Algorithms for Haplotype Phasing

Christine Lo

Abstract

A *haplotype* is the sequence of nucleotides along a single chromosome. As humans, we have 23 pairs of chromosomes. However, with current technology, it is difficult to separate the two chromosomes of a pair and we often get combined haplotype, or *genotype*, information. The objective of *haplotype phasing* is to resolve (phase) the haplotypes given genotype information. Knowing the haplotypes not only gives us a complete picture of an individual's genome, but also has other significant biological motivations.

Here we study how the haplotype phasing problem has been reformulated as technology to read the human genome has progressed. We also describe the algorithmic techniques used under the different formulations, and end by discussing the current state of the problem and future directions.

1 Introduction

Over the last two decades, there has been significant interest in understanding the genetic makeup of humans. With international efforts such as the Human Genome Project [1] and the International HapMap Project [2], technology to read the human genome has been rapidly developing. However, these technologies are still limited, and it is left to computational methods to detect and fix errors and piece together partial information from the technology. In this paper, we discuss the problem of *haplotype phasing*. As technology has progressed, the haplotype phasing problem has been reformulated to make use of the additional information gained. We discuss the history of the problem- how the technology and computational methods have changed since the first algorithm was proposed in 1990, describe the current state of the problem, and end by discussing possible future directions.

The human genome is made up of 23 pairs of chromosomes. Because humans are diploid, we have two copies of each chromosome type- one from our mother and one from our father- for a total of 46 chromosomes. The two copies are highly homologous to each other and only differ at a small fraction (0.1%) of variant sites. A *haplotype* is the nucleotide sequence along a single chromosome, but we can ignore the homozygous regions of the chromosomes and only consider the variant sites. For a chromosome with k variants, we can represent its haplotype as a string from the set $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}^k$. In fact, we assume that variants are bi-allelic, that is each variant takes one of two possible allelic values. Therefore, without loss of generality, we can represent haplotypes as a string from the set $\{\mathtt{0}, \mathtt{1}\}^k$, where 0 and 1 represent the two possible allelic values at each variant location. With current technology, it is difficult to separate a pair of chromosomes, and we often get the two haplotypes mixed together. A *genotype* is the combined haplotype information for a pair of chromosomes. We can represent it as a k-length ordered list of pairs where each pair is from the set $\{(\mathtt{0}, \mathtt{0}), (\mathtt{1}, \mathtt{1}), (\mathtt{0}, \mathtt{1})\}$. The list is ordered according to the chromosomal position of each pair, but the pairs themselves are unordered. See Figure 1 for an example. In a genomic region with k sites, there are 2^{k-1} possible haplotypes. The objective of the haplotype phasing problem is to recover the two haplotypes (out of the 2^{k-1} possible haplotypes) of an individual. We give the following biological formulation of haplotype phasing keeping in mind that the input will change with different technologies, but the output and biological objective will remain the same:

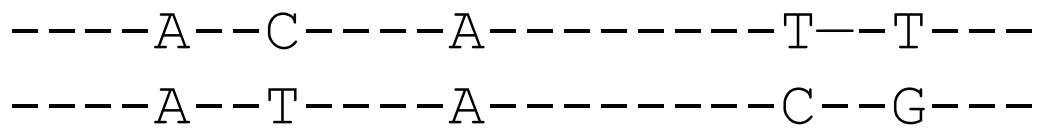


Figure 1: **Haplotypes vs Genotypes.** Example of two chromosomes of the same type. Only the variant sites are shown. The non-variant sites are represented by '-'. In this example, there are five variants. The haplotypes are '`ACATT`' ('`00000`') and '`ATACG`' ('`01011`'). The genotype is represented as a list of unordered pairs and is $\{(\texttt{A},\texttt{A}),(\texttt{C},\texttt{T}),(\texttt{A},\texttt{A}),(\texttt{T},\texttt{C}),(\texttt{T},\texttt{G})\}$ ({(0,0), (0,1), (0,0), (0,1), (0,1)}).

Haplotype Phasing Problem
Input: The genotype of an individual, $G = (g_1, g_2, ...g_k)$, where $g_i \in \{(\texttt{0},\texttt{0}),(\texttt{1},\texttt{1}),(\texttt{0},\texttt{1})\}$ for $1 \leq i \leq k$.
Output: The pair of haplotypes, $H = \{h_1, h_2\}$ for the individual, where $h_1, h_2 \in \{0,1\}^k$ and H is consistent with G. We use $\oplus$ to denote conflation; if $h_1 \oplus h_2 = G$, then we say H is consistent with G.

1.1 Evolution of Haplotypes. We get one copy of each chromosome type from our mother and one from our father. However, the evolution of haplotypes is complicated by *mutations* and *recombinations*.

A mutation will change the allelic value at a chromosome site from parent to child. The *mutation rate*, denoted as θ, is used to measure the chance of a particular site mutating. In humans, the mutation rate is around 2.5×10^{-8} [3]. The low mutation rate is why two chromosomes of the same type only differ at a small fraction of variant sites (0.1%). There are two types of variant sites- *Single Nucleotide Polymorphisms (SNPs)* and *Single Nucleotide Variants (SNVs)*. The difference between the two is defined somewhat arbitrarily. SNPs are variant sites common in a population. SNVs are variant sites that are unique to an individual. Around $1 - 10\%$ of an individual's variant sites are SNVs. In humans, the combined length of all the chromosomes is roughly 3 billion base pairs (bp). Thus, there are roughly 3 million variant sites and around $30,000 - 300,000$ SNVs.

Genetic diversity is also caused by recombination which is when the two chromosomes of a pair exchange regions of their genome. Recombination occurs during meiosis, which is the process where reproductive cells are formed in the parents. For example, a recombination at site r will cause region $[1, r]$ of one chromosome to be combined with region $[r + 1, \ell]$ of the other chromosome, where both chromosomes have length ℓ. See Figure 2 for an example. Each position along the chromosome is associated with a probability of recombination- this is called the *recombination rate*. Certain areas in the genome are more prone to recombination than others. Regions of sites with a high probability of recombining are known as *recombination hotspots*.

1.2 Motivation. Haplotypes give us a complete description of the human genome [4], and are much more informative than genotypes. Here we describe how haplotype information is useful for various application including association studies, detecting positive selection, estimating recombination rate, understanding gene function, and studying regions of the genome that are functionally related.

To elaborate, haplotypes allow us to find *associations* between a particular gene and a disease. The associations we can detect with haplotypes can not always be detected with only genotypes. See Figure 3 for an example of association.

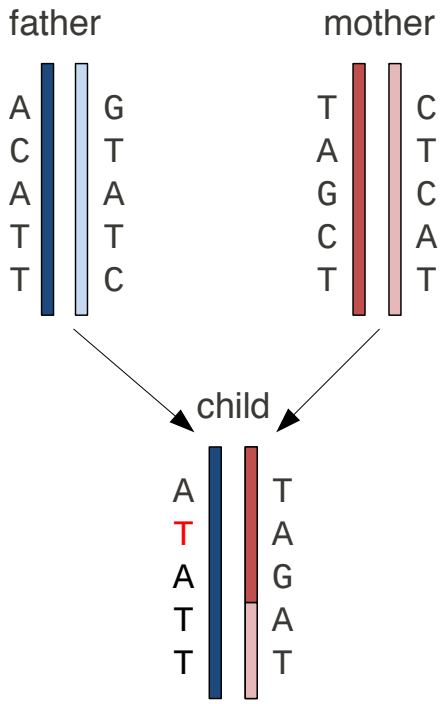


Figure 2: **Evolution of Haplotypes.** A child inherits one chromosome from the father and one from the mother. In this example, there is a mutation at the second site of the paternal chromosome (`C` → `T`). There is also a recombination on the maternal chromosome between the third and fourth site.

