haplotypes in the solutions computed by each method and the actual number of unique haplotypes in the datasets for which the true phase is known. As previously noted [6], the true number of unique haplotypes in each of the datasets is relatively small, but it is not always the most parsimonious, as seen for datasets A, F, and G.

Finally, we observed a correlation between the number of connected components in the graphs and the degree of recombination. Datasets A and B had a single component, G had six components, and the others had two or three components each.

4. DISCUSSION

In this section, the derivation of a default value for the

[2]It took less than one second to compute dataset A, B, or D when a candidate haplotype list was provided.

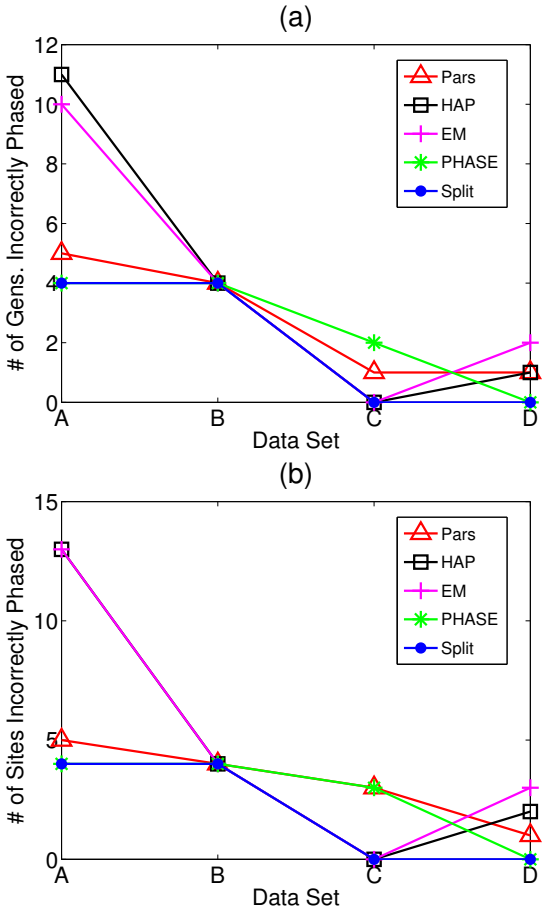

Figure 3: Results for the four datasets, A, B, C, and D, that were solvable using all five methods. The number of genotypes (a) and sites (b) incorrectly phased by Pure Parsimony (Pars), HAP, EM-DeCODER (EM), PHASE, and SplittingHeirs (Split) are shown.

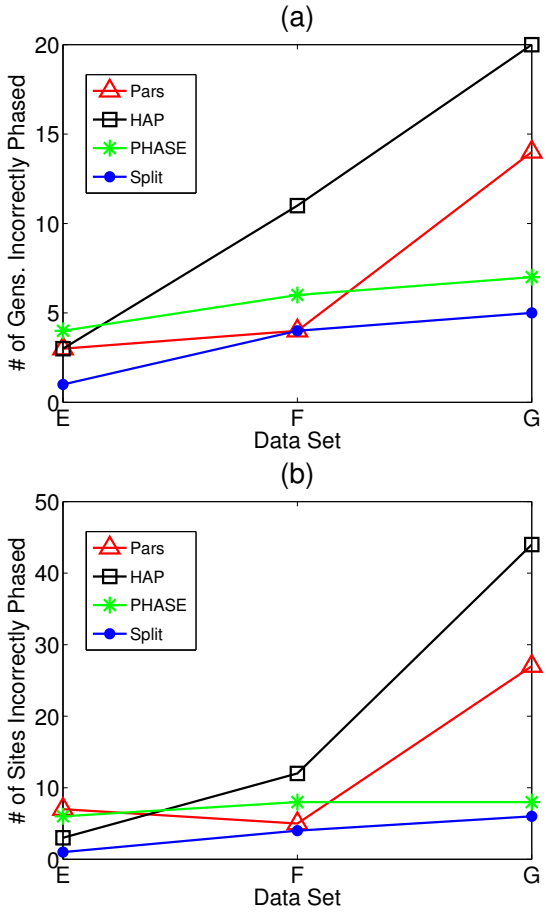

Figure 4: Results for datasets E, F, and G. The number of genotypes (a) and sites (b) incorrectly phased by Pure Parsimony (Pars), HAP, PHASE, and SplittingHeirs (Split) are shown.

density parameter α is discussed and sources of error that arise for haplotype inference are considered.

