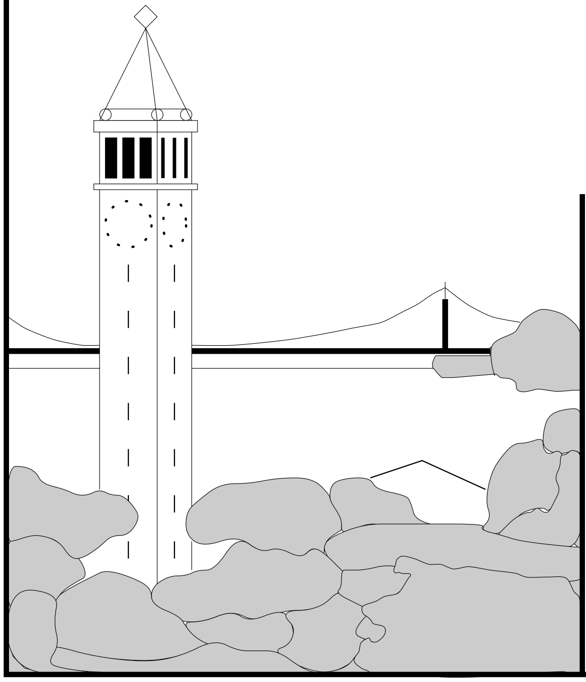
Large Scale Recovery of Haplotypes from Genotype Data using Imperfect Phylogeny

Eran Halperin *Eleazar Eskin*



Report No. UCB/CSD-2-1195

August 2002

Computer Science Division (EECS)
University of California
Berkeley, California 94720

Large Scale Recovery of Haplotypes from Genotype Data using Imperfect Phylogeny

Eran Halperin
CS Division
University of California Berkeley
eran@eecs.berkeley.edu

Eleazar Eskin
Department of Computer Science
Columbia University
eeskin@cs.columbia.edu

August 2002

Abstract

Critical to the understanding of the genetic basis for complex diseases is the modeling of human variation. Most of this variation can be characterized by single nucleotide polymorphisms (SNPs) which are mutations at a single nucleotide position. To characterize an individual's variation, we must determine an individual's *haplotype* or which nucleotide base occurs at each position of these common SNPs for each chromosome. In this paper, we present results for a highly accurate method for haplotype resolution from genotype data. Our method leverages a new insight into the underlying structure of haplotypes which shows that SNPs are organized in highly correlated "blocks". The majority of individuals have one of about four common haplotypes in each block. Our method partitions the SNPs into blocks and for each block, we predict the common haplotypes each individual's haplotype. We evaluate our method over biological data. Our method predicts the common haplotypes perfectly and has a very low error rate (0.47%) when taking into account the predictions for the uncommon haplotypes.

The algorithm is available via webserver at http://www.cs.columbia.edu/compbio/hap/[1].

Introduction

Critical to the understanding of the genetic basis for complex diseases is the modeling of human variation. Most of this variation can be characterized by single nucleotide polymorphisms (SNPs) which are mutations at a single nucleotide position that occurred once in human history and were passed on through heredity. Approximately 10 million common SNPs[14], each with a frequency of 10% to 50% account for the majority of the variation between DNA sequences of individuals[15]. To characterize an individual's variation, we must determine an individual's *haplotype* or which nucleotide base occurs at each position of these common SNPs for each chromosome. By correlating an individual's haplotypes with the presence of a disease, researchers can better understand complex diseases. The effort to characterize human variation, currently a major focus for the NIH, will be a tremendous undertaking requiring obtaining the haplotype information from a large collection of individuals from diverse populations [14].

Although the two chromosomes of an individual can be separated and analyzed independently as in Patil et al., 2001 [15], current technology suitable for large scale polymorphism screening obtains *genotype* information at each SNP. The genotype gives the bases at each SNP for both copies of the chromosome, but loses the information as to the chromosome on which each base

[1] The program will be available at the time of publication.

Haplotype	0,1 Representation	Frequency
CCGAT	00000	66
CTGAC	01001	24
ATACT	11110	10
CTGAT	01000	6
ATGAT	11000	1
ATGCC	11011	1
CCGAC	00001	1

Table 1: Block 6 from Daly et al. 2001, [2]. The block contains 5 SNPs over 11 kilobases. The horizontal line separates the common haplotypes from rare haplotypes. The first column shows the haplotypes from the transmitted chromosomes. The second column shows the same haplotypes but mapped to 0,1 representation. The 0 represents the common nucleotide at the position, while the 1 represents the rare nucleotide at the position. The third column is the frequency of the haplotype block in the transmitted chromosomes. Note that any chromosome that contained any ambiguity in the block due either to missing data or heterozygous genotypes for all members of the trio was omitted.

appears. Consider a SNP where there are two common bases, A or G. There are four possible cases for the haplotype. Two of the cases are where either both chromosomes contains A or both chromosomes contain G. We refer to these cases as *homozygous* genotypes. The other two cases are where the first chromosome contain A and the second contain G and vice versa. We refer to these cases as *heterozygous* genotypes. For this SNP, there are three possible cases for the genotype information. In the homozygous cases, the genotype will be either A or G respectively and we can infer that the base appears in both chromosomes. In the heterozygous cases, the genotype will be H (for heterozygous) and we can infer that in one chromosome, we have an A and in the other we have a G, but we can not infer on which chromosome each appears. This causes problems in reconstructing the haplotypes. Consider the example where an individual at four successive SNPs, with possible values A or G, has a genotype $AHHG$. In this case, the individual's haplotypes have two possibilities: either $AAAG$ on one chromosome and $AGGG$ on the other chromosome or $AAGG$ and $AGAG$. Without any other information, such as the genotypes from related individuals, it is impossible to determine the individual's actual haplotypes. This problem of haplotype resolution is often referred to as the phase problem.

