# Combinatorial Problems Arising in SNP and Haplotype Analysis

Bjarni V. Halldórsson, Vineet Bafna*, Nathan Edwards, Ross Lippert, Shibu Yooseph, and Sorin Istrail

Informatics Research, Celera Genomics/Applied Biosystems
45 W. Gude Drive, Rockville MD, 20850 USA
{Bjarni.Halldorsson,Nathan.Edwards,Ross.Lippert}@celera.com
{Shibu.Yooseph,Sorin.Istrail}@celera.com

**Abstract.** It is widely anticipated that the study of variation in the human genome will provide a means of predicting risk of a variety of complex diseases. This paper presents a number of algorithmic and combinatorial problems that arise when studying a very common form of genomic variation, single nucleotide polymorphisms (SNPs). We review recent results and present challenging open problems.

## 1 Introduction

Genomes can be considered to be a collection of long strings, or sequences, from the alphabet {A,C,G,T}. Each element of the alphabet encodes one of four possible *nucleotides*. With the completion of the sequencing of the human genome, efforts are underway to catalogue genomic variations across human populations. *Single Nucleotide Polymorphisms* or SNPs constitute a large class of these variations. A SNP is a single base pair position in genomic DNA at which different nucleotide variants exist in some populations; each variant is called an *allele*. In human, SNPs are almost always biallelic; that is, there are two variants at the SNP site, with the most common variant referred to as the *major allele*, and the less common variant as the *minor allele*. Each variant must be represented in a significant portion of the population to be useful.

Diploid organisms, such as humans, possess two nearly identical copies of each chromosome. In this paper, we will refer to a collection of SNP variants on a single chromosome copy as a *haplotype*. Thus, for a given set of SNPs, an individual possesses two haplotypes, one from each chromosome copy. A SNP site where both haplotypes have the same variant (nucleotide) is called a *homozygous* site; a SNP site where the haplotypes have different variants is called a *heterozygous* site. The conflated (mixed) data from the two haplotypes is called a *genotype*. Thus, in genotype data, while the nucleotide variants at homozygous and heterozygous sites are known, the information regarding which heterozygous

* Current address: The Center for Advancement of Genomics, 1901 Research Blvd., 6th floor, Rockville, MD 20850, USA, vbafna@tcag.org.

C.S. Calude et al. (Eds.): DMTCS 2003, LNCS 2731, pp. 26–47, 2003.

site SNP variants came from the same chromosome copy, is unknown. See Figure 1 for an example of these concepts. Haplotypes play a very important role in several areas of genetics, including mapping complex disease genes, genome wide association studies, and also in the study of population histories. Unfortunately, current experimental techniques to infer the haplotype of an individual are both expensive and time consuming. However, it is possible to determine the genotype of an individual quickly and inexpensively. Computational techniques offer a way of inferring the haplotypes from the genotype data.

```
----------A----------C----------A----------T------G--------
----------A----------T----------G----------T------C--------
```

**Fig. 1.** Two sequences from the same region on two nearly identical copies of a chromosome of an individual. Only the SNPs have been shown with the non-SNP positions labeled with a "-". In this example there are five SNPs. The first and the fourth SNP sites are homozygous, and the remaining three SNP sites are heterozygous. The individual has the two haplotypes $ACATG$ and $ATGTC$; the genotype is $A\{C,T\}\{A,G\}T\{G,C\}$

Out of the two nearly identical copies of each chromosome in an individual, one copy is inherited from the paternal genome and the other copy from the maternal genome. This simple picture of inheritance is complicated by a process known as *recombination*, which takes place during meiosis - a process involved in the formation of reproductive cells (or *gametes*) in the parents. During recombination, portions of the paternal and maternal chromosomes are exchanged (Figure 2). Recombination can result in haplotypes in offsprings that are different from those in the parents. The site on the chromosome where a recombination occurs is called a recombination site. On average, one or two recombinations occur per chromosome per generation [36].

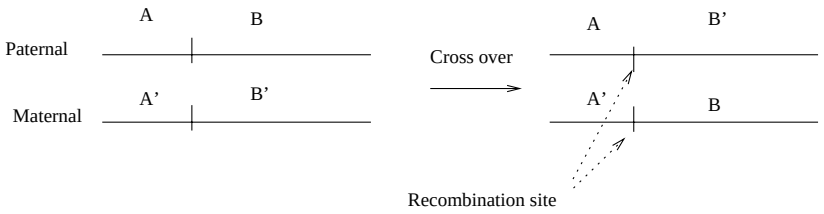


**Fig. 2.** An illustration of the recombination process that occurs during meiosis. Recombination is characterized by a cross-over event in which a portion of a paternal chromosome is exchanged with a portion of a maternal chromosome. This can result in the offspring having different haplotypes from those in the parents

In population studies, it has been shown that the likelihood that a site will act as a recombination site is not uniform across a chromosome[36], recombination sites occur much more frequently than expected in certain chromosomal regions and much less frequently in other chromosomal regions. Regions of high recombination site frequency are called *recombination hotspots*. Several recent studies [14,34,46] have suggested that human genetic variation consists largely of regions of low recombination site frequency, delineated by regions of high recombination site frequency, resulting in *blocks* of SNPs organized in mini-haplotypes.

An assumption that underlies much of population genetics, called the *infinite sites model*, requires that the mutation that results in a SNP occur only once in the history of a population, and therefore, that all individuals with a variant allele must be descendants of a single ancestor. While the infinite sites model is clearly a simplification of the true mechanism of genetic mutation, models of genetic variation built under this assumption compare favorably with empirical population genetics studies. Some of the models and algorithms in the text to follow will assume an *infinite sites model*.

## 2 The Haplotype Phasing Problem

In this section, we consider the problem of haplotype phasing: Given a set of genotypes, find a *good* set of haplotypes that resolve the set.

Generically the haplotype phasing problem can be posed as:

**Haplotype Phasing (Generic)**
**Input:** A set $G$ of genotypes.
**Output:** A set $H$ of haplotypes, such that for each $g \in G$ there exists $h_1, h_2 \in H$ such that the conflation of $h_1$ with $h_2$ is $g$.

An alternate related problem is haplotype frequency estimation. In this problem we care primarily about estimating the frequency of each potential haplotype in the population, and less so about the phasings of particular individuals.

By typing genetically related individuals one can get a better estimate of haplotypes present since the haplotype pair of a child is constrained by its inheritance from his parents. This version of the problem is considered in various software packages [1]. In this paper, we assume that such pedigree data is not available to us, however recasting the problems presented here in the presence of pedigree data is a worthwhile avenue of research.

Haplotype phasing has a variety of applications, each of which warrant different methodologies. Coarsely, one can partition haplotype phasing problems into three classes, based on their tractability:

**Small.** The number of sites is small enough that solutions requiring exponential space or time in it would be practical. It is sufficient for analyzing the SNPS in the vicinity of a single gene.

**Medium.** The number of sites is small enough that methods which are polynomial in the number of sites and individuals are practical. Number of individuals and number of sites may be on the order of $100's$. This size roughly corresponds to the number of SNPs across a region spanning several genes.

**Large.** Chromosome size, where algorithms which are linear in the number of SNPs are the only ones practical. The number of sites could be in the tens of thousands while the number of individuals sampled is small.