A second application of haplotype information is to detect *positive selection*. If a particular variant is under neutral selection, it will take a long time before many individuals in a population have the variant. During this time, recombination around the variant is likely to occur, disrupting the haplotype. However, under positive selection, the frequency of the variant in a population will rise faster and there is less time for recombination to occur around the variant. Thus, we can detect positive selection by looking for long haplotypes that are common in the population.

Haplotypes also help *estimate recombination rate*. If we know the genealogy and the haplotypes of a population, then we can detect where recombination occurs. For example, in Figure 2 we know where the recombination occurred by looking at the child's and the mother's haplotypes. Information about where recombinations occur can be used to estimate recombination rate. We can also detect recombination hotspots by looking for regions of the genome where there are a lot of recombinations in the population.

A fourth application of haplotypes is to help understand the function of a gene. The function of a gene can be determined by the way mutations occur on the two chromosomes. If the mutations occur on the same chromosome (in *cis*) then only one gene is altered, but if the mutations are in *trans* then both genes are altered. Certain events only occur if the mutations are in *trans* and the proteins that the two altered genes encode for are not produced due to the mutations- this event is known as *compound heterozygosity*. On the other hand, certain events known as *cis*-regulatory events only occur if the gene is in *cis*. Genotype information is not sufficient to differentiate if a gene is in *cis* or *trans*; we need to know the haplotypes.

Finally, studying haplotype information of genomic regions that are functionally related have many useful applications. Certain regions of the genome contain groups of genes that are

Sample	Genotype	Haplotype
Disease	{(G,C), (G,A)} {(0,1), (0,1)}	'GG' and 'CA' '00' and '11'
Disease	{(G,C), (G,G)} {(0,1), (0,0)}	'GG' and 'CG' '00' and '10'
No Disease	{(G,C), (G,A)} {(0,1), (0,1)}	'GA' and 'CG' '01' and '10'

Figure 3: **Detecting Associations with Haplotypes.** Without knowing the haplotypes, there is no association between an individual with the disease and their genotype. But if we know the haplotypes, then we can detect an association with haplotype '`GG`' ('`00`') and the disease.

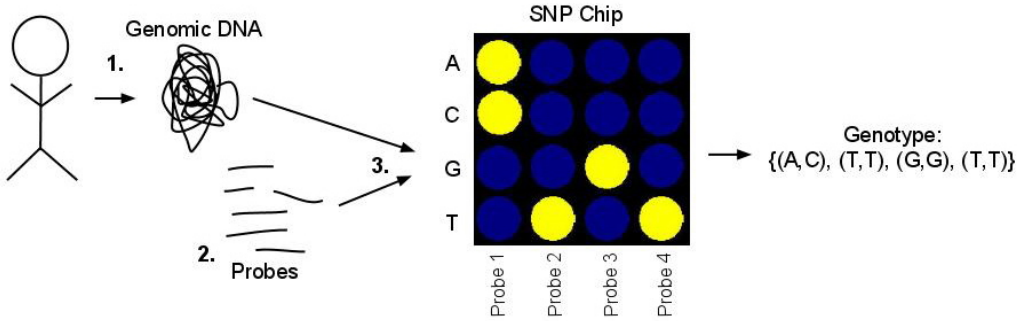
functionally related. Examples of such gene clusters include the HLA region on chromosome 6 which is 4 million bp long and the KIR region on chromosome 19 which is 3.8 million bp long. These regions are associated with autoimmune and infectious diseases. Knowing the haplotypes of these regions can help match organ donors to recipients [5, 6].

1.3 Technology. In this paper, we will discuss how the haplotype phasing problem has evolved as the technology to read the human genome has developed. The three types of technology relevant to haplotype phasing are *SNP arrays*, *sequencing platforms*, and *experimental techniques*. The rest of the paper is organized as follows.

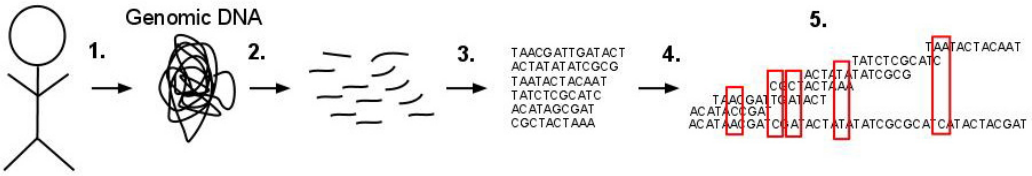
In section 2 and 3, we discuss how to build haplotypes using genotype information from *SNP arrays*. SNP arrays provide genotype information at low cost- one SNP array costs around $500 and can capture the genotype of around 1 million SNPs [7] (the whole genome has ~ 3 million SNPs) . SNP arrays take as input the DNA sample and a set of probes (these allow you to capture the genetic information at a particular location). It outputs the genotype of the DNA sample at the probe locations. See Figure 4(a). With SNP arrays we get genotype information but lose haplotype information.

In section 4, we discuss how to build haplotypes using data from *sequencing platforms*. Sequencing platforms allows one to read or sequence the whole genome directly rather than probe at specific locations. In recent years, sequencing platforms have been rapidly developing and are becoming more cost effective than they were before- sequencing an individual currently costs around $\$10,000$ [8]. Sequencing platforms take as input a DNA sample, cut up and amplify (replicate) the DNA, read each DNA fragment, and output a set of reads- each read is the nucleotide sequence corresponding to a small portion of a chromosome from the sample. These reads are mapped to the reference genome so we can get the location of each read. See Figure 4(b). With sequencing platforms we get partial haplotype information as the sequence of each read is from the same chromosome. However, these reads are short compared to the length of the chromosome (read lengths are 1000 bp long at best while chromosomes range from 51 million bp to 250 million bp) so computational methods are needed to piece together these reads to form longer haplotypes.

In section 5, we discuss *experimental techniques* to determine haplotypes directly. These are much more expensive and labor intensive than using genotyping or sequencing technologies. These methods physically separate the two copies of the chromosome before genotyping/sequencing.



(a) **SNP Chip Pipeline** 1. DNA extracted 2. Prepare Probes 3. Place probes on SNP Chip and genotype



(b) **Sequencing Pipeline** 1. DNA extracted 2. DNA cut and amplified 3. Each DNA fragment is sequenced 4. The reads of each DNA fragment is output 5. Find the genome position of each read by mapping it to a reference genome and find variants

Figure 4: **Two different technologies for reading the human genome.**

2 Phasing Populations

The invention of SNP arrays allows genotyping at low costs to be possible. Today, large amounts of population genotype data is available from the HapMap project [2] and the 1000 Genomes project [9]. Given the genotype of an individual, there are 2^{k-1} possible haplotypes. With genotypes of a population we can use assumptions about the evolution of haplotypes to phase (resolve the haplotypes of) the population.

2.1 Parsimony Approach. Due to the way haplotypes evolve, seeing a novel haplotype in a population is not as likely as seeing one of the haplotypes in the population again. The parsimony objective is to minimize the number of unique haplotypes in the population. The problem is formally defined as follows:

Haplotype Phasing with Minimum Unique Haplotypes
Input: $G = \{G_1, G_2, ...G_n\}$ is the genotype information of n individuals over k variant sites.
Output: $H = \{H_1, H_2, ...H_n\}$, the set of haplotypes such that H_i is consistent with G_i and the number of unique haplotypes in H is minimized.