4.1 Default Value for α

SplittingHeirs uses a dense graph model to evaluate solutions. A key question remaining is how dense should the graph be – i.e. what value should be used for α. This value should be related to the diversity of the sample. In general, the overall diversity of the haplotypes in a population may vary from one study to another. This diversity can be affected by the age and size of the population, the degree of gene flow, and other population level properties. Furthermore, given a population, the overall diversity of the haplotype pool can also vary over different regions of the genome. Some regions are highly conserved while others undergo a high rate of mutations.

The density of the graph, e/h, is required to be greater than or equal to α. Choosing a large α value forces greater density and should only be used in cases when the diversity of the sample is expected to be small. Conversely, α can be reduced for samples that are expected to be exceptionally diverse. The default value of α is 2.0. Using the five datasets with little or no recombination, we tested the sensitivity of α by experimenting with various values. The solutions for datasets A and B did not change for α values ranging from 1.0 to 2.8. The solutions for the other three datasets were best in the range from 1.8 to 2.4. In the Results section, we used the default value of 2.0 for all trials.

4.2 Sources of Error

Haplotype inference is particularly challenging as it has several sources of error, some of which can be minimized and others that may be inherently unavoidable.

First of all, the assumptions that are made by a model may introduce error. It is not known what qualities of a solution are important and to what degree they should be pursued. For example, Clark's Subtraction Method, Pure Parsimony, and EM completely disregard the similarities between haplotypes, PHASE favors pair-wise similarities with prescribed weights, and SplittingHeirs favors cluster-wide similarities with a different prescribed weight.

Second, the method for solving a model may introduce error. For instance, statistical methods can miss peaks on the likelihood/probability surface. Another source of error

Table 2: The number of unique haplotypes in each of the data sets for which the phase is known (True) and the number of unique haplotypes in the solutions found by Pure Parsimony (Pars), HAP, EM-DeCODER (EM), PHASE, and SplittingHeirs (Split). The true solutions don't always have the most parsimonious number of haplotypes, as seen for sets A, F, and G.

Data	True	Pars	HAP	EM	PHASE	Split
A	17	15	20	20	16	15
B	7	7	7	7	7	7
C	12	12	12	12	12	12
D	7	7	7	8	7	7
E	16	16	20	-	17	16
F	18	17	25	-	20	18
G	32	28	40	-	28	28

is the use of a technique called *partition-ligation* [28, 35]. This strategy was introduced for HAPLOTYPER [28] and later added to PHASE [35]. When using partition-ligation, genotypes are cut up into short segments, solved, and recombined. While this approach generally reduces computation time, it introduces additional sources for error. On the other hand, Splitting Heirs, Pure Parsimony, and Perfect Phylogeny compute globally optimal solutions.

Finally, there is a source of error that will always persist as long as the input consists solely of a set of genotypes. Even if the model precisely and accurately quantified true biological assumptions and it computed the optimal solution exactly, the stochasticity of nature introduces uncertainty.

Dataset B illustrates this uncertainty. Eight of the 39 genotypes are identical. Each has two heterozygous sites and is encoded as 02002. However, four of these genotypes are conflations of 01000 and 00001 and the other four are conflations of 01001 and 00000. All four of these haplotypes were in the solutions for all five methods. The best an algorithm can do in this case is to identify both pairs as probable, use biological assumptions to rank the pairs, notify the user of the other possibilities, and assign the "best" pair to all eight genotypes. SplittingHeirs and PHASE correctly identified both pairs as probable. Interestingly, they chose opposite pairs as the "best" – resulting with the same score. Pure Parsimony, HAP, and EM-DeCODER chose the same pair as SplittingHeirs for all eight genotypes, but those implementations don't explicitly identify other probable pairs.