Additionally, many of the population genetics assumptions that hold for the small problems will not extend easily to the medium and large problems where the effects of recombination become significant. Different measures of success are appropriate depending on the problem size. Given a set of genotypes with a priori phasing information, a natural questions to ask is whether the algorithm retrieves the correct phasing. For small and medium problems, appropriate measures include the number of haplotypes that are predicted correctly or the difference in population frequency of the haplotypes in the known and the predicted set. For very large problems it is likely that these measures will be blunt and all methods will not perform well. An alternate measure suggested in [39] is the number of crossovers to explain the correct haplotypes from the predicted haplotypes.

When presenting the problems, we will assume that the genotype information we have is accurate. However, in practice, this is not the case, current genotyping technologies will fairly frequently not call genotypes (missing data) and less frequently miscall a genotype (wrong data). A practical algorithm needs to deal with these problems, in particular the missing data problem. The discussion in this paper is in terms of SNP's, most of the results and methods also will apply, perhaps with some modification, to studies of alternate genetic variations (markers) such as microsatellites.

*Notation.* We will follow notation by Gusfield [21] for haplotypes and genotypes. We will arbitrarily label the two alleles of any SNP 0 and 1. A genotype, representing a pair of haplotypes, can take three values for each SNP, corresponding to the observation of $\{0\}, \{1\}, \{0,1\}$. To simplify notation we will use 0 for $\{0\}$, 1 for $\{1\}$ and 2 for $\{0,1\}$. We will say that a SNP is *ambiguous* in a genotype if it has value 2. A genotype is *ambiguous* if it contains more than one *ambiguous* SNP.

We will generally use subscripts for objects associated with haplotypes and superscripts for objects associated with genotype. For example, the probability of observing the genotype $g$ in a given population might be given as $\phi^g$ and the haplotype probabilities as $\phi_h$. Since superscripts are possibly confused with exponentiation, explicit parentheses will be placed around exponentiated quantities to disambiguate this.

We will use $+$ to denote conflation and write $h + \bar{h} = g$ if the conflation of $h$ and $\bar{h}$ is $g$. To capture the notion that two haplotypes combine to make a genotype, we will, when convenient to do so, use the Kronecker delta, $\delta^g_{h+\bar{h}} = 1$ if $h + \bar{h} = g$ and 0 else.

We will denote the number of genotypes with $n$ and the number of SNP sites with $m$.

### 2.1 Clark's Rule

In a seminal paper[11], Clark proposed a common sense approach to phasing, that has become known as *Clark's rule*. Clark's rule is an inference method that resolves genotypes to their haplotype pairs. First, all homozygous and single ambiguous site genotypes are identified. The haplotypes that phase these genotypes are completely determined, forming an initial set of haplotypes supported by the data. Given a set of haplotypes $H$ representing the resolved genotypes, Clark's rule finds $g \in G$ and $h \in H$ such that $g = h + \bar{h}$ for some $\bar{h}$. The haplotype $\bar{h}$ is added to $H$. The process continues until either all genotypes are resolved, or no suitable pair of unresolved genotype and resolving haplotype $(g, h)$ exists.

Note that it may not even be possible to get this algorithm started if there are no homozygous or single ambiguous site genotypes. Further, there is no guarantee that a particular sequence of applications of Clark's rule will resolve all genotypes. Genotypes that remains unresolved after a maximal sequence of applications of Clark's rule are called *orphans*.

It should be clear from the description of Clark's rule that it describes a *class* of algorithms, each of which uses a different protocol for selecting a genotype-haplotype pair from which to infer a (typically) new haplotype. Clark's paper applies a greedy approach, in which the known haplotypes are tested against the unresolved genotypes in turn. The first genotype that Clark's rule can be applied to is resolved, potentially adding a new haplotype to the set of known haplotypes for the next iteration.

It is natural to ask for a Clark's rule application sequence that results in the fewest number of orphans. Clark's experiments [11] on real and simulated data suggest that the sequence of applications of Clark's rule that resolves the most genotypes generates fewest incorrect haplotype assignments.

*Problem 1 (Minimizing Orphans).* Find a sequence of Clark's rule applications that results in the fewest orphans.

Biological intuition about the nature of haplotypes present in human populations prompt us to think about versions of problem 1 that produce solutions that respect this intuition.

*Problem 2 (Maximizing Unique Resolutions).* Find a sequence of Clark's rule applications that maximizes the number of resolutions subject to the constraint that the final set of haplotypes must provide a single unique resolution to each genotype.

*Problem 3 (Minimizing Inference Distance).* Find a sequence of Clark's rule applications that minimizes the number of Clark's rule applications necessary to generate the genotypes' haplotypes.

Gusfield [21,22] studied a slightly restricted version of this problem, in which each genotype can participate in at most one Clark's rule application. Gusfield showed that finding an optimal Clark's rule application sequence is NP-hard,

but that in practice, on medium-sized instances, this version of the problem can be solved by a combination of careful enumeration and linear programming. Gusfield also evaluated the effectiveness of an algorithm incorporating a greedy application of Clark's rule with mixed results.

### 2.2 Maximum Likelihood

Hardy-Weinberg equilibrium (HWE) is the condition that the probability of observing a genotype is equal to the product of the probabilities of observing its constituent haplotypes (see [26]). Under this hypothesis, the probability of genotype $g$ in the population is related to the haplotype probabilities by the compact expression

$$\phi^g = \sum_{h+\bar{h}=g} \phi_h \phi_{\bar{h}}$$

where $\phi_h$ is the probability of haplotype $h$ in the population.

The *maximum likelihood method* of [17,27,41,60] estimates the haplotype probabilities $\phi_H = (\phi_h, \phi_{\bar{h}}, \ldots, \phi_{h'})$ from observed genotype frequencies $\hat{\phi}^G$ in $n$ individuals. The approach assumes HWE and a uniform prior on the $\phi_h$'s. The likelihood function of the observed is then

$$L(\phi_H) = \prod_{g \in G} (\phi^g)^{n\hat{\phi}^g} \tag{1}$$

where $\phi^g = \sum_{h+\bar{h}=g} \phi_h \phi_{\bar{h}}$. The estimated $\phi_H$ is a maximum of $L$ subject to the constraints that $\sum_{h \in H} \phi_h = 1$ and $\phi_h \geq 0$, $\forall h \in H$.

There is a great deal of literature on the maximization of this polynomial, for example the method of *Expectation Maximization* is a linearly convergent method guaranteed to locate a local maximum of $L$ from almost every (feasible) starting point.

However, a naïve implementation of the EM method requires exponential space, since there are $2^m$ unknown haplotype probabilities which must be stored for $m$ variant sites. One notes note that, for $n$ sampled individuals, $\Omega(n)$ haplotypes are expected to have significant probability. An efficient way to discover those haplotypes which contribute significantly to the maximizer of $L$ would make this approach much more efficient.

*Problem 4 (Haplotype Support Problem).* Given observed genotype frequencies $\phi^g$, and $\epsilon > 0$, find $H' \subset H$, such that one can guarantee that there exists a $\phi_H$ that is a global maximizer of $L$ and that $h \notin H'$ implies $\phi_h < \epsilon$.