Notice that the haplotype of an individual is *unambiguous* if he is homozygous at all sites or heterozygous in at most one site. For example, given the genotype $\{(0,0),(1,1)\}$, the haplotypes are unambiguously '01' and '01'. On the other hand, given the genotype $\{(0,1),(0,1)\}$ there are several possible haplotypes so we say the haplotypes are *ambiguous*. Using unambiguous haplotypes, we can phase populations as follows. Say the given population has two individuals with genotypes G_1 and G_2. If the first individual has an unambiguous haplotype, h and the second individual had a genotype such that $h \oplus h' = G_2$ for some haplotype h'. To minimize the number of unique haplotypes in the population, we would say the second individual's haplotypes

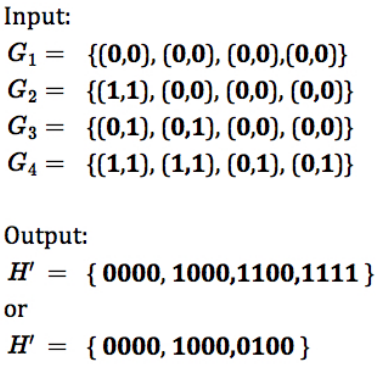


Figure 5: **Example of Clark's algorithm.** The input is the genotype of three individuals. The G_1 and G_2's haplotypes are unambiguous. G_3's haplotype can be inferred using either the unambiguous haplotype from G_1 or G_2. Using G_1's haplotype, G_3 and G_4 can be phased, however, if G_3 was phased using G_2's haplotype, then G_4 can not be phased. Clark's algorithm outputs four or three resolved haplotypes in H'.

are h and h'.

This objective problem is NP-hard by reduction from minimum clique cover [10]. However, one of the first algorithms for haplotype phasing is a greedy heuristic known as **Clark's algorithm** [11]. The algorithm starts by finding all unambiguous haplotypes in a population. The unambiguous haplotypes are added to the set, H', of haplotypes seen so far. In the example in Figure 5, there are two unambiguous haplotypes, '0000' and '1000' from the genotypes $G_1 = \{(0,0),(0,0),(0,0),(0,0)\}$ and $G_2 = \{(1,1),(0,0),(0,0),(0,0)\}$. The algorithm works iteratively by seeing if the genotype of any unphased individual, G_i can be explained by one of the haplotypes in H', that is if $G_i = h \oplus h'$ for $h \in H'$ and some haplotype h'. If so, it adds h' to H' if it is not already in there. In the example, we phase G_3 because G_3 ='0000' $\oplus$ '1111'. The algorithm continues until no more individuals can be phased. In the example, we phase G_4 because G_4 ='1111' $\oplus$ '1100'. H can be constructed during the process as we resolve an individual's haplotype.

Clark's algorithm relies on unambiguous haplotypes to start, but for a large enough population and a small enough region, the probability of someone having an unambiguous haplotype is high. Therefore, Clark's algorithm is usually used to phase short regions of the genome.

Application of short range haplotypes. Algorithms that produce short range haplotypes are primarily used to phase specific genes. For example, Clark's algorithm was used to find that certain haplotypes of the beta(2) adrenergic receptor gene are associated with bronchodilator response in asthmatics while individual SNPs in this region did not show association [12]. This illustrates the increase in association power that haplotypes have over individual SNPs. Another application of Clark's algorithm found the methylenetetrahydrofolate reductase polymorphism was associated with a haplotype found in Africa, Asian, and European populations [13]. This result indicates a selective advantage of the haplotype. This polymorphism, along with sufficient folic acid intake, has been shown to protect against colon cancer, leukemia, and fetal loss.

Clark's algorithm has $O(n)$ iterations. At each iteration it compares $O(n)$ unphased genotypes to $O(2n)$ phased haplotypes in H' in $O(n^2k)$. The total run time of Clark's algorithm is $O(n^3k)$. Although, Clark's algorithm runs in polynomial time, it provides no guarantees. If there are no unambiguous haplotypes, the algorithm can not start. The probability that there

is an unambiguous haplotype in the population increases as the population size increases but decreases as the size of the region increases. This fact limits the size of the region Clark's algorithm can phase. Also, there is no guarantee that the algorithm will phase all the individuals- the algorithm will stop if non of the haplotypes in H' can explain the remaining individuals. In fact, the algorithm is non-deterministic. The order in which the input is processed affects the number of resolved haplotypes. Take the example in Figure 5. If we used G_2 instead of G_1 to phase G_3 (‘1000’$\oplus$‘0100’), then we would not be able to phase G_4. Instead of using a heuristic to solve the problem, Gusfield [10] gives an integer linear program that outputs the optimal solution. However, the number of variables is exponential in the number of variant sites, and this algorithm does not scale well as the number of variant sites increases. Both algorithms also assume that there is no recombination in the region we are phasing. Assuming no recombination is valid as long as the region we are phasing is small.

2.2 Stochastic Approach. In conjunction with a parsimonious approach, another common approach uses a stochastic method where the objective is to find the most likely set of haplotypes H given the input genotypes, G. Formally, this objective is

$$\arg\max_{H} \Pr(H|G)$$

In the stochastic algorithms we describe, $\Pr(H|G)$ is based on assumption of how haplotypes evolve and rely on the mutation rate, θ, and recombination rate ρ. Here is the general formula:

$$\Pr(H|G) = \Pr(H_1|H\backslash H_1, G)\Pr(H_2|H\backslash\{H_1, H_2\}, G)...\Pr(H_n|G)$$

If we know the general distribution of $\Pr(H_i|H^*, G)$ for some haplotyope set H^*, then we can approximate $\Pr(H|G)$ using a Markov Chain Monte Carlo (MCMC) approach. Each state in the Markov chain is a possible set of haplotype pairs, H. At each state transition, only one individual's haplotypes changes. The MCMC algorithm is shown in Algorithm 1. The key part of this algorithm is step 2, where we sample H_i from the conditional probability $\Pr(H_i|H\backslash H_i, G)$. So far we assumed the distribution of $\Pr(H_i|H\backslash H_i, G)$ is known, but in fact it is unknown and stochastic algorithms aim to approximate it based on the following assumptions about the evolution of haplotypes as stated in Li *et al.* [14]:

- A new haplotype in the population is more likely to match an existing haplotype that has occurred more frequently in the population.
- The probability of seeing a novel haplotype increases as the sample size, n, increases.
- The probability of seeing a novel haplotype increases as mutation rate, θ, increases.
- If the next haplotype is not exactly the same as an existing haplotype, it will tend to differ by a small number of mutations from an existing haplotype rather than be completely different from all existing haplotypes.
- Due to recombination, the next haplotype will tend to look somewhat similar to existing haplotypes over contiguous genomic regions. The length of the region is longer in areas where ρ is smaller.

Next we describe two different methods of approximating $\Pr(H_i|H\backslash H_i, G)$.

In the **PHASE** algorithm [15], the approximate $\Pr(H_i|H\backslash H_i, G)$ by approximating the general distribution $\Pr(H_i|H^*, G)$ as follows. If H_i is consistent with G_i, then we have

$$\Pr(H_i|H\backslash H_i, G) \propto \Pr(h_{i1}|H\backslash H_i)\Pr(h_{i2}|(H\backslash H_i)\cap h_{i1})$$

Algorithm 1 MCMC Algorithm to estimate most likely H

Initialize H (randomly but still consistent with G)
For $t = 1, 2, ...$
1: Choose an individual, i, at random
2: Sample H_i^{t+1} from $\Pr(H_i|H^{p,t}\backslash H_i, G)$
3: Construct $H^{p,t+1}$

They propose the following approximation for the general probability, $\Pr(h^*|H^*)$, that is the probability of seeing a haplotype, h^*, given a set of haplotypes, H^*.

$$\Pr(h^*|H^*) = \sum_{h' \in H^*} \frac{n_{h'}}{n}(1-\lambda_n)(I-\lambda_n P)^{-1}_{h^*h'}$$

where $n = |H^*|$, $n_{h'}$ is the number of occurrences of h' in H^*, $\lambda_n = \frac{\theta}{n+\theta}$, θ is the mutation rate, I is the identity matrix, and P is the transition matrix. Notice that recombination is not incorporated into $\Pr(h^*|H^*)$. Later versions of PHASE incorporate recombination into their approximation for $\Pr(H_i|G, H\backslash H_i)$ [16].