Recent studies in linkage disequilibrium [6, 16] characterizing haplotype structure have shown that SNPs are grouped into "blocks" of limited diversity with the regions between blocks being "hot spots" of recombination. In each block containing n SNPs, typically around four haplotypes account for the majority of the haplotypes in the population. Consider the haplotype block shown in Table 1 consisting of 5 SNPs over 11 kilobases from a recent paper, Daly et al, 2001. We can map each of these haplotypes to the 0,1 representation where 0 represents the common nucleotide and 1 represents the rare nucleotide. The 0,1 representation for block 6 is also shown in Table 1. Note that 90% of the individuals contain one of four common haplotypes.

Haplotypes can be resolved from genotype data by making the assumption that most of the haplotypes within a block will loosely fit the perfect phylogeny model. This method for resolving haplotypes was first proposed in Gusfield, 2002 [9]. The perfect phylogeny model assumes an infinite site mutation model and allows no recombinations [11]. The infinite site mutation model makes the assumption that at each SNP site, a mutation only happened once in human history. This model forbids recurrent mutations or back mutations. The assumptions of the model imply that a chromosome with a mutation at a SNP is a direct descendant from the chromosome of the ancestor in which the mutation occurred. Likewise, any chromosome without the mutation can not be a descendant of a chromosome that has the mutation. Clearly, these assumptions are not realistic although it is reasonable to assume that recombinations and recurrent mutations are relatively rare events within a block. Thus, we consider a relaxed model which allows for a certain number of recurrent mutations and recombinations within a block.

In this paper, we present results for a highly accurate method for haplotype resolution from genotype data. Our method takes as input a population of genotypes and decomposes the SNPs into blocks. For each block we predict the common haplotypes as well as the haplotypes of each individual in the population. We also show that the common haplotypes roughly fit a perfect phylogeny model. Essentially, our method can effectively predict the haplotypes for *unrelated* individuals. This ability significantly reduces the costs and difficulties of characterizing human variation since it eliminates the need for collecting genotype data from complete trios.

Existing methods effectively assume all of the SNPs are in a single block. These methods include a variety of methods including the parsimony approach of Clark [1] and related approaches [7, 8, 12], maximum likelihood methods [4, 5, 10, 13] and statistical methods such as PHASE [18], and perfect phylogeny-based approaches [9]. All of these approaches suffer from the fact that they do not explicitly take into account the haplotype block structure. In addition, these methods are often too inefficient to be practical for large datasets. These methods typically can not scale to data that contains more than 30 sites. A similar approach to ours is the strict perfect phylogeny model approach of Gusfield, 2002 [9]. However, in this paper, we show that only if we relax the assumptions of the perfect phylogeny model does the algorithm work in practice.

Results

Predicting Haplotypes from Genotypes We performed our experiments over the data presented in Daly et al., 2001 [2]. Our first set of experiments assumes that we are given the block partition for the 11 blocks in the data. Our second set of experiments assumes that we have no prior information about the block partition of the 103 SNPs and are only given their genotypes. We apply our algorithm to determine the block partition. We also use a tiling technique to extend the haplotype predictions across block boundaries. We evaluate our predictions by comparing them to the correct haplotypes inferred from the trios.

Predicting Haplotypes within a Block In the first set of experiments, for each of the 11 blocks as defined in Daly et al., 2001 [2] we predicted the common haplotypes from the genotypes of the children in the trios as well as each child's haplotype. In all cases but the fourth block, the predictions for the common haplotypes are consistent with the published predictions as shown in Table 2. Note, in some cases, the published results use a $*$ character to denote two haplotypes with a single base difference between them. A significant portion of the data is missing, 10.03% of the total genotype data. This missing data comes from various sources of experimental error. We resolve the missing data. Over the 129 individuals and 103 SNPs, our error rate is only 0.53% in terms of bases, significantly lower the the amount of missing genotype data. If we ignore the missing data, our error rate in terms of bases is only 0.30%. Over individuals that contain the common haplotypes, our predictions are perfect. The errors only occur for individuals who have an uncommon haplotype. The program takes only a few seconds to make each of these haplotype predictions.

SNPs	Actual Common Haplotypes	Predicted Common Haplotypes	Frequency	Error Rate
1-8	GGACAACC	GGACAACC	215	0.0000
	AATTCGTG	AATTCGTG	38	
10-14	TTACG	TTACG	217	0.0000
	CCCAA	CCCAA	35	
16-24	CGGAGACGA	CGGAGACGA	139	0.0078
	GACTGGTCG	GACTGGTCG	52	
	CGCAGACGA	CGCAGACGA	34	
		CGGATACGA	15	
25-35	CGCGCCCGGAT	CGCGCCCGGAT	142	0.0021
	CTGCTATAACC	CTGCTATAACC	39	
	TTGCCCCGGCT*	CTGCCCCGGCT	35	
	CTGCCCCAACC*	TTGCCCCAACC	25	
36-40	CCAGC	CCAGC	146	0.0062
	CCACC	CCACC	51	
	GCGCT	GCGCT	30	
	CAACC	CAACC	12	
41-45	CCGAT	CCGAT	152	0.0140
	CTGAC	CTGAC	63	
	ATACT	ATACT	31	
78-84	CGTTTAG	CGTTTAG	142	0.0044
	TGTT*GA	TGTTTGA	53	
	TGATTAG	TGATTAG	20	
	CGTCTAG	CGTCTAG	12	
		TGTTGGA	10	
86-91	ACAACA	ACAACA	145	0.0129
	GCGGTG	GCGGTG	71	
	ACGGTG	ACGGTG	14	
	GTGACG	GTGACG	13	
92-98	GTTCTGA	GTTCTGA	142	0.0078
	TGTGTAA	TGTGTAA	49	
	TG*GCGG	TGTGCGG	32	
		TGCGTAA	15	
99-103	CGGCG	CGGCG	112	0.0031
	TATAG	TATAG	105	
	TATCA	TATCA	35	