**The Phasing Polynomial.** We will now give a combinatorial interpretation of $L$. We assume that $\phi^G$ comes from counts of individual observed genotypes,

and thus $n\phi^g$ is integral for each genotype $g$. We may then formulate $L$ in terms of a product over $n$ observed individual genotypes $g_i$ $(1 \leq i \leq n)$, i.e.

$$L = \prod_{i=1}^{n} \phi^{g_i} = \prod_{i=1}^{n} \Big( \sum_{h+\bar{h}=g_i} \phi_h \phi_{\bar{h}} \Big)$$

Interchanging product and summation this becomes

$$L = \sum_{h_1,h_2,\cdots h_{2n}} \delta^{g_1}_{h_1+h_2} \delta^{g_2}_{h_3+h_4} \cdots \delta^{g_n}_{h_{2n-1}+h_{2n}} \phi_{h_1} \phi_{h_2} \cdots \phi_{h_{2n}}$$

Let an *explanation* of the genotypes $\boldsymbol{g} = (g_1, \ldots, g_n)$ be a sequence of $2n$ haplotypes $\boldsymbol{h} = (h_1, h_2, \ldots, h_{2n})$ such that $h_{2i-1} + h_{2i} = g_i$. Then the polynomial above can be more compactly expressed as

$$L = \sum_{\boldsymbol{h} \text{ explains } \boldsymbol{g}} \phi_{h_1} \phi_{h_2} \cdots \phi_{h_{2n}}$$

with the sum ranging over all explanations of $\boldsymbol{g}$. The likelihood function is a polynomial with a term of coefficient 1 for each possible explanation of the observed genotypes. Thus, a solution to the genotype phasing problem corresponds to a particular term in this polynomial.

The maximum likelihood approach seeks frequencies $\phi_H$ which maximize $L$. This problem is known to be NP-hard [29]. Also note that the problem does not directly address the problem of computing the individual phasings for each genotype. However, approximations can be made which recover the combinatorial nature of the phasing problem.

**A Discrete Approximation.** Let us collect the terms of $L$, and use a multi-index $\boldsymbol{P}$ (a vector of non-negative integers indexed by $H$) to keep track of the exponents, then

$$L = \sum_{\boldsymbol{P}} K(\boldsymbol{P}, \boldsymbol{g}) \prod_{h \in H} (\phi_h)^{P_h},$$

where $K(\boldsymbol{P}, \boldsymbol{g})$ denotes the number of explanations of the observed genotype counts $\boldsymbol{g}$ which have $\boldsymbol{P}$ haplotype counts.

Since the $\phi_h$ are constrained to lie between 0 and 1, most of the terms in $L$ are expected to be small. We may approximate $L$ with its largest term:

$$L \sim L_{MAX} = \max_{\boldsymbol{P}} \left\{ K(\boldsymbol{P}, \boldsymbol{g}) \prod_{h \in H} (\phi_h)^{P_h} \right\}.$$

The maximization of $L_{MAX}$ with respect to the $\phi_h$ is trivial, since any monomial $\prod_{h \in H} (\phi_h)^{P_h}$ in probabilities, $\phi_h$, is maximized by $\phi_h = P_h/(2n)$. Thus

$$\max_{\phi_H} L \sim L_{MAX} = \max_{\boldsymbol{P}} \left\{ (2n)^{-2n} K(\boldsymbol{P}, \boldsymbol{g}) \prod_{h} (P_h)^{P_h} \right\}.$$

Thus, we see that the maximization of the maximal term of the likelihood polynomial reduces to a discrete problem. The solution of this problem does not give a phasing, but a collection of possible phasings with identical counts. The solution may also be a good initial point for an iterative maximum likelihood method, such as expectation maximization.

The objective function in this optimization problem is

$$F(\boldsymbol{P}, N\hat{g}^k) = K(\boldsymbol{P}, G) \prod_h (P_h)^{P_h},$$

where $\sum_i P_i = 2N$, which counts the number of ways to select $2N$ haplotypes from a bag with counts $\boldsymbol{P}$ *with replacement* to form an explanation of the genotypes $G$.

We are not aware of any results about the complexity of evaluating $F$ or its maximum. In fact, there is a feasibility problem to which we have found no easy answer as well.

*Problem 5 (Haplotype Count Feasibility).* Given genotypes $\boldsymbol{g} = (g_1, \ldots, g_n)$ and a vector of counts $\boldsymbol{P}$ over $H$, decide whether there exists an explanation of $\boldsymbol{g}$ with counts $\boldsymbol{P}$.

*Problem 6 (Counting Arrangements $K(\boldsymbol{P}, \boldsymbol{g})$).* Given genotypes $\boldsymbol{g} = (g_1, \ldots, g_n)$ and a vector of counts $\boldsymbol{P}$, count how many distinct explanations, $\boldsymbol{h} = (h_1, h_2, \ldots, h_{2n-1}, h_{2n})$, exist for $\boldsymbol{g}$ with counts $\boldsymbol{P}$.

*Problem 7 (Maximizing Arrangements).* Given $\boldsymbol{g} = (g_1, \ldots, g_n)$, find counts $\boldsymbol{P}$, such that $K(\boldsymbol{P}, \boldsymbol{g}) \prod (P_h)^{P_h}$ is maximized.

**Links to Clark's Rule.** One method for breaking ties in the application Clark's rule is to allow the haplotype frequencies to serve as probabilities, and randomly select $g$'s and $h$'s to which to apply it. In such a scheme, one would still resolve the homozygotes and the single-site heterozygotes, since they are unambiguous, but, when faced with a choice between multiple phasings, one randomly selects the phasing $h, \bar{h}$ with probability $\phi_h \phi_{\bar{h}} / \phi^{h+\bar{h}}$. Since this procedure is still strongly dependent on the order of consideration for the ambiguous genotypes, one draws them, and re-draws them, uniformly at random, from the sampled individuals.

This process can be viewed as a means to generate a sample from a stationary point of the maximum likelihood function. To see this, we view the individual samples as all having some phase, which we rephase through random applications of Clark's rule with random tie-breaking as above. In the continuum limit, new instances of haplotype $h$ are introduced at a rate

$$\Delta\phi_h = \sum_{g, \bar{h}: h+\bar{h}=g} \frac{\hat{\phi}^g}{\phi^g} \phi_h \phi_{\bar{h}} \tag{2}$$

where $\phi^g = \sum_{h+\bar{h}=g} \phi_h \phi_{\bar{h}}$, while instances of haplotype $h$ are being removed (by individuals with haplotype $h$ being re-drawn) at a rate

$$\Delta\phi_h = -\phi_h.$$

A steady state occurs when the two processes balance, i.e.

$$\phi_h = \sum_{g,\bar{h}:h+\bar{h}=g} \frac{\hat{\phi}^g}{\phi^g} \phi_h \phi_{\bar{h}}$$

which is a sufficient condition for $\phi_H$ to be a local maximum of the likelihood function of equation 1. Thus, the haplotypes sampled in this random application of Clark's rules are distributed according to some stationary distribution of $L$.

### 2.3 Clark's Rule and Population Models

The observation that the maximum likelihood method could be modeled by a certain probabilistic application of Clark's rule was known to researchers, Stephens, Smith, and Donnelly [54], who proposed a modification of the ML sampling procedure of the previous section. Their modification introduces an approximate population genetics model [53] as a prior for observing the set of haplotypes.