Another method of approximating $\Pr(H_i|H\backslash H_i, G)$, uses a *Hidden Markov Model (HMM)*. In the HMM model, observed data is generated by an unobserved Markov process. See Figure 6. Here the unknown Markov process is the phase on sites $\{1, 2, ...k\}$, and the genotype is the observed data. We can think of the haplotypes in $H\backslash H_i$ as the template haplotypes. At each step of the Markov process there are several possible hidden states (template haplotypes) we can be in. Because humans are diploid, each state is actually a pair of single haplotypes. If $|H\backslash H_i| = n$, there are $2n$ single haplotypes and $(2n)^2$ pairs of haplotypes, giving us a total of $(2n)^2$ possible hidden states at each site. Let X_j denote the set of possible states at the jth location- biologically it represents the set of possible template haplotype that H_i descended from at site j. It remains to define the transition and emission probabilities. Let $X_j[p, q]$ denote the hidden state corresponding to the haplotype pair of the pth and qth haplotype of $H\backslash H_i$ at the jth position. The transition probabilities from $X_j[p, q]$ to $X_{j+1}[p', q']$ are a function of the recombination rate while the emission probabilities, $\Pr(G_i|X_j[p, q])$ is a function of the mutation rate. Finally, we can run the forward-backward algorithm to get the approximation for the distribution of $\Pr(H_i|H\backslash H_i, G)$.

Application of long range haplotypes. Using population phasing methods, Auton *et al.* [17] phased 4000 individuals from four continental regions at 400,000 sites and found patterns of diversity on a genome wide scale over four continents. For instance, they found that Japanese populations are less diverse than Taiwanese populations. This difference in diversity can be explained by less migration to Japan or perhaps a bigger/more recent bottleneck event in Japan. With many samples coming from different parts of Europe, they were able to study haplotype diversity on a finer scale. In particular, they found a north-south gradient in haplotype diversity within Europe. More haplotype diversity in the south compared to the north can be explained by migration to Europe from Africa and a harsher climate in the north which causes population bottlenecks. Another application to genome-wide haplotypes is to find indications of positive selection by looking for unusually long haplotypes. Sabeti *et al.* found genes that were under positive selection in different populations. In the European population, they found two genes ($SLC24A5$ and $SLC45A2$) involved in skin pigmentation were under positive selection. In particular, the $L374F$ substitution in $SLC45A2$ is associated with fair skin and non-black hair and is at 100% frequency in European populations but is absent in Asians and Africans. A third application is genome wide haplotype association studies. In Tregouet *et al.* [18] they found that the $SLC22A3 - LPAL2 - LPA$ gene cluster is associated

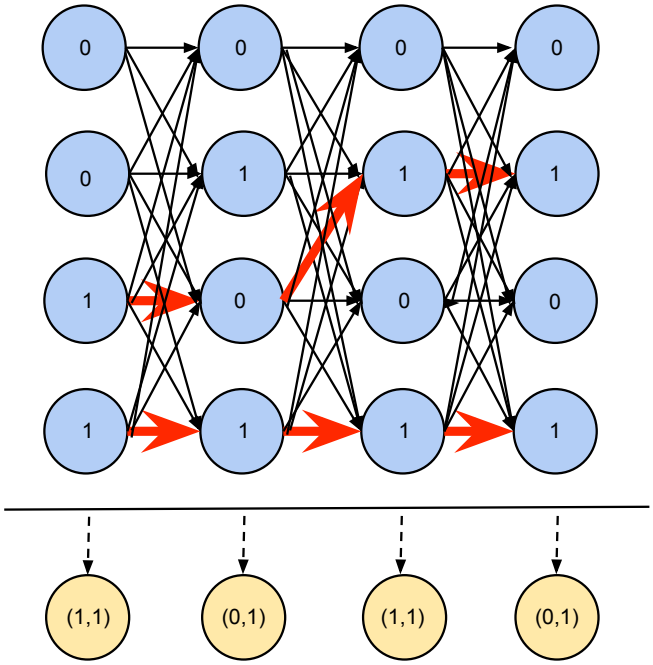


Figure 6: **Hidden Markov Model for** $\Pr(H_i|G, H\backslash H_i)$. In this example, we are given the genotype information (observed data) over 4 sites (columns) and 4 template haplotypes (rows) from the population. Transition probabilities are associated with the solid arrows while emission probabilities are associated with the dashed arrows. (There are actually 4 emission probabilities per site from each of the template haplotypes). The two red paths represent the most likely haplotypes for this example found by running the forward-backward algorithm on the HMM. Lastly, we only show 4 template haplotypes; however, the real HMM will have $4^2 = 16$ rows- each row representing a pair of single haplotypes.

with coronary artery disease. Lastly, several public reference panels such as HapMap [2] and 1000 Genomes [9] have used population based methods to phase.

PHASE was considered the gold standard for accuracy for some time. In fact, PHASE v2.1 was used to phase the data of the HapMap [2] project. However, like Clark's algorithm, PHASE assumes no recombination. The assumption of no recombination is still valid in shorter regions of the genome since the probability of recombination in a region decreases as the size of the region decreases. In a population, the average length of a region with no recombination is $22,000$ bps [19]. Although quite accurate for regions with no recombination, the run time for each iteration in PHASE is $O(2^k)$. This is due to the fact that there are 2^k possible haplotypes we have to calculate conditional probabilities for at each iteration.

Using HMMs is a straight forward way to incorporate recombination into the approximation of the distribution. Two different algorithms that use this approach are **MACH** [20] and **IMPUTE** [21]. There are $(2n)^2$ pairs of template haplotypes, and k sites for a total of $(2n)^2k$ hidden states. Notice that the run time grows linearly in the number of variant sites. This coupled with the fact that recombination is incorporated into the approximation allows phasing of very long regions. Although, these algorithms are faster than $PHASE$, the run time still

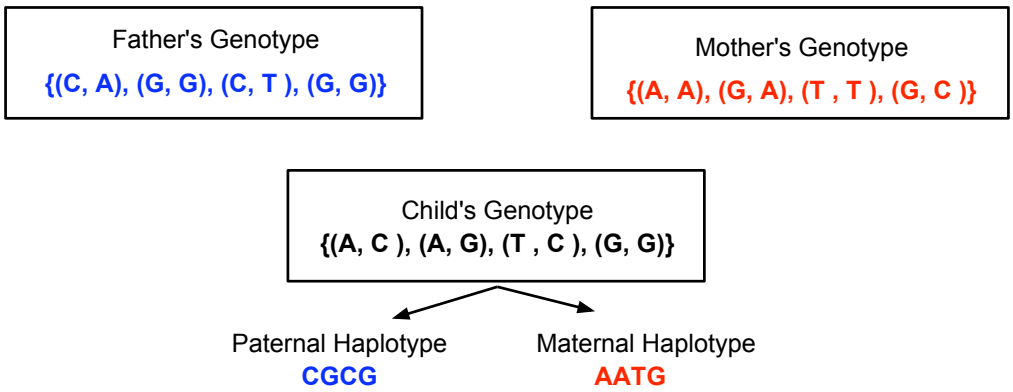


Figure 7: **Identical By Descent.** Using IBD we can infer the phase of the child is $AATG$ from the mother and $CGCG$ from the father.

grows quadratically as the number of haplotypes in H increase. In order to limit the size of H, MACH uses a random subset of known haplotypes in H while IMPUTE uses a subset of individuals similar to the current individual. Using only a subset of template haplotypes limits the size of H but pays the price of accuracy.

3 Phasing Related Individuals

As genotyping and sequencing technologies become cheaper, we not only get more population data, but we also are able to genotype families. Recall that an individual inherits one set of chromosomes from the mother and another from the father. Assuming no mutations, whole haplotypes are *identical by descent (IBD)* between a parent and a child. This means that we can phase the child using genotype data from duos (parent and child) or trios (both parents and child) using the following rules.