Table 2: Predictions over blocks defined by Daly et al. 2001, [2]. The second column shows the common haplotypes as presented in Daly et al. 2001. The third column shows the predicted common haplotypes and the fourth gives their frequencies. The fifth column gives the error rate in terms of bases after resolving all missing data. The error rate is the total number of errors in the predictions for the divided by the total number of bases in the block. The error rate includes predictions for the uncommon haplotypes. The overall error rate over the 103 SNPs resolving missing data is 0.53%.

SNPs	Actual Common Haplotypes	Predicted Common Haplotypes	Frequency	Error Rate
46-76	CCCTGCTTACGGTGCAGTGGCACGTATT*CA	CCCTGCTTACGGTGCAGTGGCACGTATTGCA	137	0.0056
	TCCCATCCATCATGGTCGAATGCGTACATTA	TCCCATCCATCATGGTCGAATGCGTACATTA	59	
	CCCCGCTTACGGTGCAGTGGCACGTATATCA	CCCCGCTTACGGTGCAGTGGCACGTATATCA	19	
	CATCACTCCCCAGACTGTGATGTTAGTATCT	CATCACTCCCCAGACTGTGATGTTAGTATCT	10	
		CCCTGCTTACGGTGCAGTGGCACGTATTTCA	9	

Table 3: Predictions over data from Daly et al. 2001, [2] (continued).

Predicting Blocks from Genotypes Typically, we must determine the block partition directly from the genotype data. We first make haplotype predictions for all possible blocks of up to length 30 using the local haplotype prediction algorithm and discard any blocks with more than five common haplotypes. This leaves 1140 potential blocks.

Since there are only a few haplotypes in each block, we do not need to check every SNP in the block to determine which of the common haplotypes an individual has. For each block, we can define a set of representative SNPs that are sufficient to determine an individual's haplotypes. In Table 1, the second, third and fifth SNPs are sufficient to determine the haplotype. For example, if we observe T, A, and T in these SNPs, we can infer that the individual has the third haplotype (assuming the individual has one of the common haplotypes). On the other hand, if we observe T, G, and C, we can infer the individual has the second haplotype. For this block, there are other possibilities for a set of representative SNPs such as the first, second and fifth SNPs or the second, fourth and fifth haplotypes. However, for this block, the minimum number of representative SNPs is three. That is, no two SNPs can distinguish the four common haplotypes.

A criterion for determining a good block partition is minimizing the sum of the number of representative SNPs over all blocks. This criterion has been used to partition blocks on a larger scale [15, 19]. The reasoning behind minimizing the number of representative SNPs is to reduce the cost of obtaining an individual's haplotype. If we assume that an individual has only the common haplotypes, then it is enough to determine what common haplotypes the individual has at each block position. To determine this, we need to obtain information about the individual's SNPs only at the positions of the representative SNPs. This is significantly cheaper than obtaining information about all of the individual's SNPs.

Using dynamic programming, we choose the best block partition for the data from Daly et al. 2001, [2] where the objective is to minimize the number of representative SNPs over the entire block partition. Table 4 shows the predicted block partition which contains 27 representative SNPs. For each block in the partition, Table 4 gives the number of representative SNPs in the block, the common haplotypes, as well as the error rate after resolving the missing data. Over the blocks chosen by the algorithm, the error rate for the 103 SNPs is 0.73% and only 0.47% if we ignore the missing data. The block partition varies from the partition described in Daly et al., 2001 [2] (shown in Table 2) since the criterion for defining block partitions vary. Our criterion, to minimize the number of representative SNPs, is consistent with the criterion in [15] while the criterion in [2] determines blocks by estimating the recombination frequencies between blocks.

Tiling Block Predictions Over the predicted blocks, for each individual, we can accurately predict which haplotypes the individual contains for each block. A more difficult problem is to recover the complete haplotypes of the individual. Given the haplotype block predictions, this problem reduces to determining which of the individuals haplotypes are on one chromosome and which are on the other. At each block boundary, there are 2 possibilities for how the haplotype blocks are arranged on the chromosome. Using the predictions of Table 4, since there are 11 blocks and each of the two predicted haplotypes can be on either chromosome, there are a total of 2^{10} possible complete haplotypes.