Instead of phasing randomly selected individuals with probabilities weighted by $\phi_h \phi_{\bar{h}}$, they proposed a more complicated probability rule, where the weight of phasing $h, \bar{h}$ for $g$ is given by

$$\delta^g_{h+\bar{h}} \pi_h(\boldsymbol{h}\backslash h) \cdot \pi_{\bar{h}}(\boldsymbol{h}\backslash h\backslash \bar{h}) \tag{3}$$

where $\boldsymbol{h}\backslash h$ is the sequence of haplotypes $\boldsymbol{h}$ with one occurrence of $h$ removed. The function $\pi_h(\boldsymbol{h})$ approximates the probability that haplotype $h$ might be generated either by direct descent or mutation from a population with haplotype counts of $\boldsymbol{h}$.

It should be noted that equation 2 applies only when $N$ is much larger than the number of haplotype variants in the population. Thus it is not strictly applicable for small populations where a substantial portion of variants occur only once. It is not an issue for this Markov Chain Monte Carlo (MCMC) approach.

The algorithm they propose is to iteratively modify the explanation of the given genotypes, selecting the explaining haplotypes $h, \bar{h}$ for a random individual with genotype $g$, and replacing that pair with a pair generated randomly with weights from equation 3, updating the current frequencies, $\phi_H$, of the variants in the sample. Statistics of the sampled set of phasings are then used to select the phasings of the individuals.

It remains to define the approximation of $\pi_h(\boldsymbol{h})$, for which they propose

$$\pi_h(\boldsymbol{h}) = \sum_{h'} (h \cdot \boldsymbol{h})(I - \frac{\theta}{2N+\theta} M)^{-1}_{hh'} \tag{4}$$

where $h \cdot \boldsymbol{h}$ counts the number of occurrences of $h$ in $\boldsymbol{h}$, $\theta$ is an estimate for the per site mutation rate, $I$ is the identity matrix, and $M$ is the single site mutation

matrix, $M_{hh'} = 1$ iff $h$ and $h'$ have exactly one mismatch and 0 otherwise. This is an approximation to the probability that $h$ can come from some $h'$ after a geometrically distributed number of single site mutations. This approximation arose from considering a random population model in [32]. It should be noted that while the matrix $M$ appears to be of exponential size, an arbitrary element of $(I - \frac{\theta}{2N+\theta}M)^{-1}$ can be computed in $O(m)$ time.

An implementation of this algorithm by Stephens, Smith, and Donnelly is *PHASE*. An alternative algorithm, which more closely follows the maximum likelihood method was produced by Niu et al. [44]. *PHASE* works well on medium problems with a small population.

### 2.4 Parsimony Formulations

Extending Clark's basic intuition that unresolved haplotypes are to look like known ones, a variety of parsimony objectives can be considered.

In the context of haplotype phasing, the most *parsimonious* phasing refers to the solution that uses the fewest haplotypes. Hubbell [30] showed that this version of the problem is NP-hard, in general, by a reduction from minimum clique cover. Gusfield [24] solved the problem via an (exponentially large) integer programming formulation that is solvable in many cases, even for medium-sized problems. An intriguing open problem is to determine whether there are practical instances when this problem can be solved efficiently (for example if the perfect phylogeny condition holds, see section 2.5).

*Problem 8 (Restricted Parsimony).* Find a restriction on the input to the haplotype phasing problem that most real world instances satisfy, for which the most parsimonious haplotype assignment can be found in polynomial time.

Diversity is another commonly used parsimony objective in population genetics. Haplotype diversity is defined as the probability that two haplotypes drawn uniformly at random from the population are not the same.

*Problem 9 (Haplotype Diversity Minimization).* Devise an algorithm for the haplotype phasing under the objective of minimizing haplotype diversity.

Graph theoretically, this problem can be posed as constructing a graph with a node for every haplotype in the observed population (two nodes for each observed genotype), an edge between every pair of haplotypes that are not equal and then minimizing the number of edges in the graph.

We observe that Clark's rule is not effective for parsimony.

**Lemma 1.** *Clark's rule does not yield an effective approximation algorithm for parsimony.*

Let $n_d$ be the number of distinct genotypes in the $G$. The trivial algorithm of arbitrarily phasing each distinct genotype will return a phasing with at most $2n_d$ haplotypes. $\Omega(\sqrt{n_d})$ is a lower bound on the number of haplotypes as each

$$\left\{\begin{array}{l}(1111)+(1111)\\(1111)+(0011)\\(1111)+(1001)\\(1111)+(1100)\\(1111)+(1010)\\(1111)+(0101)\\(1111)+(0110)\end{array}\right\} \overset{\mathcal{P}_C}{\Longleftarrow} \left\{\begin{array}{l}(1111)\\(2211)\\(1221)\\(1122)\\(1212)\\(2121)\\(2112)\end{array}\right\} \overset{\mathcal{P}_P}{\Longrightarrow} \left\{\begin{array}{l}(1111)+(1111)\\(0111)+(1011)\\(1011)+(1101)\\(1101)+(1110)\\(1011)+(1110)\\(0111)+(1101)\\(0111)+(1110)\end{array}\right\}$$

**Fig. 3.** Set of 7 genotypes with 7 haplotype Clark's rule resolution $\mathcal{P}_C$, and 4 haplotype parsimony resolution $\mathcal{P}_P$.

genotype is made of at most two distinct haplotypes. A worst case approximation guarantee is thus $O(\sqrt{n_d})$, we will give such an example.

Let $m$ be the number of SNPs and let $G$ be comprised of genotype that has all ones and all $\binom{m}{2}$ possible genotypes that have exactly two 2s and all other SNPs as 1s. Clark's inference rule will initially infer the haplotype of all ones and then infer the $\binom{m}{2}$ haplotypes that have all but 2 SNPs as 1s. The resolution with the minimum number of haplotypes however has the $m$ haplotypes with all but 1 SNP as 1. An example when $m = 4$ is given in Figure 3. □

The Hamming distance between a pair of haplotypes is, under the infinite sites model, the number of mutations that occurred in the evolutionary history between the pair of haplotypes. If we consider an evolutionary history to be a tree whose nodes are the unknown haplotype sequences of the observed genotype sequences, then a likelihood function which approximates it [45] in terms of Hamming distance is given by:

$$L(\boldsymbol{h}) \propto \sum_{T} \prod_{e \in \text{Edges}(T)} f(D(e)) \tag{5}$$

where $T$ ranges over all trees on the $2n$ nodes with unknown haplotypes $h_i \in H$, $1 \leq i \leq 2n$, $e$ ranges over all $2n-1$ edges in $T$, $D(e)$ is the Hamming distance between the $h_i$ and $h_j$ which are joined by $e$, and $f$ is a monotonic function. One reasonable choice might by $f(x) = e^{-\beta x}$ where $\beta$ plays the role of the mutation rate, or one might take $f$ from equation 4.

This sum over all trees of products of edge weights can be evaluated in polynomial time (using Kirchoff's matrix-tree theorem [9,35]). Methods for sampling from this and related distributions can be found in [8,13].

If we take $f(x) = e^{-\beta x}$, then we can interpret equation 5 as a partition function from statistical mechanics,

$$Z(\boldsymbol{h}; \beta) = \sum_{T} e^{-\beta E(T, \boldsymbol{h})}$$

where $E(T, \boldsymbol{h})$ is the sum of the Hamming distances on all the edges in $T$.