In addition to having multiple pairs of haplotypes in the solution set that may resolve a single genotype, there are other sources of stochastic error associated with this problem. For instance, if the number of genotypes is small in comparison to the size of the population, the sample may not contain the same proportions of haplotypes that exist in the entire population.

Another consideration is gene flow between populations. Modern technology has put diverse populations in closer contact than ever before. Recent gene flow between populations can confound some haplotype inference models. For instance, a phylogenetic tree that captures all of the haplotypes in a diverse population requires that a common ancestor from deep in the past is identified. Such a requirement invites opportunities for errors. A favorable feature of the dense graph model used by SplittingHeirs is that disjoint subgraphs are possible, as illustrated by the example problem in the Methods section. This model does not force a universal connectivity on the solution.

Sometimes the error introduced by assumptions implied by the model, error from compromising optimality in the computation method, and/or error due to stochasticity in nature are captured by considering all of the solutions with confidence levels above a threshold as provided by statistical methods. However, statistical methods may completely miss the peak that is closest to the true solution. In these cases, the true solution won't be captured even if a very large number of solutions are considered. When accuracy is paramount, identification of all global optimal and near optimal solutions is more promising. A future version of SplittingHeirs will offer this option.

4.3 Conclusion

The haplotype inference problem has a rich history and has been approached using a number of various combinatorial and statistical methods. Due to consequences for vitally important studies, the benefits of accuracy for the haplotype inference problem reach far beyond mere financial gains. SplittingHeirs finds globally optimal solutions for this problem that favor low cardinality of unique haplotypes as well as similarities across clusters of haplotypes. Favoring cluster-wide similarities is biologically intuitive and this assumption is experimentally validated using known haplotype data. In this paper, we used seven sets of data for which the true phase is known to test the accuracy of Pure Parsimony, HAP, EM-DeCODER, PHASE, and SplittingHeirs. SplittingHeirs tied for highest accuracy for four of the datasets and outperformed all of the methods tested for the remaining three datasets. Furthermore, SplittingHeirs is robust and had higher accuracy than the other haplotype inference methods for genotypes with high recombination rates.

5. ACKNOWLEDGMENTS

This research was supported in part by an Olin Fellowship; NIH grants P50-GM065509, R01-GM02871924A2, and U01-GM063340; NSF grants IIS-0535257 and DBI-0743797; and a grant from the Alzheimer's Association. Thanks to Taylor Maxwell for many useful discussions, and to Gerold Jäger, who provided the Pure Parsimony solutions in this paper.

6. REFERENCES

[1] A. M. Andrés, A. G. Clark, E. Boerwinkle, C. F. Sing, and J. E. Hixson. Understanding the accuracy of statistical haplotype inference with sequence data of known phase. *Genet. Epi.*, 31:659–671, 2007.

[2] M. R. Barnes. Navigating the HapMap. *Briefings in Bioinformatics*, 7:211–224, 2006.

[3] D. G. Brown and I. M. Harrower. Integer programming approaches to haplotype inference by pure parsimony. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3(2):141–154, April-June 2006.

[4] M. Cargill, D. Altshuler, J. Ireland, P. Sklar, K. Ardlie, N. Patil, N. Shaw, C. R. Lane, E. P. Lim, N. Kalyanaraman, J. Nemesh, L. Ziaugra, L. Friedland, A. Rolfe, J. Warrington, R. Lipshutz, G. Q. Daley, and E. S. Lander. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet.*, 22(3):231–238, July 1999.

[5] A. G. Clark. Inference of haplotypes from PCR-amplified samples of diploid populations. *Molecular Biology and Evolution*, 7:111–122, 1990.

[6] S. Climer, G. Jäger, A. R. Templeton, and W. Zhang. How frugal is Mother Nature with haplotypes? *Bioinformatics*, 25(1):68–74, 2009.