We use a "tiling" technique to extend the haplotype predictions across block boundaries. We make predictions for a set of "tiling" blocks. These blocks span our original block boundaries. At a block boundary, we have the predictions for both haplotypes on each side of the boundary as well as the predictions for the haplotypes of the tiling block that spans this boundary. There are four possible ways to arrange these six blocks on the two chromosomes. We choose the arrangement that has the least number of inconsistencies. In some cases it is impossible to determine how to

SNPs	Num Rep SNPs	Predicted Common Haplotypes	Frequency	Error Rate
1-10	1	GGACAACCGT	199	0.0016
		AATTCGTGGC	34	
11-15	2	TACGC	134	0.0108
		TACGT	85	
		CCAAC	34	
16-24	3	CGGAGACGA	139	0.0078
		GACTGGTCG	52	
		CGCAGACGA	34	
25-36	3	CGCGCCCGGATC	141	0.0019
		CTGCCCCGGCTC	35	
		CTGCTATAACCG	34	
		TTGCCCCAACCC	23	
37-46	4	CAGCCCGATC	139	0.0101
		CGCTCTGACT	36	
		CACCATACTC	29	
		CACCCTGACT	18	
		AACCCTGACC	14	
47-76	3	CCTGCTTACGGTGCAGTGGCACGTATTGCA	137	0.0044
		CCCATCCATCATGGTCGAATGCGTACATTA	58	
		CCCGCTTACGGTGCAGTGGCACGTATATCA	21	
		ATCACTCCCCAGACTGTGATGTTAGTATCT	12	
		CCTGCTTACGGTGCAGTGGCACGTATTTCA	9	
77-79	2	CCG	151	0.0232
		GTG	51	
		CTG	40	
80-82	1	TTT	205	0.0000
		ATT	29	
83-91	3	AGCACAACA	144	0.0121
		GACGCGGTG	48	
		GAGGCGGTG	18	
		GGCGTGACG	13	
		AGCACGGTG	13	
92-98	3	GTTCTGA	142	0.0078
		TGTGTAA	49	
		TGTGCGG	33	
		TGCGTAA	15	
99-103	2	CGGCG	112	0.0031
		TATAG	105	
		TATCA	35	

Table 4: Predicted block partition over data from Daly et al. 2001, [2]. The second column gives the number of representative SNPs. The third column shows the predictions for the common haplotypes and the fourth gives their frequencies. The fifth column gives the error rate after resolving missing data. The error rate is the total number of errors in the predictions for the divided by the total number of bases in the block. The error rate includes predictions for the uncommon haplotypes. The overall error rate over the 103 SNPs resolving missing data is 0.73%.

connect the blocks because two possibilities have the minimum number of inconsistencies. We refer to these cases as *unresolvable*. Consider the case where one of the blocks contains two copies of the same haplotype. Since the haplotypes are the same, it is impossible to determine to how to connect the neighboring haplotypes.

For each individual in the data, we must make 10 choices on how to connect their blocks. Over the data set of 129 individuals, 784 of these choices are unresolvable. Over the remaining 506 choices, we correctly predict 473 of them for an error rate of 6.5%.

Comparison with PHASE We applied the widely used program PHASE [18] to reconstruct the haplotypes in order to provide a comparison with our method. PHASE is able to make local predictions for each of the 11 blocks, however, these predictions took hours for each block. In practice, since we do not know the block partition, we must make predictions for each of more than $2,500$ candidate blocks in order to determine the optimal block partition. This makes PHASE impractical for predicting the block partition from the genotypes.

Discussion

Recent studies in haplotype structure have shown that haplotypes are structured into blocks with limited diversity. A recent paper by Gusfield 2002 [9] suggested the use of perfect phylogeny to reconstruct the haplotypes. However, the actual haplotypes do not fit the perfect phylogeny model. For a given set of haplotypes, we can measure the percentage of conflicts to the perfect phylogeny model. If we consider all haplotypes, even the uncommon ones, the data does not fit the perfect phylogeny model as shown in Figure 1A. If we consider instead a relaxed perfect phylogeny model using an error threshold which allows a small number of haplotypes to be excluded when determining conflicts, we notice that the haplotypes fit the model much better. The results for the Chromosome 5p31 data are in Figures 1A-C. Clearly, as the error threshold increases, the number of conflicts significantly decreases. This is due to the fact that the infrequent haplotypes cause the majority of the conflicts with the perfect phylogeny model.

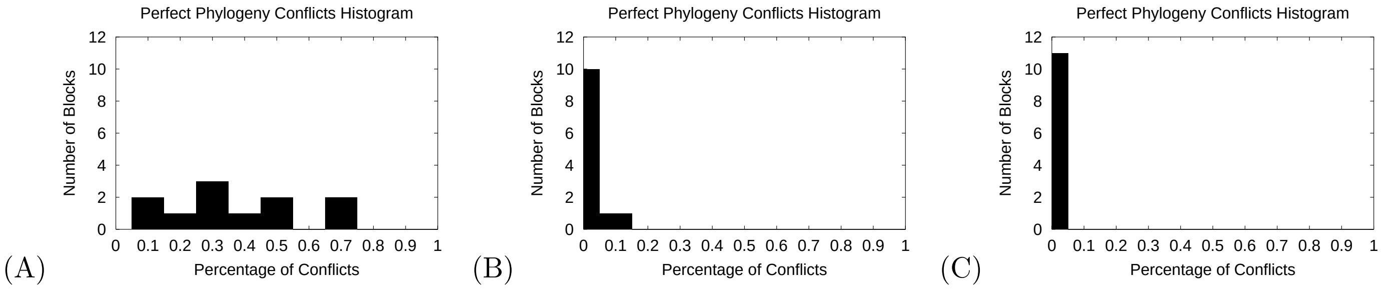


Figure 1: Histograms of percentage of conflicts under different error thresholds for the blocks defined in Daly et al., 2001 [2]. Thresholds are (A) 0% (B) 5% and (C) 10%.

We have demonstrated our method over actual haplotype data collected from 129 trios and verified the accuracy of our predictions to the correct haplotypes. Our method is highly accurate and efficient. The predictions differ from the actual haplotypes by less than 1% even after resolving approximately 10% of the missing genotype data. We also present a method for determining the block partition from genotype data and a method for extending haplotype predictions beyond single blocks.