*Problem 10 (Partition Function Maximization).* Devise an algorithm which maximizes

$$Z(\boldsymbol{h};\beta) = \sum_{T} e^{-\beta E(T,\boldsymbol{h})} \tag{6}$$

over all $\boldsymbol{h}$ explaining $\boldsymbol{g}$.

This problem has two asymptotic regimes.

The first is the *low temperature* regime $\beta \to \infty$, where, one can approximate the summation with maximization,

$$\begin{aligned} Z(\boldsymbol{h};\beta \sim \infty) &\sim \max_{T} e^{-\beta E(T,\boldsymbol{h})} \\ &= \exp\{-\beta \min_{T} E(T,\boldsymbol{h})\} \end{aligned}$$

and approximate the partition function with the minimum weight tree.

*Problem 11 (Tree Minimization).* Devise an algorithm which finds

$$\min_{T,\boldsymbol{h}} E(T,\boldsymbol{h}) \tag{7}$$

over all $\boldsymbol{h}$ explaining $\boldsymbol{g}$ and all trees $T$.

The second is the *high temperature* regime $\beta \sim 0$

$$Z(\boldsymbol{h};\beta) \sim \sum_{T}(1-\beta E(T,\boldsymbol{h})) = (2n)^{2n-2}(1-\frac{1}{2n}\sum_{h_1,h_2\in\boldsymbol{h}} D(h_1,h_2))$$

where $D(h_1,h_2)$ is the Hamming distance between $h_1$ and $h_2$. In this extreme, the approximate problem is the minimization of the sum of all pairwise Hamming distances.

*Problem 12 (Sum of Pairs Hamming Distance Minimization).* Devise an algorithm which finds

$$\min_{\boldsymbol{h}} \sum_{h_1,h_2\in\boldsymbol{h}} D(h_1,h_2) \tag{8}$$

over all $\boldsymbol{h}$ explaining $\boldsymbol{g}$ and all trees $T$.

Figure 4 gives an example where the sum of pairs Hamming distance minimization does not yield the same phasing parsimony.

At the time of this writing, we are not familiar with any progress on these problems.

$$\left\{\begin{array}{l}(11111111)+(11111100)\\(11111111)+(11111001)\\(11111111)+(11110011)\\(11111111)+(11001111)\\(11111111)+(10011111)\\(11111111)+(00111111)\end{array}\right\} \underset{\Longleftarrow}{\mathcal{P}_P} \left\{\begin{array}{l}(11111122)\\(11111221)\\(11112211)\\(11221111)\\(12211111)\\(22111111)\end{array}\right\} \underset{\Longrightarrow}{\mathcal{P}_H} \left\{\begin{array}{l}(11111101)+(11111110)\\(11111011)+(11111101)\\(11110111)+(11111011)\\(11011111)+(11101111)\\(10111111)+(11011111)\\(01111111)+(10111111)\end{array}\right\}$$

**Fig. 4.** Set of 6 genotypes with 7 haplotype parsimony phasing $\mathcal{P}_P$, and 8 haplotype minimum sum of paired Hamming distances phasing $\mathcal{P}_H$.

### 2.5 Perfect Phylogeny

The concept of a *perfect phylogeny* [15,51,5] has also been used to formulate constraints on haplotype phasings. A (binary) perfect phylogeny is defined as follows: Let $S$ be a set of $n$ sequences (haplotypes) each drawn from $\Sigma^m$, where the alphabet $\Sigma = \{0, 1\}$. We say that $S$ admits a *perfect phylogeny* if there exists a tree $T$ with $n$ leaves that has the following properties: (1) Each leaf of $T$ is uniquely labeled with a sequence from $S$, (2) Every internal node $v$ in $T$ is labeled with a sequence from $\Sigma^m$, and (3) For each sequence position $i$ (where $1 \leq i \leq m$) and for each $a \in \Sigma$, the set of nodes whose sequence labels each have the symbol $a$ at position $i$, forms a subtree of $T$. The tree $T$ is said to be a perfect phylogeny for $S$.

Gusfield [23] introduced a haplotype phasing problem that was motivated by studies on the haplotype structure of the human genome that reveal the genome to be *blocky* in nature ([14,28,52,18]), i.e., these studies show that human genomic DNA can be partitioned into long blocks where genetic recombination has been rare, leading to strikingly fewer distinct haplotypes in the population than previously expected. This *no-recombination in long blocks* observation together with the standard population genetic assumption of infinite sites, motivates a model of haplotype evolution where the haplotypes in a population are assumed to evolve along a coalescent, which as a rooted tree is a *perfect phylogeny*. Informally, this means that each SNP changed from a 0 to a 1 at most once in this rooted tree (here we are assuming that 0 is the ancestral state for a SNP). This motivates the following algorithmic problem called Perfect Phylogeny Haplotyping problem (PPH) - given $n$ genotypes of length $m$ each, does there exist a set $S$ of at most $2n$ haplotypes such that each genotype is explained by a pair of haplotypes from $S$, and such that $S$ admits a perfect phylogeny?

In [23], it was shown that the PPH problem can be solved in polynomial time by reducing it to a graph realization problem. The algorithm runs in $O(nm\alpha(nm))$, where $\alpha$ is the inverse Ackerman function, and hence this time bound is almost linear in the input size $nm$. The algorithm also builds a linear-space data structure that represents all the solutions, so that each solution can be generated in linear time. Although the reduction described in [23] is simple and the total running time is nearly optimal, the algorithm taken as a whole is very

difficult to implement, primarily due to the complexity of the graph realization component.

Following the work in [23], additional algorithms [3,16] have been proposed to solve the PPH problem that are simpler, easy to program and yet still efficient. These algorithms also produce linear-space data structures to represent all solutions for the given instance. Though they use quite different approaches, the algorithms in [3] and [16] take $O(nm^2)$ time. In [3], a non-trivial upper bound on the number of PPH solutions is also proved, showing that the number is vastly smaller than the number of haplotype solutions when the perfect phylogeny requirement is not imposed; furthermore, a biologically appealing representation is proposed that aids in visualizing the set of all solutions. In [16], an approach is also provided to deal with parent-child genotype data.

There are several interesting questions posed as a result of the works of [23,3,16]. We list three of them here.

*Problem 13 (Optimal PPH).* Can the PPH problem be solved in $O(nm)$? If so, is a practical algorithm possible?

*Problem 14 (PPH with Missing Data).* Devise solutions to deal with missing data and errors in the input.

The above problem is important as real data, very often, contains both missing data and errors. There are several directions that one could pursue here. For example, one could ask the question, can each missing value be set to one of $0, 1$, or 2 so that the resulting instance has a perfect phylogeny? Alternatively, as in [16], one could study the complexity of the problem of removing a minimum number of genotypes so that the phasing of the remaining genotypes admits a perfect phylogeny.

*Problem 15 (PPH with Recombination).* What is the complexity of the problem when small deviations from the *no-recombination* model are allowed? For instance, allowing for a small number of recombination events, can we still phase the genotypes efficiently in this framework? Allowing recombination events means that the solution is no longer a tree but a network (i.e. a graph with cycles) [57].

## 3 Haplotype Assembly

The need to infer haplotypes directly from genotypes is based on the assumption that biotechnologies for haplotype determination are unlikely to be available in the short term. This may not be the case. Various approaches to single molecule sequencing have been described recently [42,43,10] and some of these may mature to the point that phasing based solely on genotype analysis becomes unnecessary.