- For all sites where the child is homozygous, the phase is trivial.
- For sites where the mother is homozygous and the child is heterozygous, we know the allele that was inherited from the mother.
- For sites where the father is homozygous and the child is heterozygous, we know the allele that was inherited from the father.

In duos we can phase all sites of the child except those where the parent and the child are heterozygous. And in trios we can phase all sites of the child except those where both parents and the child are heterozygous. To phase these sites, we can use population haplotype data to infer the phase [22]. Because of recombination, we can only phase the child. Of course if we had genotype data from grandparents, then we could phase the parents.

So far, we have assumed no mutation. Here we briefly describe how we can infer mutation and recombination sites. If we know a particular allele was inherited from the mother, we can infer that the other allele at the same site was inherited from the father. This fact can be used to detect mutations. If we have inferred that a child inherited a particular allele at a site, but the child's haplotype does not match the parent's, then we know a mutation must have occurred.

In order to phase the parents, we need to find the recombination sites. To do this, we need genotype information from other family members. For example, if we knew the genotype of the grandparents, then we could phase the parent, and then compare the parent's haplotype to the child's and to find the recombination sites. If we can not phase the grandparents, we could also phase other members of the family to infer recombination sites. For instance, if we know the genotype of a sibling, then we can also phase the sibling. By comparing the paternal haplotype of both children, we can infer the recombination sites on the paternal haplotypes of the two children. We can infer the recombination sites on the maternal haplotypes of the two children similarly.

Application of family phased haplotypes. Using family information to phase provides accurate, long range haplotypes that can span regions with recombination without sacrificing accuracy. It provides more accurate haplotypes as one can use family data detect and recover errors from genotyping/sequencing. With familial information, one can also identify exact recombination location. In Kong *et al.* they phased the Icelandic population where the genealogy is known. By looking at the phase between siblings and their parents, they were able to detect recombination locations. They also used familial information to study the inheritance of recurrent structural deletion related with schizophrenia and inferred the origin of the deletion. Family information is important when detecting *compound heterozygosity*. In the same gene, usually it is only one chromosome copy that gets mutated. However, compound heterozygosity is when there is a mutation on both copies. The phase information will determine the presence of compound heterozygosity. Roach *et al.* [23] used the phase of trios to identify the causal gene by detecting compound heterozygosity.

With family data, we can get whole genome haplotypes at high accuracy. However, family data is not always available and can increase the cost of haplotyping two to three folds.

4 Phasing using Sequencing Technology

Phasing using sequencing data is very attractive given the proliferation of inexpensive techniques that have the throughput to sequence entire human genomes. The output reads of the sequencing platform each represent a small section of one chromosome, thus giving us partial phase information. We can build a haplotype by connecting overlapping reads. To gain an intuition about how reads can be connected to form haplotypes we define a *SNP-graph*. Each variant site corresponds to a node in the graph. When a read overlaps two variant sites, we add an edge to the corresponding nodes. See Figure 8(b). It is easy to see that two sites can be phased if and only if they are connected in the SNP-graph. The *length* of the haplotypes depend upon the size of the connected components. Unfortunately, sequencing platforms are erroneous. The majority of *sequencing errors* occur when the nucleotide found in a read differs from the actual nucleotide in the DNA sample (*i.e.* a read has the symbol `A` at a position where it should be a `C`). The *error rate* measures the probability of a read error at a particular position. Error rates differ for different sequencing platforms and can range from $1\% - 10\%$ and greatly affect the *accuracy* of the haplotypes we phase. To some extent, the *accuracy* and the *length* of the haplotype can be decoupled. The algorithms focus on outputting accurate haplotypes despite sequencing error while the length of the haplotype is mostly determined by the different sequencing platforms which we will describe later.

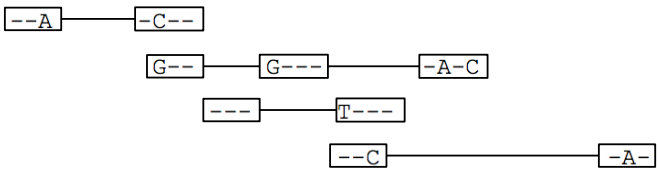
To formalize the problem, the set of sequenced reads can be represented as a *SNP matrix*, X, where the reads are rows and the variants are columns. The variants are ordered based on their natural genomic location. If a read does not cover a particular site, we label the matrix entry with '-'. The entries of the SNP matrix, X, are in the alphabet $\{0, 1, -\}$. See Figure 8(c) for an example of how a set of mapped reads can be represented as a SNP Matrix.

The most common formulation of the problem is *Minimum Error Correction (MEC)*. Let the MEC score given a set of reads X and a haplotype $H = \{h_1, h_2\}$ be defined as follows:

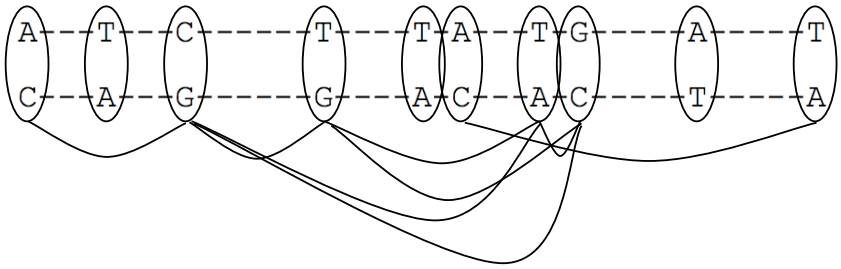
$$MEC(X, H) = \sum_i MEC(X_i, H) = \sum_i \min\{d(X_i, h_1), d(X_i, h_2)\}$$

When phasing an individual, we can discard all homozygous variants, so that h_2 is the compliment of h_1 ($h_2 = \overline{h}_1$). Under the MEC objective, we have the following formulation:

Haplotype Phasing with Minimum Error Correction
Input: SNP Matrix, X



(a) Mapped input Reads from Sequencing Technology



(b) Mapped input Reads from Sequencing Technology

0	-	0	-	-	-	-	-	-	-
-	-	1	1	-	-	1	1	-	-
-	-	-	-	0	-	-	-	-	-
-	-	-	-	-	1	-	-	-	1

(c) Reads represented as a SNP Matrix

Figure 8: **Different representations of sequencing reads.**

Output: The haplotypes, $H = \{h, \overline{h}\}$, such that the MEC score, $MEC(X, H)$, is minimized

The MEC objective is NP Hard- shown by Cilibrasi *et al.*[24] by giving a polynomial time reduction from *Max-cut*.

4.1 Greedy Algorithms. In the context of haplotyping using sequencing technology, the first algorithms proposed were heuristics that greedily cluster the reads into two sets corresponding to the two haplotypes. Fast Hare [25] was one of the first of these algorithms. Fast Hare begins by sorting the reads in X by their starting position. It uses the first fragment to initialize the two sets, R_1 and R_2- one set contains the first fragment and the other set contains its compliment. It iterates through the rest of the reads in order. For each read, r, it calculates the partial consensus haplotype of each set, $H(R_1)$ and $H(R_2)$ and calculates a score between r and each set, $D(r, H(R_1))$ and $D(r, H(R_2))$. For a read r and a haplotype h, the score is defined as follows:

$$D(r, h) = \sum_{i=1}^{k} d(r[i], h[i])$$

$$d(r[i], h[i]) = \begin{cases} 0 & \text{if } r[i] = \text{'-'} \\ 1 & \text{if } h[i] = \text{'-'} \\ -1 & \text{if } r[i] \neq h[i] \end{cases}$$

For the set R_1 or R_2 that produces the highest score, it add the read to the set. After processing all reads, the algorithm outputs the consensus haplotypes of the two sets.