The program for predicting haplotype structure is publicly available via a webserver at `http://www.cs.columbia.edu/compbio/haplotype/`[2].

Methods

Dataset Description The data set over which we perform our experiments is a 500 kilobase region of chromosome 5p31 containing 103 SNPs from the studies of Daly et al., 2001 [2] and Rioux et al., 2001 [17]. In this study, genotypes for the 103 SNPS are collected from 129 mother, father, child trios from a European-derived population in an attempt to identify a genetic risk factor for Crohn's disease. A significant portion of the genotype data (10.03%) is missing with an average of 10 SNPs per individual's genotype missing. The 103 SNPs were split into 11 blocks containing from 5 to 31 SNPs and ranging from 3 to 92 kilobases. For each of these blocks, four haplotypes correspond to 90% of the individual chromosomes. Since this set consists of trios, we can infer each individual's haplotypes. This data is publicly available at `http://www-genome.wi.mit.edu/humgen/IBD5/`.

We use the data to show that the perfect phylogeny model roughly fits the common haplotypes. To evaluate our predictions of haplotypes, we make predictions over the genotype data of the individuals and then compare our predictions to the correct haplotypes inferred from the trios.

Inferring Haplotypes from Trios We use data collected in trios to measure the accuracy of our method. Given the genotypes for a mother, father, child trio, in most cases, we can infer the haplotypes for each of the individuals. We infer the haplotypes at each SNP independently using Mendelian genetics. We define each parent to have a transmitted chromosome and an untransmitted chromosome. The child has both transmitted chromosomes from the parents. Each SNP for each chromosome can be represented by either 0 or 1 for the common base or mutation base respectively. For these four chromosomes, there are a total of 16 possibilities. Each SNP in the genotype can be denoted either 0, 1 or 2 which represents homozygous with the common base, homozygous with the mutation base, or heterozygous respectively. Although there are 27 possible genotypes for each trio at a given SNP, many of them are invalid such as the case where the father and child are homozygous for the common base and the mother is homozygous for the mutation base. In any valid case where at least one of the genotypes in the trio is homozygous, we can uniquely determine the haplotypes for that SNP. For example, consider the case where the genotypes for a mother, father, child trio are 2, 2, 0 respectively. From the child we know that both transmitted chromosomes contain the common base. This implies that both of the parents must have transmitted the common base to the child on the transmitted chromosomes and their untransmitted chromosomes contain the mutation base.

Only in the case where all three of the genotypes are heterozygous is there more than one possible resolution. For example, if the genotypes for the mother, father, child trio are 2, 2, 2, then it is possible that the mother's transmitted chromosome contains the mutation base and the father's transmitted chromosome contains the common base, or vice versa.

In the case where there is missing data, only if some of the genotypes are homozygous can we infer portions of the trio. For example, if both parents are homozygous, then we can infer the child's haplotypes even if the child's genotype is missing.

[2] The program will be available at the time of publication.

Measuring Perfect Phylogeny from Haplotypes The perfect phylogeny model implicitly defines a phylogenetic tree for the haplotype data such as the one for the four common haplotypes from Table 1 shown in Figure 2. At each edge of the tree, we have a mutation labeled with the position of the mutation. Under the perfect phylogeny model assumptions, there can only be one edge for each site in the tree. Once a mutation occurs at an edge, the mutation must be present in each individual in the subtree rooted at that edge and only in the subtree. Each haplotype at a node contains all of the mutations along the path from the root node to the current node.

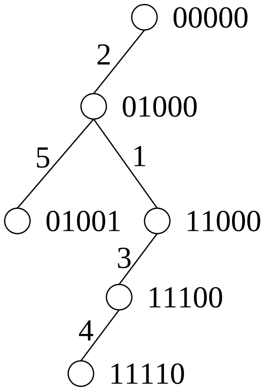


Figure 2: The Perfect Phylogeny Tree for the data from Table 1

We can measure how well a set of haplotypes fits the perfect phylogeny model by constructing a phylogenetic tree for the haplotypes. Typically there is more than one possible phylogenetic tree for the haplotypes in the data. These trees can be determined by inferring the relations between sites from the individual's haplotypes. If there exists an individual which has a mutation at both sites i and j, we can infer that sites i and j must be along the same path in the tree. We refer to this relation as *descendant* (parent-child). For example, in Table 1, site 2 and 5 are descendants since the second haplotype has a mutation at both sites. If two sites have the descendant relation and there is also another individual that has a mutation at site i and no mutation at site j, we can infer that site i must be the *parent* of site j in the phylogenetic tree, i.e. the mutation at site i must have occurred before the mutation at site j. For example, from the third haplotype we can infer that site 2 is the parent of site 5. If there is a pair of sites i and j such that there is an individual that has a mutation at site i and no mutation at site j and another individual that has a mutation at site j and no mutation at site i, then we can infer that sites i and j are *siblings* in the tree. The common ancestor to any individuals that have either of these mutations has neither of the mutations. For example, consider sites 1 and 5 in Table 1. ¿From the second and third haplotypes, we can infer that sites 1 and 5 are siblings.