An increase in the current read length ($\sim 500$) in a sequencing reaction to a few thousand basepairs, make it possible to phase large regions of a chromosome. Assuming that a SNP occurs every 1000 basepairs, many fragments will contain

multiple SNPs. Consider a sequence assembly containing fragments $f_1, f_2, f_3$ from a single individual. If $f_1$ and $f_2$ differ in a SNP, they must come from different chromosomes. Likewise if $f_2$, and $f_3$ also differ in (some other) SNP, they come from different chromosomes. However, for a diploid organism, this must imply that $f_1$ and $f_3$ come from the same chromosome, and we have therefore *phased* the individual in this region (see Figure 3). Even current technology can produce reads of over 1000 basepairs, by creating a gapped read, where only the ends of the fragment are actually read, leaving a large gap in the middle.

Formally, define a SNP matrix M with rows representing fragments, and columns representing SNPs. Thus $M[f, s] \in \{0, 1, -\}$ is the value of SNP $s$ in fragments $f$. Gapped reads are modeled as single fragments with gaps $(-)$ in SNPs that in the gap. Two fragments $f$ and $g$ conflict if there exists SNP $s$ such that $M[f, s] = 0$, and $M[g, s] = 1$ or vice-versa. Based on this, a SNP matrix $M$ can be used to define a fragment conflict graph $G_{\mathcal{F}}$. Each fragment is a node in $G_{\mathcal{F}}$. Two nodes are connected by an edge if they have different values at a SNP. It is easy to see that $G_{\mathcal{F}}$ is bipartite in the absence of errors. In the presence of errors, we can formulate combinatorial problems that involve deleting a minimum number of nodes (poor quality fragments), or edges (bad SNP calls), so that the resulting graph is bipartite (can be phased trivially).
In [48,37], the following is shown:

1. The minimum fragment removal and minimum SNP removal problems are tractable if the underlying matrix $M$ has the consecutive ones property, i.e., there is no gap within a fragment.
2. They are NP-hard in the case of gaps, even when limited to at most one gap per fragment. The problems are shown to be tractable under a fixed parameter. That parameter being the total length of gaps in a fragment.

The algorithms are thus not tractable for dealing with the case of fragments with gaps, and it is an interesting open problem to design heuristics/approximations that give good results in practice. Some branch and bound heuristics were reported to work very well on real and simulated assembly data in [40]. Li et al. [38] give a statistical reformulation of this problem.

A fairly immediate extension to this problem, is the problem of simultaneous assembly multiple haplotypes. This will occur when studying multiploidal or when simultaneously assembling related organisms. For practical consideration it may be easier to sequence multiple related organisms simultaneously, for example to assemble different strains of a bacteria simultaneously.

*Problem 16 (Multiploidal Haplotype Assembly).* Devise an algorithm for assembling multiple haplotypes simultaneously.

## 4 Haplotype Block Detection

The haplotype block conjecture is that the genome consists of regions of relatively low recombination rate, between which there is relatively high rate of recombination.


**Fig. 5.** An illustration of the construction of long-range haplotypes from assembly data. A) the fragment assembly of a diploid organism. B) the identification of SNPs. C) the partitioning of the fragments into two consistent groups, introducing a long-range phasing

Several methods have been suggested for detecting recombination. The first and the most basic method was suggest by Hudson and Kaplan [33] who showed that under the infinite sites model, if all four possible values that a pair of SNPs can take are observed then recombination between the pair is implied. A drawback of this model is its lack of robustness, a genotyping error or a violation of the infinite sites model may imply falsely detecting a recombination.

A variety of alternate block detection methods have been suggested [33,46,19], which give similar and statistically concordant, but not same blocks [50]. None of the block detection methods has been shown to be consistent, i.e. will converge to the "true" answer given sufficient data.

*Problem 17 (Block Detection).* Devise a consistent algorithm for detecting haplotype blocks.

Given a haplotype block detection method, the problem of finding a partitioning of a chromosome into blocks arises. This problem has been shown to be solvable using dynamic programming [59,49]. Given a haplotype block partition it is natural to ask about the evolutionary history of the chromosomes. Working concurrently Schwartz et al. and Ukkonen [49,55] gave an efficient algorithm for the haplotype block coloring problem, partitioning a chromosome into blocks and coloring sequences to signify likely ancestral sequences of each segment.

In the presence of recombination events, the evolutionary history of a set of haplotypes can be modeled using an *Ancestral Recombination Graph* (ARG) [20,31]. An ARG for a set $S$ of haplotypes is a directed graph $G = (V, E)$. A node $v \in V$ represents a haplotype (either ancestral or extant); an edge $(u, v) \in E$ indicates that $u$ is a parent of $v$ (equivalently, $v$ is a child of $u$). $G$ has a special node designated as the root node $r$. The leaves of $G$ are in bijection with the haplotypes in $S$. Node $u$ is an *ancestor* of node $v$ (equivalently $v$ is a *descendant* of $u$) if $u$ is on the path from $r$ to $v$. Graph $G$ has the additional property that each node (except for $r$) can have either one or two parents; if a node $x$ has two parents $y$ and $z$, then it is the case that $y$ is not an ancestor of $z$ and $z$ is not an ancestor of $y$.

ARGs have been used in population genetics to study mutation rates, recombination rates, and time to the Most Recent Common Ancestor (MRCA) [20,45].

*Problem 18 (ARG Reconstruction).* Devise algorithms for inferring ARGs under realistic models.

For instance, assuming the infinite-sites model leads to the Perfect Phylogenetic Network inference problem [57].

The ARG inference problem can also be viewed as a problem of combining trees to produce a graph under some optimization criterion. This is motivated by the concept of haplotype blocks. Given a set of haplotypes, one could use some block detection algorithm to infer blocks [33,46,19], and then construct an evolutionary tree for each block; these evolutionary trees could then be combined to produce a graph. For instance, we could insist on a graph containing the minimum number of edges such that each tree is a subgraph of this graph.

## 5 SNP Selection

The problem of selecting informative subsets of SNPs is directly and indirectly related to the haplotype phasing problem. Closely spaced SNPs are generally highly correlated and hence a small number of SNPs is usually sufficient for characterizing a haplotype block. The haplotype tagging approach [12,2] makes use of this fact and starts by partitioning a chromosome into a set of blocks and tagging SNPs are then selected within each block. The selection SNPs within each haplotype block can be relaxed to the test cover problem [25,7,6]. Although this problem is NP-hard in the number of SNPs, blocks, in practice, contain only a few SNPs, making the problem tractable.

Since a An extra level of uncertainty is added to the SNP selection problem by first partitioning a chromosome into blocks as a chromosome cannot consistently be partitioned into haplotype blocks [50]. An alternate approach is taken in [4], where a measure for selecting an informative subset of SNPs in a block free model is developed. The general version of this problem is NP-hard, but there exist efficient algorithms for two important special cases of this problem. For each SNP a set of *predictive* SNPs are defined. For the case when the SNPs can be ordered such that the distance (measured in number of SNPs) between a SNP and its predictive SNP is bounded by a constant, $w$, a $O(2^{2w}nm)$ algorithm is given. In the case when the predictive set contains only SNPs that pairwise obey the predictive set obeys a perfect phylogeny condition, it is shown that a single SNP can be predicted using an efficient algorithm. The problem of predicting a set of SNPs when the predictive set obeys a perfect phylogeny condition is open.