Another greedy algorithm of historical interest is one proposed by Levy *et al.*. This algorithm was used to phase Craig Venter's genome which was sequenced using Sanger sequencing. The main difference between the two algorithms is the order in which the reads are processed. In Levy *et al.*, instead of sorting the reads and picking the first one, they pick the read that phases the most variants (row with the most non-'-' entries in the matrix) to start the initial haplotype. At each iteration, they calculate $D(r, H(R_1))$ and $D(r, H(R_2))$ for each $r \notin R_1 \cup R_2$ and adds the r which has the highest score to the corresponding set. It outputs the consensus haplotypes of both sets.

Without errors, both Fast Hare and Levy *et al.*'s algorithm will output the correct haplotypes. These algorithms also run fast. Fast Hare runs in $O(nlogn + nk)$ time. The $O(nlogn)$ is the time to sort the reads and the $O(nk)$ is the time to construct the haplotypes. The Levy's greedy heuristic runs in $O(n^2k)$ time as there are n iterations and in each iteration it calculates a score of each read. However, given that these algorithms are heuristics, there are no guarantees that the output haplotype has minimal or even close to minimal MEC score. In fact, the MEC score can be quite large. For example, Fast Hare processes the fragments in a fixed order according to position. Therefore, early errors could cause incorrect partition assignment of later fragments. These early errors can accumulate. Instead of choosing reads in a fixed order, Levy *et al.* starts with the fragment with the largest phase. This is also not necessarily the read with the lowest chance of error. In fact, if the sequencing error is per base, the longer the fragment, the lower the probability that the fragment is correct. And thus there is a higher probability that an error will accumulate and produce inaccurate haplotypes.

4.2 Stochastic Approach. The next algorithm uses a stochastic approach to find the most likely haplotype under the MEC objective. Formally, it aims to find the most likely haplotype $H = \{h, \overline{h}\}$ conditional on the given sequencing data:

$$\arg\max_{H} \Pr(X|H)$$

HASH, proposed by Bansal *et al.* [26], takes a *Markov Chain Monte Carlo (MCMC)* approach to search through the space of all possible haplotypes. They sample from the probability distribution $\Pr(H|X)$. Note that $\Pr(H|X) \propto \Pr(X|H)$. Each state in the Markov Chain is a haplotype and its compliment. The subset of sites, S, where two haplotypes differ are used to describe the transition from one state, H, to another H_S. See Figure 9 for an example transition in the Markov chain. The general algorithm is described in Algorithm 2.

Algorithm 2 HASH

1: Choose an initial haplotype configuration H^0
2: For $t = 1, 2, ...$
3: With probability 1/2, set $H^{t+1} = H^t$
4: Otherwise, sample a subset S from Γ with probability $1/|\Gamma|$
5: With probability min , set $H^{t+1} = H^t_S$
6: Otherwise, set $H^{t+1} = H^t$

In step 4, we sample S from Γ, which is the set of all subset of sites S corresponding to transitions in the Markov Chain. The key to making the algorithm run fast is to choose a good

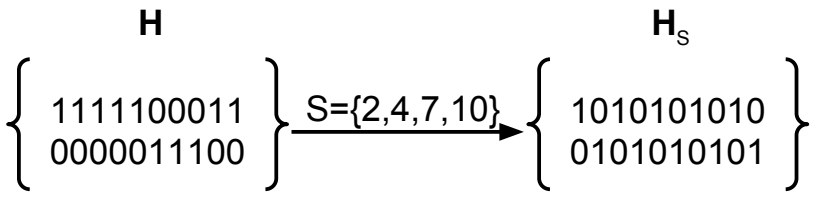


Figure 9: **Markov chain transition.** In this transition, the phase is switched at sites $2, 4, 7$, and 10.

Γ. The choice of Γ must satisfy certain constraints in order for the corresponding Markov Chain to be ergodic and for $\Pr(X|H)$ to have a stationary distribution. Γ should also be chosen so that the mixing time is fast. The minimal choice for Γ is $\Gamma_1 = \{\{1\}, \{2\}, ..\{k\}\}$. Choosing $\Gamma = \Gamma_1$ means at each step of the Markov chain, only one site changes phase. Clearly the mixing time (time to convergence) for Γ_1 is large. In Bansal *et al.* they show that any Γ such that $\Gamma_1 \subset \Gamma$, will satisfy the necessary constraints. In Bansal *et al.* they propose a graph partitioning algorithm to find a good Γ. The graph used in the algorithm is a weighted version of the SNP-Graph described earlier.

The results in Bansal *et al.* show that HASH outputs haplotypes with lower MEC scores than greedy algorithms. However, the running time of HASH can be quite slow especially for a bad choice of Γ.

4.3 Combinatorial Approach. Bansal and Bafna [27] proposed the algorithm **HapCUT** that iteratively finds haplotypes with lower MEC scores by constructing a weighted SNP-graph at each iteration and finding a cut with maximum score. Let $G_X(H)$ be a weighted SNP-graph. Recall that the nodes of the SNP-graph are the variant sites of X and there is an edge between two sites if a read in X covers them. The weight of an edge between two nodes, $w(j, k)$, is proportional to the difference in number of reads inconsistent with the phase of j and k given by H and those consistent with it. A cut is a partition of the vertices into two sets, S and $V(G_X(H)) - S$. The weight of a cut in the graph, $w(S)$, is naturally defined as $w(S) = \sum_{j \in S, k \in V(G_X(H))-S} w(j, k)$. This algorithm is shown in Algorithm 3. The algorithm is based off the following claim proved in their paper.

Algorithm 3 HapCUT

1: Choose an initial haplotype configuration H^1 randomly
2: For $t = 1$, 2, ...
3: Construct the graph $G_X(H^t)$
4: Compute a cut S in $G_X(H^t)$ such that $w_H(S) \geq 0$
5: If $MEC(H^t_S) \leq MEC(H^t), H^{t+1} = H^t_S$
6. Else $H^{t+1} = H^t$

Claim 1. *[27] For any haplotype pair H, let $S \in X$ be a positive weighted cut $w_h(S) > 0$ in the graph $G_X(H)$. Then $MEC(H_S) = MEC(H) - w_H(S) < MEC(H)$. If S is a max-cut in the graph $G_X(H)$, the H_S is an optimal MEC solution for X.*

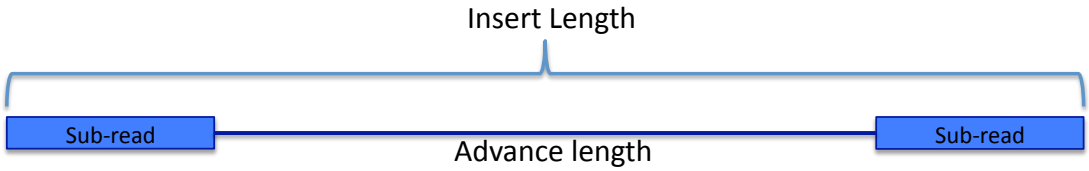


Figure 10: **Diagram of a sequencing read.**

Using this claim, the max-cut of $G_X(H)$ would give us the optimal haplotype, but the max-cut problem is also NP-hard. The claim also suggests that as long as S is a subset that produces a positive weight cut, then switching values of the current haplotype, H, at all sites in S will decrease the MEC score. The larger the weight of the cut, the more the MEC score will improve. To find a good cut with positive weight (step 4) Bansal and Bafna use a greedy heuristic.

In practice, HapCUT runs faster than HASH and produces haplotypes with comparable MEC scores. It is worth it to note that greedy heuristics for max-cut are designed for graphs with only positive edges. $G_X(H)$ contains some negative edges especially if the current haplotype H is close to the optimal MEC haplotype. However, the greedy heuristic used in HapCUT takes negative weights into account and works well in practice.