[7] S. Climer and W. Zhang. Cut-and-solve: An iterative search strategy for combinatorial optimization problems. *Artificial Intelligence*, 170:714–738, June 2006.

[8] F. S. Collins, M. S. Guyer, and A. Chakravarti. Variations on a theme: Cataloging human DNA sequence variation. *Science*, 278(5343):1580–1581, November 1997.

[9] D. N. Cooper. *Human Gene Evolution*. BIOS Scientific Publishers, Oxford, 1999.

[10] C. M. Drysdale, D. W. McGraw, C. B. Stack, J. C. Stephens, R. S. Judson, K. Nandabalan, K. Arnold, G. Ruano, and S. B. Liggett. Complex promoter and coding region b_2-adrenergic receptor haplotypes alter receptor expression and predict *in vivo* responsiveness. *Proceedings of the National Academy of Science*, 97:10483–10488, September 2000.

[11] E. Eskin, E. Halperin, and R. M. Karp. Large scale reconstruction of haplotypes from genotype data. In *The Seventh Annual International Conference on Computational Biology*, pages 104–113, 2003.

[12] L. Excoffier and M. Slatkin. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol*, 12(5):921–927, September 1995.

[13] D. Fernández-Baca and J. Lagergren. A polynomial-time algorithm for near-perfect phylogeny. *SIAM Journal of Computing*, 32(5):1115–1127, 2003.

[14] S. B. Gabriel, S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. Defelice, A. Lochner, M. Faggart, S. N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E. S. Lander, M. J. Daly, and D. Altshuler. The structure of haplotype blocks in the human genome. *Science*, 296:2225–2229, 2002.

[15] D. Gusfield. Haplotyping as perfect phylogeny: Conceptual framework and efficient solutions. In *Research in Computational Molecular Biology (RECOMB '02)*, pages 166–175, 2002.

[16] D. Gusfield. Haplotype inference by pure parsimony. In *14th Annual Symposium on Combinatorial Pattern Matching (CPM'03)*, pages 144–155, 2003.

[17] D. Gusfield and S. H. Orzack. Haplotype inference. In S. Aluru, editor, *Handbook on Bioinformatics*. CRC, 2005.

[18] E. Halperin and E. Eskin. Haplotype reconstruction from genotype data using imperfect phylogeny. *Bioinformatics*, 20:1842–1849, 2004.

[19] M. K. Halushka, J. B. Fan, K. Bentley, L. Hsie, N. Shen, A. Weder, R. Cooper, R. Lipshutz, and A. Chakravarti. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genetics*, 22:239–247, 1999.

[20] D. A. Hinds, L. L. Stuve, G. B. Nilsen, E. Halperin, E. Eskin, D. G. Ballinger, K. A. Frazer, and D. R. Cox. Whole-genome patterns of common DNA variation in three human populations. *Science*, 307:1072–1079, February 2005.

[21] G. Kimmel and R. Shamir. GERBIL: Genotype resolution and block identification using likelihood. *Proceedings of National Academy of Science USA*, 102:158–162, 2005.

[22] M. Kimura. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61:893–903, 1969.

[23] L. Kruglyak and D. A. Nickerson. Variation is the spice of life. *Nature Genetics*, 27:234–236, 2001.

[24] W. H. Li and L. A. Sadler. Low nucleotide diversity in man. *Genetics*, 129:513–523, 1991.

[25] S. Lin, D. J. Cutler, M. E. Zwick, and A. Chakravarti. Haplotype inference in random population samples. *The American Journal of Human Genetics*, 71:1129–1137, 2002.

[26] J. Marchini, D. Cutler, N. Patterson, M. Stephens, E. Eskin, E. Halperin, S. Lin, Z. Qin, H. Munro, G. Abecasis, P. Donnelly, and I. H. C. (2006). A comparison of phasing algorithms for trios and unrelated individuals. *Amercan Journal of Human Genetics*, 78:437–450, 2006.