*Problem 19 (Perfect Phylogeny SNP Detection).* Devise an efficient algorithm (or show that one does not exist) for predicting a set of SNPs when the predictive set obeys a perfect phylogeny condition.

Both of the above mentioned approaches assume that SNPs are selected from haplotypes. This requires that either the haplotypes are known or the haplotypes have been inferred. Inferring the haplotypes first necessarily creates an extra level of uncertainty to the problem. A more interesting approach would be to select the SNPs directly from the genotypes.

*Problem 20 (Genotype SNP Selection).* Devise an algorithm for selecting informative SNPs directly from genotype data.

A simplistic method for solving this problem is described in [56].

As the endgoal is generally to use the SNPs for detecting disease an objective function for selecting the SNPs should consider how likely it is that we will detect a SNP given a model for how the disease is likely to be found in the population. A common assumption is the common disease/common variant model for disease [47], this model however is currently under dispute in the populations genetics community [58].

## 6 Discussion

While the subject of haplotype phasing, or frequency inference may be of interest on purely statistical and mathematical grounds, the desired end result generally is an understanding of the implications of genetic variation in individual pathology, development, etc. As such, these variances are one part of a larger set of interactions which include individual environment, history, and chance.

Although the media often carries stories of "genes" (actually alleles) being found for some popular diseases, biologists agree that such stories are rather the exception than the rule when it comes to disease causation. It is suspected to be more the case that an individual's genetic variations interact with each other as well as other factors in excruciatingly complex and sensitive ways.

The future open problems of SNP analysis are those regarding the interactions of genetic variation with external factors. Substantial progress in multifactorial testing, or a tailoring of current multifactorial testing to this setting, is required if we are to see an impact of haplotype analysis on human health care.

**Acknowledgments**

We would like to thank Mark Adams, Sam Broder, Michelle Cargill, Andy Clark, Francis Kalush, Giuseppe Lancia, Kit Lau, Russell Schwartz, Francisco de la Vega, and Mike Waterman for many valuable discussions on this subject.

## References

1. G. R. Abecasis, S. S. Cherny, W. O. Cookson, and L. R. Cardon. Merlin — rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, 30(1):97–101, 2002.
2. H. I. Avi-Itzhak, X. Su, and F. M. De La Vega. Selection of minimum subsets of single nucleotide polymorphisms to capture haplotype block diversity. In *Proceedings of Pacific Symposium on Biocomputing*, volume 8, pages 466–477, 2003.
3. V. Bafna, D. Gusfield, G. Lancia, and S. Yooseph. Haplotyping as a perfect phylogeny. A direct approach. *Journal of Computational Biology*, 2003. To appear.
4. V. Bafna, B. V. Halldórsson, R. S. Schwartz, A. G. Clark, and S. Istrail. Haplotypes and informative SNP selection algorithms: Don't block out information. In *Proceedings of the Seventh Annual International Conference on Computational Molecular Biology (RECOMB)*, 2003. To appear.
5. H. Bodlaender, M. Fellows, and T. Warnow. Two strikes against perfect phylogeny. In *Proceedings of the 19th International Colloquium on Automata, Languages, and Programming (ICALP)*, Lecture Notes in Computer Science, pages 273–283. Springer Verlag, 1992.
6. K. M. J. De Bontridder, B. V. Halldórsson, M. M. Halldórsson, C. A. J. Hurkens, J. K. Lenstra, R. Ravi, and L. Stougie. Approximation algorithms for the minimum test cover problem. *Mathematical Programming-B*, 2003. To appear.
7. K. M. J. De Bontridder, B. J. Lageweg, J. K. Lenstra, J. B. Orlin, and L. Stougie. Branch-and-bound algorithms for the test cover problem. In *Proceedings of the Tenth Annual European Symposium on Algorithms (ESA)*, pages 223–233, 2002.
8. A. Broder. Generating random spanning trees. In *Proceedings of the IEEE 30th Annual Symposium on Foundations of Computer Science*, pages 442–447, 1989.
9. S. Chaiken. A combinatorial proof of the all-minors matrix tree theorem. *SIAM Journal on Algebraic and Discrete Methods*, 3:319–329, 1982.
10. E. Y. Chen. Methods and products for analyzing polymers. U.S. Patent 6,355,420.
11. A. G. Clark. Inference of haplotypes from PCR-amplified samples of diploid populations. *Molecular Biology and Evolution*, 7(2):111–122, 1990.
12. D. Clayton. Choosing a set of haplotype tagging SNPs from a larger set of diallelic loci. *Nature Genetics*, 29(2), 2001. URL: `www.nature.com/ng/journal/v29/n2/extref/ng1001-233-S10.pdf`.
13. H. Cohn, R. Pemantle, and J. Propp. Generating a random sink-free orientation in quadratic time. *Electronic Journal of Combinatorics*, 9(1), 2002.

14. M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson, and E. S. Lander. High-resolution haplotype structure in the human genome. *Nature Genetics*, 29:229–232, 2001.
15. W. H. E. Day and D. Sankoff. Computational complexity of inferring phylogenies by compatibility. *Systematic Zoology*, 35(2):224–229, 1986.
16. E. Eskin, E. Halperin, and R. M. Karp. Efficient reconstruction of haplotype structure via perfect phylogeny. Technical report, Columbia University Department of Computer Science, 2002. URL: `http://www.cs.columbia.edu/compbio/hap`. Update of UCB technical report with the same title.
17. L. Excoffier and M. Slatkin. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution*, 12(5):921–927, 1995.
18. L. Frisse, R. Hudson, A. Bartoszewicz, J. Wall, T. Donfalk, and A. Di Rienzo. Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *American Journal of Human Genetics*, 69:831–843, 2001.
19. 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. Altschuler. The structure of haplotype blocks in the human genome. *Science*, 296(5576):2225–2229, 2002.
20. R. C. Griffiths and P. Marjoram. Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology*, 3(4):479–502, 1996.
21. D. Gusfield. A practical algorithm for optimal inference of haplotypes from diploid populations. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB)*, pages 183–189, 2000.
22. D. Gusfield. Inference of haplotypes from samples of diploid populations: Complexity and algorithms. *Journal of Computational Biology*, 8(3):305–324, 2001.
23. D. Gusfield. Haplotyping as perfect phylogeny: Conceptual framework and efficient solutions (Extended abstract). In *Proceedings of the Sixth Annual International Conference on Computational Molecular Biology (RECOMB)*, pages 166–175, 2002.
24. D. Gusfield. Haplotyping by pure parsimony. In *Proceedings of the 2003 Combinatorial Pattern Matching Conference*, 2003. To appear.
25. B. V. Halldórsson, M. M. Halldórsson, and R. Ravi. On the approximability of the minimum test collection problem. In *Proceedings of the Ninth Annual European Symposium on Algorithms (ESA)*, pages 158–169, 2001.
26. D. L. Hartl and A. G. Clark. *Principles of Population Genetics*. Sinauer Associates, 1997.
27. M. E. Hawley and K. K. Kidd. HAPLO: A program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *Journal of Heredity*, 86:409–411, 1995.
28. L. Helmuth. Genome research: Map of the human genome 3.0. *Science*, 293(5530):583–585, 2001.
29. E. Hubbell. Finding a maximum likelihood solution to haplotype phases is difficult. Personal communication.
30. E. Hubbell. Finding a parsimony solution to haplotype phase is NP-hard. Personal communication.
31. R. R. Hudson. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23:183–201, 1983.