4.4 Sequencing Technology and Haplotype Length. So far we have discussed algorithms that focus on the accuracy of haplotypes. The length of the haplotypes are determined by the parameters of the sequencing technology which we now discuss.

Metric for Haplotypes Phased with Sequencing Data. First, we define the standard metric used to measure haplotype length called *N50*. When phasing with sequencing technology data, many haplotype contigs will be constructed. Let *span* be the physical distance of a haplotype contig in base pairs. The N50 is the span such that 50% of all sites are in blocks of span N50 or greater. The N50 metric tends to inflate the haplotype size when there are contigs of long distance that do no phase many SNPs (many gaps). Therefore, we use the *Adjusted N50 (AN50)* metric which weights the span of each contig by the fraction of variants phased versus variants spanned.

All sequencing platforms output a set of reads, but the size and structure of the reads can vary across different platforms. Figure 10 shows the general dichotomy of a read. Essentially, a *read* can have several *sub-reads*. We use k to denote the number of sub-reads in a read. Mate-pair reads have two sub-reads ($k = 2$). The *read length*, L is the total length of the sub-reads in a read. For a read with k sub-reads, the length of each sub-read will be $\frac{L}{k}$. The *advance length*, A, between two sub-read is the distance between the first and second sub-read. The *insert length*, I is the total length of the read. For a mate-pair read, $I = L + A$. Sequencing coverage, c is the average number of a times a particular site is sequenced. Different sequencing technologies provide different values for these parameters which affects the haplotype length.

Development of Sequencing Technology. The first sequence of the genomic individual, was produced using *Sanger sequencing*. The sequencing was paired-end with a read length of around 1000 bp with two sub-reads linked at $2,000$ bp, $10,000$ bp, and $50,000$ bp apart, and $c = 6$x. The phasing was quite effective with an AN50 of $270,000$ bp [28]. Sanger sequencing provides long and accurate reads but low throughput and expensive library preparation. By contrast,

Sequencing Platform	Read Length (bp)	Advance Length (bp)	Cost/Coverage
Sanger Sequencing	1000	Fixed: $2,000$, $10,000$, and $50,000$	\$70 million/7.5x
Next Generation Sequencing	100	Fixed: $200 - 5000$	$\$10,000$/30x
Strobe Sequencing	1000	Variable: $0 - 9000$	\$100/0.01x

Table 1: **Comparison of Sequencing Technology**

Next-Generation Sequencing platforms are cheaper allowing for higher coverage ($c = 30$x), but have much shorter reads ($30 - 100$ bp), shorter fixed advance lengths ($200 - 500$ bp), and are more error prone. The haplotype lengths achievable with next-generation sequencing does not compare with Sanger sequencing. *Third-Generation Sequencing* platforms are the newest development. One example of a third-generation platform is the *Strobe Sequencer* developed by Pacific Biosciences. In strobe sequencing the read lengths and insert sizes are longer than next-generation sequencers (currently ≤ 900 bp and $\leq 20,000$ bp respectively). But they also allow for variable advance lengths and multiple sub-reads (strobes). See Table 1 for a complete comparison of the different sequencing technologies. With the ability to vary parameters such as advance length and number of sub-read, we can design sequencing methods to achieve longer haplotypes. Next, we examine the effect of different parameters- advance length distribution, maximum insert size, read length, and strobes- on haplotype lengths.

Third-generation sequencers offer more flexibility than Sanger and Next-Generation sequencers. In particular, third-generation sequencers allow for variable *advance lengths*. It turns out that variability in advance lengths greatly affects haplotype length. For example, if all of the reads had an advance length of 9000 bp ($L = 900$ bp, $c = 20$x), we get an AN50 of $6,700$ bp. However, if half of the reads have an advance length of 9000 bp and the other half 3000 bp[1], the AN50 increases by an order of magnitude. This leads to the question: "What distribution of advance lengths will give us the highest AN50?". Lo *et al.* [28] show that changing the distribution of advance lengths from a uniform distribution to one that is skewed towards longer advance lengths gives a two-fold increase (from $60,000$ bp to $150,000$ bp) in AN50. See Table 2 to see how the distribution of advance length affects AN50 [28]. Figure 11 shows the effects of different parameters on the haplotype length. Another key parameter to haplotype phasing is the *maximum insert length*. In Figure 11(a), we see that AN50 increases as max insert length increases. In Figure 11(b), we see that AN50 increases as *coverage* increases but reaches saturation quickly. We can think of increasing coverage as increasing the probability of an edge existing in the SNP graph. An increase in edge probability leads to larger connected components (or longer haplotype contigs). In Figure 11(c), we see that AN50 increases as read lengths increase but also saturate quickly. Increasing the read length also increases the edge probability in the SNP graph, but once read lengths are sufficient, the haplotype lengths saturate. In Figure 11(d), we show how different number of *strobes* effect haplotype length ($c = 10$x here). To compare the design of different number of strobes, we fix the total read length (for a read with k strobes, each sub-read will have length L/k). Longer sub-read lengths help cover the relatively high proportion of variants that are clustered close together. Therefore, increasing number of strobes helps increase the variation in advance lengths against the penalty of smaller sub-reads.

Applications of phased haplotypes using sequencing data. An advantage of using sequencing data instead of genotype data from SNP arrays, is that we are able to capture SNVs.

[1] unless otherwise noted all experiments in this section are conducted with $L = 900$ bp, $c = 20$x, and Maximum-$A = 9000$ bp.

Distribution of Advance Lengths	AN50
100% of reads with $A = 3000$	$5,400$
100% of reads with $A = 9000$	$6,700$
50% of reads with a $A = 3000$, 50% $A = 9,000$	$54,000$
Uniform Distribution	60,000
Beta Distribution, $\alpha = 1.6$, $\beta = 0.5$	150,000

Table 2: **Distribution of Advance Lengths vs. Haplotype Length**

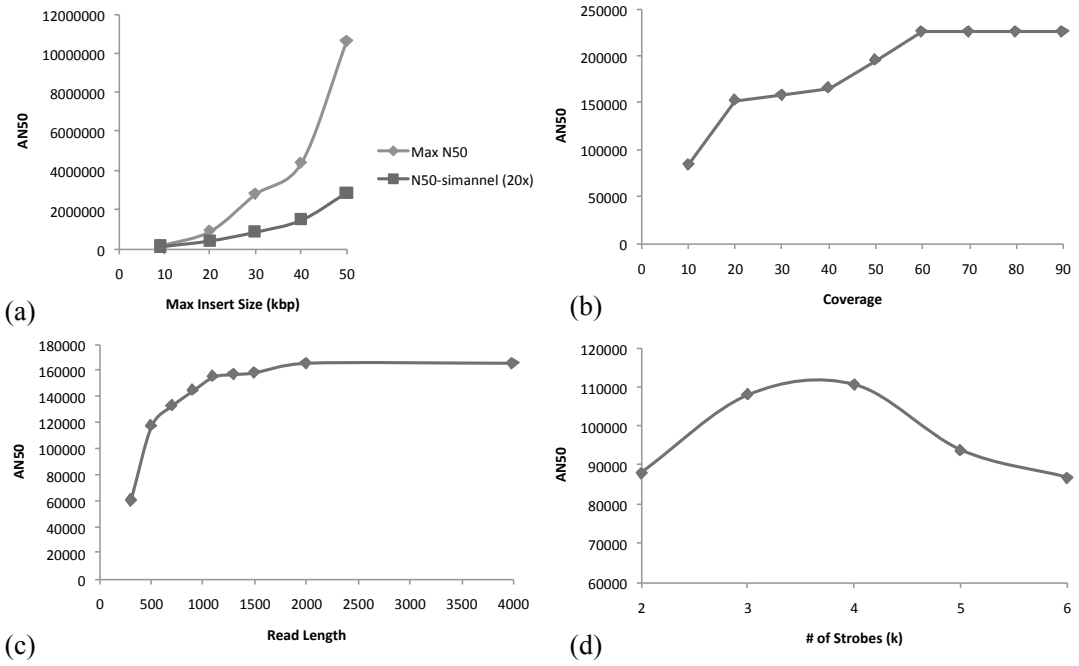


Figure 11: **Other Parameters and AN50**

Since SNVs are unique to an individual, their location is not known and it is harder to get the necessary probe for the SNP array. However, sequencing data reads all of the genome and as a result, we can detect SNVs. In addition, we can phase across regions with high recombination rate without losing accuracy. Thus, we can phase large regions of the genome. For example, the *human leukocyte antigen (HLA)* region on chromosome 6 spans about 4 million bp . It is a highly polymorphic region containing many genes involved in our immune system. One application of phase information for this region is to match organ donors to hosts. With current sequencing technology ($c = 10$x, $L = 900$ bp, $A = 20,000$ bp), we can span 80% of the region with 5 haplotype contigs [28].