Not all haplotypes fit the perfect phylogeny model. Consider the sixth haplotype in Table 1 which implies that sites 1 and 5 are descendants. However, as above, sites 1 and 5 are siblings. For the haplotypes to fit a perfect phylogeny model, a pair of sites can not have both the sibling and descendant relation. More formally, a conflict with the perfect phylogeny model occurs whenever there is a pair of sites i and j under the following condition. There is an individual which has both mutations (denoted 11), an individual with a mutation at site i and no mutation at site j (10) and an individual with a mutation at site j and no mutation at site i (01). If we consider the uncommon haplotypes in Table 1, there are many conflicts with the perfect phylogeny model. The sixth haplotype causes conflicts with site 5 and sites 1, 2, and 4. The seventh haplotype causes conflicts with site 5 and 2. Note that these conflicts only occur if we consider the uncommon haplotypes.

In general, we can use the frequencies of the haplotypes to help determine the relations in the presence of conflicts with the model. Consider sites 1 and 5 in Table 1. 24 haplotypes have a mutation at site 5 and no mutation at site 1. 10 haplotypes have a mutation at site 1 and no mutation at site 5. Only a single haplotype has a mutation at both site 1 and site 5. In this case, there is much more evidence to support that sites 1 and 5 are siblings rather than descendants.

We can measure how well a block fits perfect phylogeny by counting the number of conflicts between pairs of sites within the block. For a block containing n SNPs, we can normalize the count by $\binom{n}{2}$ to compare blocks that contain a different number of SNPs. In general, the infrequent haplotypes cause many conflicts with the perfect phylogeny model. In fact, a single individual can cause many conflicts which can conceal the fact that the haplotypes of the remaining individuals fit the perfect phylogeny model. We adapt this measure to evaluate how well the majority of the data fits the perfect phylogeny model, by introducing an error threshold. We consider a pair of sites to have a conflict if the number of individuals that contain 11, 10 and 01 are all above the error threshold. For example consider sites 4 and 5 in Table 1 considering all of the haplotypes. For these two sites we have 25 individuals that have 01, 10 individuals that have 10 and only a single individual that has 11. If the error threshold is 0, this would be a conflict. However, if the error threshold is 1 or higher, the we would not consider the individual who has 11 and there would be no conflict. Typically, the error threshold is set to a fraction of the size of the data.

Haplotype Resolution Via the Perfect Phylogeny Model The problem of haplotype resolution as perfect phylogeny was proposed in Gusfield, 2002 [9]. An overview of the algorithm is given in Appendix A and the complete algorithm is given in [3].

The basic idea behind the algorithm is to make haplotype resolutions that are consistent with the perfect phylogeny model. In order to avoid conflicts with the perfect phylogeny model, heterozygous genotypes at pairs of sites must be resolved consistently. If the pair of sites are descendant (parent-child), then the pair of heterozygous sites must be resolved such that both mutation bases occur on one chromosome and both common bases appear on the other chromosome. This type of resolution is called *equal* resolution. Likewise, if a pair of sites are siblings, the mutations must occur on different chromosomes. This type of resolution is called *unequal* resolution. For example, consider an individual with genotype 0220. If the second and third sites are siblings in the perfect phylogeny tree, then the haplotypes for the individual are 0100 and 0010. On the other hand if the pair of sites are descendant, then the haplotypes are 0110 and 0000. Multiple individuals must have the same resolution between pairs of sites, otherwise there is a conflict. This reduces the number of possible resolutions from 2^d where d is the number of SNPs where individuals have heterozygous genotypes in the population to 2^n where n is the number of sites. In the case of the block in Table 5, the number of possible resolutions in general is 2^{173} while if we restrict to the perfect phylogeny model, the number of possible resolutions is at most 64.

Many of the relations between the sites can be inferred from the genotype data. For example, if there exists and individual with a 11, 12, or 21 at a pair of sites, we know that the two sites must be descendant. Likewise, if we know that there are two individuals where one has a 10 or 20 and a second individual that has a 01 or 02, then we can infer the two sites are siblings. If we can infer the relationships between sites, we can resolve the pairs of heterozygous genotypes that occur in those sites. For example, consider the first and third sites of Table 5. They are descendants since we have several individuals with the pair 12. We would resolve all pairs of individuals which have heterozygous genotypes in both positions with equal resolution. For the pairs of sites where we can

Genotype	0,1 Representation	Frequency
CHGAH	02002	23
CCGAT	00000	20
HHHHT	22220	11
HTHHH	21222	4
CTGAH	01002	3
CTGAC	01001	2
CCGAH	00002	1
AHHHT	12220	1
HTHHT	21220	1
ATHCH	11212	1
CHGAT	02000	1

Table 5: Genotypes from block 6 from Daly et al. 2001, [2]. The block contains 5 SNPs over 11 kilobases. The block represents SNPs 41-45 of the 103 SNPs. The first column shows the genotypes from the 129 children with H representing the heterozygous genotype. The second column shows the same genotypes but mapped to 0,1,2 representation. The 0 represents the homozygous genotype of the common nucleotide at the position, while the 1 represents the homozygous genotype of the rare nucleotide at the position. A 2 represents the heterozygous genotype. The third column is the frequency of the genotype among the 129 children. Note that any genotypes that contained any missing data were omitted.

not infer the relations, we iterate over possible relations to obtain all of the solutions that fit the perfect phylogeny model.