32. R. R. Hudson. Gene genealogies and the coalescent process. In D. Futuyma and J. Antonovics, editors, *Oxford surveys in evolutionary biology*, volume 7, pages 1–44. Oxford University Press, 1990.
33. R. R. Hudson and N. L. Kaplan. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, 111:147–164, 1985.
34. A. J. Jeffreys, L. Kauppi, and R. Neumann. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genetics*, 29(2):217–222, 2001.
35. G. Kirchhoff. Über die auflösung der gleichungen, auf welche man bei der untersuchung der linearen verteilung galvanischer ströme geführt wird. *Annalen für der Physik und der Chemie*, 72:497–508, 1847.
36. A. Kong, D. F. Gudbjartsson, J. Sainz, G. M. Jonsdottir, S. A. Gudjonsson, B. Richardsson, S. Sigurdardottir, J. Barnard, B. Hallbeck, G. Masson, A. Shlien, S. T. Palsson, M. L. Frigge, T. E. Thorgeirsson, J. R. Gulcher, and K. Stefansson. A high-resolution recombination map of the human genome. *Nature Genetics*, 31(3):241–247, 2002.
37. G. Lancia, V. Bafna, S. Istrail, R. Lippert, and R. Schwartz. SNPs problems, complexity and algorithms. In *Proceedings of the Ninth Annual European Symposium on Algorithms (ESA)*, pages 182–193, 2001.
38. L. Li, J. H. Kim, and M. S. Waterman. Haplotype reconstruction from SNP alignment. In *Proceedings of the Seventh Annual International Conference on Computational Molecular Biology (RECOMB)*, 2003. To appear.
39. S. Lin, D. J. Cutler, M. E. Zwick, and A. Chakravarti. Haplotype inference in random population samples. *American Journal of Human Genetics*, 71:1129–1137, 2002.
40. R. Lippert, R. Schwartz, G. Lancia, and S. Istrail. Algorithmic strategies for the single nucleotide polymorphism haplotype assembly problem. *Briefings in Bioinformatics*, 3(1):23–31, 2002.
41. J. C. Long, R. C. Williams, and M. Urbanek. An E-M algorithm and testing strategy for multiple-locus haplotypes. *American Journal of Human Genetics*, 56(2):799–810, 1995.
42. R. Mitra, V. Butty, J. Shendure, B. R. Williams, D. E. Housman, and G. M. Church. Digital genotyping and haplotyping with polymerase colonies. *Proceedings of the National Academy of Sciences*. To appear.
43. R. Mitra and G. M. Church. In situ localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Research*, 27(e34):1–6, 1999.
44. T. Niu, Z. S. Qin, X. Xu, and J. S. Liu. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *American Journal of Human Genetics*, 70:157–169, 2002.
45. M. Nordborg. *Handbook of Statistical Genetics*, chapter Coalescent Theory. John Wiley & Sons, Ltd, 2001.
46. N. Patil, A. J. Berno, D. A. Hinds, W. A. Barrett, J. M. Doshi, C. R. Hacker, C. R. Kautzer, D. H. Lee, C. Marjoribanks, D. P. McDonough, B. T. N. Nguyen, M. C. Norris, J. B. Sheehan, N. Shen, D. Stern, R. P. Stokowski, D. J. Thomas, M. O. Trulson, K. R. Vyas, K. A. Frazer, S. P. A. Fodor, and D. R. Cox. Blocks of limited haplotype diversity revealed by high resolution scanning of human chromosome 21. *Science*, 294:1719–1723, 2001.
47. D. E. Reich and E. S. Lander. On the allelic spectrum of human disease. *Trends in Genetics*, 17(9):502–510, 2001.

48. R. Rizzi, V. Bafna, S. Istrail, and G. Lancia. Practical algorithms and fixed-parameter tractability for the single individual SNP haplotyping problem. In *Proceedings of the Second International Workshop on Algorithms in Bioinformatics (WABI)*, pages 29–43, 2002.
49. R. S. Schwartz, A. G. Clark, and S. Istrail. Methods for inferring block-wise ancestral history from haploid sequences. In *Proceedings of the Second International Workshop on Algorithms in Bioinformatics (WABI)*, pages 44–59, 2002.
50. R. S. Schwartz, B. V. Halldórsson, V. Bafna, A. G. Clark, and S. Istrail. Robustness of inference of haplotype block structure. *Journal of Computational Biology*, 10(1):13–20, 2003.
51. M. A. Steel. The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of Classification*, 9:91–116, 1992.
52. J. C. Stephens, J. A. Schneider, D. A. Tanguay, J. Choi, T. Acharya, S. E. Stanley, R. Jiang, C. J. Messer, A. Chew, J.-H. Han, J. Duan, J. L. Carr, M. S. Lee, B. Koshy, A. M. Kumar, G. Zhang, W. R. Newell, A. Windemuth, C. Xu, T. S. Kalbfleisch, S. L. Shaner, K. Arnold, V. Schulz, C. M. Drysdale, K. Nandabalan, R. S. Judson, G. Ruano, and G. F. Vovis. Haplotype variation and linkage disequilibrium in 313 human genes. *Science*, 293(5529):489–493, 2001.
53. M. Stephens and P. Donnelly. Inference in molecular population genetics. *Journal of the Royal Statistical Society, Series B*, 62(4):605–635, 2000.
54. M. Stephens, N. J. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, 68:978–989, 2001.
55. E. Ukkonen. Finding founder sequences from a set of recombinants. In *Proceedings of the Second International Workshop on Algorithms in Bioinformatics (WABI)*, pages 277–286, 2002.
56. F. M. De La Vega, X. Su, H. Avi-Itzhak, B. V. Halldórsson, D. Gordon, A. Collins, R. A. Lippert, R. Schwartz, C. Scafe, Y. Wang, M. Laig-Webster, R. T. Koehler, J. Ziegle, L. Wogan, J. F. Stevens, K. M. Leinen, S. J. Olson, K. J. Guegler, X. You, L. Xu., H. G. Hemken, F. Kalush, A. G. Clark, S. Istrail, M. W. Hunkapiller, E. G. Spier, and D. A. Gilbert. The profile of linkage disequilibrium across human chromosomes 6, 21, and 22 in African-American and Caucasian populations. In preparation.
57. L. Wang, K. Zhang, and L. Zhang. Perfect phylogenetic networks with recombination. *Journal of Computational Biology*, 8(1):69–78, 2001.
58. K. M. Weiss and A. G. Clark. Linkage disequilibrium and the mapping of complex human traits. *Trends in Genetics*, 18(1):19–24, 2002.
59. K. Zhang, M. Deng, T. Chen, M. S. Waterman, and F. Sun. A dynamic programming algorithm for haplotype block partitioning. *Proceedings of the National Academy of Sciences*, 99(11):7335–7339, 2002.
60. P. Zhang, H. Sheng, A. Morabia, and T. C. Gilliam. Optimal step length EM algorithm (OSLEM) for the estimation of haplotype frequency and its application in lipoprotein lipase genotyping. *BMC Bioinformatics*, 4(3), 2003.