[27] T. J. Maxwell, K. E. Hyma, L. C. Shimmin, E. Boerwinkle, J. E. Hixson, and A. R. Templeton. The impact of nonrandom mutation, recombination, and gene conversion on shaping haplotype variation in the KLK region of human chromosome 19 and its implications for association studies. To appear.

[28] T. Niu, Z. Qin, X. Xu, and J. Liu. Bayesian haplotype inference for multiple linked single nucleotide polymorphisms. *The American Journal of Human Genetics*, 70:157–169, 2002.

[29] S. H. Orzack, D. Gusfield, J. Olson, S. Nesbitt, L. Subrahmanyan, and V. P. S. Jr. Analysis and exploration of the use of rule-based algorithms and consensus methods for the inferral of haplotypes. *Genetics*, 165:915–928, October 2003.

[30] Z. S. Qin, T. Niu, and J. S. Liu. Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *Am J Hum Genet*, 71:1242–1247, November 2002.

[31] R. Redon, S. Ishikawa, K. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. Carson, W. Chen, E. K. Cho, S. Dallaire, J. F. J. Gonzalez, M. Gratacos, J. Huang, D. Kalaitzopoulos, D. Komura, J. MacDonald, C. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, J. Zhang, L. Armengol, D. Conrad, X. Estivill, C. Tyler-Smith, N. Carter, H. Aburatani, C. Lee, K. J. KW, S. Scherer, and M. H. ME. Global variation in copy number in the human genome. *Nature*, 444(7118):444–454, November 2006.

[32] M. J. Rieder, S. L. Taylor, A. G. Clark, and D. A. Nickerson. Sequence variation in the human angiotensin converting enzyme. *Nature Genetics*, 22:59–62, 1999.

[33] P. Scheet and M. Stephens. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and

haplotypic phase. *The American Journal of Human Genetics*, 78:629–644, 2006.

[34] Y. S. Song, Y. Wu, and D. Gusfield. Algorithms for imperfect phylogeny hapltoyping (IPPH) with a single homoplasy or recombination event. *Workshop on Algorithms in Bioinformatics 2005. Lecture Notes in Computer Science*, 3692:152–164, 2005.

[35] M. Stephens and P. Donnelly. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *The American Journal of Human Genetics*, 73:1162–1169, 2003.

[36] M. Stephens, N. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, 68:978–989, 2001.

[37] A. R. Templeton. Haplotype trees and modern human origins. *Yearbook of Physical Anthropology*, 48:33–59, 2005.

[38] A. R. Templeton and N. J. Georgiadis. A landscape approach to conservation genetics: conserving evolutionary processes in the african bovidae. In J. C. Avise and J. L. Hamrick, editors, *Conservation Genetics: Case Histories From Nature*, pages 398–430. Chapman & Hall, New York, 1996.

[39] A. R. Templeton, T. Maxwell, D. Posada, J. H. Stengard, E. Boerwinkle, and C. F. Sing. Tree scanning: a method for using haplotype trees in genotype/phenotype association studies. *Genetics*, 169:441–453, 2005.

[40] A. R. Templeton, C. F. Sing, A. Kessling, and S. Humphries. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. II. The analysis of natural populations. *Genetics*, 120:1145–1154, 1988.

[41] The Celera Genomics Sequencing Team. The sequence of the human genome. *Science*, 291:1304–1351, February 2001.

[42] The International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437:1299–1320, 2005.

[43] The International HapMap Consortium. A second generation human haplotype map of over 3.1 million snps. *Nature*, 449:851–861, 2007.

[44] The International Human Genome Mapping Consortium. A physical map of the human genome. *Nature*, 409:934–941, February 2001.

[45] D. G. Wang, J. B. Fan, C. J. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour, N. Perkins, E. Winchester, J. Spencer, L. Kruglyak, L. Stein, L. Hsie, T. Topaloglou, E. Hubbell, E. Robinson, M. Mittmann, M. S. Morris, N. Shen, D. Kilburn, J. Rioux, C. Nusbaum, S. Rozen, T. J. Hudson, R. Lipshutz, M. Chee, and E. S. Lander. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, 280(5366):1077–1082, May 1998.