Not all genotype data fits the perfect phylogeny model. One type of conflict is a conflict between a pair of sites referred to as a *column conflict*. This conflict is analogous to the conflicts described in building the phylogenetic tree from the haplotypes. These are conflicts between pairs of sites which arise if there is evidence to support that the pair of sites have both the sibling and descendant relation. Consider the example of a column conflict between the second and fifth sites in Table 5. The third and seventh haplotypes imply that the sites are siblings, while the fourth, fifth, sixth, and tenth imply that they are descendants. A second type of conflict arises because in some cases, there is no possible genotype resolution to fit the perfect phylogeny model. For example, consider three sites i, j, and k, and an individual which is has a heterozygous genotype in each site. If i and j, j and k and i and k are all siblings, there is no valid resolution. If i and j must have unequal resolution, and j and k must have unequal resolution, then i and k must have equal resolution, but since i and k are siblings this is a conflict. We refer to this type of conflict as a *graph conflict*. We describe how the algorithm is modified to make predictions in the presence of noise in the data and conflicts in Appendix A.

Maximum Likelihood Model for Local Haplotype Reconstruction We choose the "best" solution from the set of candidate solutions that roughly fit the perfect phylogeny model using a maximum likelihood model. The maximum likelihood model estimates the likelihood of observing the population of genotypes given the predicted haplotype frequencies. The likelihood model assumes independence between the haplotypes of an individual.

Given a population of n individuals, we denote the two haplotypes of the ith individual as i_1 and i_2. We use the notation $f(i_1)$ to denote the frequency of the haplotype i_1 in the population. The likelihood of a haplotype i_1 is $\frac{f(i_1)}{2n}$. The likelihood for each genotype of an individual is simply the product of the likelihoods of their two haplotypes $\frac{f(i_1)f(i_2)}{(2n)^2}$. The likelihood of a candidate solution for a population of genotypes is

$$L = \prod_{i=1}^{n} \frac{f(i_1)f(i_2)}{(2n)^2} \tag{1}$$

This model is consistent with previous maximum likelihood models for choosing haplotypes from genotypes [4, 10, 13, 5, 18]. The main caveat with these previous approaches is that they do not restrict the possible solutions to the ones that roughly fit the perfect phylogeny model. This results in far to many possible haplotype resolutions that need to be evaluated using the maximum likelihood model, and thus the previous algorithms are inefficient and not practical.

Resolving Missing Data Missing data is resolved after the algorithm resolves heterozygous genotypes. When determining the relations between sites, individuals with missing genotype data for one of those sites are ignored. Once the maximum likelihood model chooses the best haplotype resolutions, there are typically several common haplotypes which account for the majority of the population. At this point missing genotypes are resolved by choosing the most likely SNP based on the maximum likelihood model. Effectively, we resolve the missing data by choosing the SNP to match the common haplotypes.

Computing Block Partitions from Genotypes Our method predicts block partitions directly from the genotype data. We first define a set of candidate blocks. Given a maximum block length, we slide a window across the data for each block length to define our candidate blocks. For each candidate block, we apply the local haplotype prediction algorithm to predict the haplotypes. Our algorithm accurately predicts haplotypes only if there is limited diversity within a block. To ensure accuracy of our predictions, we discard all candidate blocks that have more than five common haplotypes. For each remaining candidate block, we determine the number of representative SNPs. This is done by enumerating over all subsets of the SNPs in the block and checking to see if they distinguish between the common haplotypes.

To compute the block boundaries for the haplotypes, we use a straightforward dynamic programming technique similar to the one presented in [19]. The main difference is that in our setting, there is no missing data since it is resolved by the local prediction algorithm. Note that the block partition in Daly et al., 2001 [2] does not assign several SNPs to blocks. We can easily modify the dynamic programming algorithm to optimize a block partition where several SNPs are allowed to be left out.

Tiling Local Blocks to Obtain Global Haplotype Structure In order to reconstruct the complete haplotype, we use a tiling technique. For each haplotype block boundary, we make a predictions for a tiling block of length 6 which spans the boundary and 3 SNPs on either side. For each individual, we must determine how to connect the two haplotypes on either side of the block boundary. We make the choice that is most consistent with the prediction of the tiling block.

Consider the following example where an individual has the haplotypes $TACGC$ and $TACGT$ for predicted block 2 in Table 4 and $CGGAGACGA$ and $GACTGGTCG$. The prediction of the tiling block is $CGCGAC$ and $CGTCGG$. Using these predictions, it is clear that the complete haplotype is $TACCGCGACTGGTCG$ and $TACGTCGGAGACGA$. Note that the actual predictions of the tiling block may not be as accurate as the block predictions. However, as the results show, they are accurate enough to make a decision between two choices on how to join together the haplotypes.

A Appendix

Haplotype Resolution via Perfect Phylogeny The complete algorithm as well as proofs of correctness are given in [3]. Here we give a summary of the algorithm.

The basic idea behind the algorithm is to determine the relations between the sites which defines whether we use equal or unequal resolution to resolve individuals containing pairs of heterozygous genotypes at the sites.

The data consists of the genotypes for n individuals at m sites. For each pair of sites i and j, we determine whether the pair of sites are either descendant, siblings or ambiguous. If the pair of sites contains an individual with the genotype either 11, 21, 12 then sites i and j have the relation descendant. Each of those genotypes implies that there is a haplotype 11. On the other hand, if the data contains an individual with the genotypes 10 or 20 and another individual with the genotypes 01 or 02, then sites i and j are siblings. If neither of these cases occur, then the relation is ambiguous.

The algorithm checks for "graph" conflicts and enumerates the possible solutions simultaneously. For each site i, we construct the following graph. The graph will determine what the relations of each site is with i. Each of the m sites is a node in the graph. We connect two nodes with an edge if both sites contain an individual that has a heterozygous genotype at each site as well as a heterozygous genotype at node i. For the data to fit the perfect phylogeny model, each graph must be bipartite. Since the graph is bipartite, we can partition each connected component of the graph into two parts. One of the parts will correspond to the sites that are siblings and the other part correspond to the descendants. If either part contains a site where there is a unambiguous relation, we can determine the remaining relations for the sites in the connected component. For these connected components, we can resolve the pairs of heterozygous genotypes accordingly. For the connected components that are still ambiguous, both relations are possible which will correspond to different solutions. If we iterate over the assignments of these relations and resolve all pairs of heterozygous haplotypes accordingly, we will arrive to the complete set of possible haplotype resolutions.

Handling Noisy Data In many cases, the uncommon haplotypes cause conflicts with the perfect phylogeny model. We adapt the algorithm by requiring more evidence to determine whether a relation is descendant or sibling. For each pair of sites we count the number of genotypes 11, 21 or 12, the number of genotypes 01 or 02, and the number of genotypes 10 or 20. If the use the strict perfect phylogeny model, a single individual 11, 21 or 12 will cause the sites to have the relation descendant even if this individual has an uncommon haplotype. We adapt the algorithm by introducing an error threshold. If any of these counts are below the error threshold, we set the counts to 0. For example, consider the genotypes for sites 2 and 5 in Table 5. There are 11 individuals with 13 individuals with genotype 20, 8 individuals with genotype 12 and only 2 individuals with genotype 21 and a single individual with 11. If the error threshold is 3, then we would determine that the sites have the relation descendant.

In some cases, even with an error threshold, there are individuals that provide evidence that a pair of sites are both siblings and descendants. In this case we resolve the conflict by choosing the relation where there is more evidence to support it based on the counts.

References

[1] AG Clark. Inference of haplotypes from pcr-amplified samples of diploid populations. *Journal of Molecular Biology and Evolution*, 7(2):111–22, Mar 1990.

[2] MJ Daly, JD Rioux, SF Schaffner, TJ Hudson, and ES Lander. High-resolution haplotype structure in the human genome. *Nature Genetics*, 29(2):229–32, Oct 2001.

[3] Eleazar Eskin, Eran Halperin, and Richard M. Karp. Efficient reconstruction of haplotype structure via perfect phylogeny. *Technical Report. UC Berkeley Computer Science*, 2002.

[4] L Excoffier and M Slatkin. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution*, 12(5):921–7, Sept 1995.

[5] D Fallin and NJ Schork. Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. *American Journal of Human Genetics*, 67(4):947–59, Oct 2000.

[6] DB Goldstein and ME Weale. Population genomics: linkage disequilibrium holds the key. *Current Biology*, 11:R576–R579, 2001.

[7] 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*, 2000.

[8] D Gusfield. Inference of haplotypes from samples of diploid populations: complexity and algorithms. *Journal of Computational Biology*, 8(3):305–23, 2001.

[9] Dan Gusfield. Haplotyping as perfect phylogeny: Conceptual framework and efficient solutions (extended abstract). In *Proceedings of the 6th International Conference on Computational Molecular Biology (RECOMB 2002)*, 2002.

[10] ME Hawley and KK Kidd. Haplo: a program using the em algorithm to estimate the frequencies of multi-site haplotypes. *Journal of Heredity*, 86(5):409–11, Sep-Oct 1995.

[11] R Hudson. Gene genealogies and the coalescent process. *Oxford Survey of Evolutionary Biology*, 7:1–44, 1990.

[12] G. Lancia, V. Bafna, S. Istrail, R. Lippert, and R. Schwartz. Snps problems, algorithms and complexity, european symposium on algorithms. In Springer-Verlag, editor, *Proceedings of the European Symposium on Algorithms (ESA-2001), Lecture Notes in Computer Science*, volume 2161, pages 182–193, 2001.

[13] JC Long, RC Williams, and M Urbanek. An e-m algorithm and testing strategy for multiple-locus haplotypes. *American Journal of Human Genetics*, 56(3):799–810, Mar 1995.

[14] NIH. Large-scale genotyping for the haplotype map of the human genome. RFA: HG-02-005.

[15] N Patil, AJ Berno, DA Hinds, WA Barrett, JM Doshi, CR Hacker, CR Kautzer, DH Lee, C Marjoribanks, DP McDonough, BT Nguyen, MC Norris, JB Sheehan, N Shen, D Stern, RP Stokowski, DJ Thomas, MO Trulson, KR Vyas, KA Frazer, SP Fodor, and DR Cox. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, 294(5547):1719–23, Nov 23 2001.

[16] DE Reich, M Cargill, S Bolk, J Ireland, PC Sabeti, DJ Richter, T Lavery, R Kouyoumjian, SF Farhadian, R Ward, and ES Lander. Linkage disequilibrium in the human genome. *Nature*, 411(6834):199–204, May 10 2001.

[17] JD Rioux, MJ Daly, MS Silverberg, K Lindblad, H Steinhart, Z Cohen, T Delmonte, K Kocher, K Miller, S Guschwan, EJ Kulbokas, S O'Leary, E Winchester, K Dewar, T Green, V Stone, C Chow, A Cohen, D Langelier, G Lapointe, Gaudet D, J Faith, N Branco, SB Bull, RS McLeod, AM Griffiths, A Bitton, GR Greenberg, ES Lander, KA Siminovitch, and TJ Hudson. Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to crohn disease. *Nature Genetics*, 29(2):223–8, Oct 2001.

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

[19] K Zhang, M Deng, T Chen, MS Waterman, and F Sun. A dynamic programming algorithm for haplotype block partitioning. *Proceedings of the Nationall Acadamy of Science*, 99(11):7335–9, May 28 2002.