5 Experimental Phasing

Recently, many experimental phasing methods have developed. These are considerably more expensive than genotyping and sequencing, but they provide whole genome haplotypes. Here we provide a brief overview of recent work in the area. There are two main types of experimental phasing. One method separates pairs of chromosomes in the metaphase stage. The other uses a technique similar to sequencing to get longer phases.

5.1 Microfluidics. Fan *et al.* [29] developed a custom microfluidics chip that captures a cell in metaphase and isolates the 46 individual chromosomes. To facilitate extracting information from these chromosomes, they amplify each chromosome to provide more genetic material. After finding the identity of the chromosomes, they separate the chromosomes into two sets such that haplotype information is preserved. They genotype each set using SNP arrays. Since they genotype each set separately, the output genotypes will all be homologous since there is only one of each pair of chromosome in a set- thus, providing haplotype information.

Fan *et al.* reported to phase the whole genome with 99.8% accuracy. However, the haplotypes provided are very sparse (only about 25% of the SNPs were phased). The density of the haplotypes could be improved if whole genome sequencing was performed instead of whole genome genotyping.

While no concrete cost was presented in the paper, microfluidic chips are custom made and as a result generally expensive. Labor is also involved in this method- for example, they microscopically capture the cell in metaphase. Furthermore, the method must currently be repeated about $2-3$ times to ensure accuracy. This makes it difficult for an average lab to buy and operate this method.

5.2 Single Genome Amplification. Zhang *et al.* [30] is currently developing an experimental phasing method that starts by capturing several cells in metaphase to separate the pairs of chromosomes. This method is more cost effective than the microfluidics approach since it does not use a custom chip. Instead of separating each chromosome individually, they pool several chromosomes into one tube. They choose the number of chromosomes allowed in one tube such that with low probability, two different chromosome copies do not fall in the same tube. They pool chromosomes several times- in a typical experiment they have around 10 tubes. Each tube is then amplified. Because of amplification bias, some parts of a chromosome may not get amplified. The amount of amplification bias determines the density of the haplotypes. After amplification, they either perform sequencing or genotype each tube. After collecting sequencing data from each tube, statistical test can be used to determine if two copies are in the same tube.

Both Single Genome Amplification and Microfluidics give sparse haplotypes. The haplotype density can be improved by performing whole genome sequencing on the individual and phasing the sequenced data using methods described in section 4. Then we can combine this to fill in the gaps of the sparse haplotypes from experimental phasing.

5.3 Fosmids. Sequencing technologies provide phase information per read. However, because each read only spans a few SNPs at best, the read by itself does not give very useful haplotype information. Sanger sequencing uses fosmid libraries to get long reads. Kitzman*et al.* and Suk *et al.* [31, 6] also use fosmids to get long DNA segments, but instead of reading each fosmid directly, which is costly, they use cheaper next-generation sequencing technology. In Kitzman *et al.* they created single fosmid libraries with insert sizes of $\sim 37,000$ bp. The fosmids were randomly divided into 115 pools ($\sim$ 5000 fosmids in each pool) such that the probability of a chromosome pair in the same pool is low. Each pool was barcoded and then they were all sequenced together. The contents of a pools can be recovered due to the barcoding. Each pool essentially provides a haplotype for about $\sim 3\%$ of the genome. Overlaps between pools are used to assemble longer haplotypes (contigs). Having to assemble the haplotype contigs of each pool could potentially cause errors. However, it is more cost effective than the custom microfluidics technique. Furthermore, they combined the assembled haplotypes from fosmids with regular sequencing data using a method similar to HapCUT. In addition to the cost of sequencing, Kitzman *et al.* report an additional cost of $4000.

Using this method, they were able to phase 94% of a Gujarati Indian woman with an N50

Input Data	**AN50 (bp)**	**Cost**
Population Genotype Data	$22,000$	\$700 (1 million SNPs)
Family Genotype Data	whole chromosome length * 0.95 [32] * 0.33	$2-3$x cost of genotyping an individual
Sanger Sequencing	$270,000$	\$70 million
Next Generation Sequencing	1,800	$\$10,000$
Strobe Sequencing	$150,000$	\$2 million
Family Sequence Data	whole chromosome length * 0.95	$2-3$x cost of sequencing an individual
Microfluidics	whole chromosome length * 0.25	unpublished
Fosmids	$386,000(94\%) = 362,840$	\$4000 + cost of sequencing

Table 3: **Comparison of Phasing Algorithms**

of $386,000$ bp. With the phase information, they were able to lower the number of candidate genes associated with a recessive Mendelian disorder from 44 to 10 by identifying 10 genes with compound heterozygosity.

6 Conclusion and Future Direction

Table 3 compares the lengths of the haplotypes that are phased using various methods. Genotyping population data and phasing is the least expensive way to phase. But it is limited by recombination. Population phasing methods that incorporate recombination are time consuming and not as accurate as other phasing methods. When family data is available, computational phasing can produce whole genome haplotypes. However, family data is not always available. Phasing using sequencing data is becoming more and more feasible as sequencing platforms are rapidly developing. Although costly and labor intensive, experimental phasing techniques are able to directly determine whole genome haplotypes.

The haplotype phasing problem has come a long way over the past 20 years. We now have the potential to get whole genome haplotypes. The current bottleneck of haplotype phasing are the high costs of experimental methods and the limitations of sequencing technology. Hybrid experimental and sequencing methods are currently in development and the next computational problems will likely involve combining sequencing data and experimentally phased haplotypes efficiently. For example, the microfluidic approach of Fan *et al.* and the amplification approach of Zhang *et al.* provide sparse but whole genome haplotypes. By combining these with sequencing data, we can potentially "fill in the gaps". The challenge will be resolving errors in the sequencing data and the experimental techniques. The fosmid approach of Kitzman *et al.* is a much cheaper alternative to experimental phasing. However, overlapping fosmids need to be combined in order to get whole genome haplotypes. We can combine fosmids using computational techniques described in section 4 or there may be other heuristic methods that work more efficiently with fosmids as they are longer and potentially more sparse than reads. Furthermore, since fosmids are not sequenced directly, but sequenced using next-generation sequencing platforms, the short reads need to be combined computationally to reconstruct the fosmid. Using efficient computational methods to lower the cost of experimental phasing will be the next step in phasing humans on a genome-wide scale.

In this paper we discussed the haplotype phasing problem in humans which are diploid organisms and have two haplotypes that need to be resolved. However, phasing polyploid organisms may be of interest. For example, most viruses such as influenza and HIV are

haploid (one haplotype); however, viruses replicate quickly and have a high mutation rate which means that there could be several versions of a virus (quasi-species) infecting an individual at once. When we sequence a virus sample, it can contain several quasi-species. Finding the haplotypes of each quasi-species is like phasing an organism with several haplotypes. Knowing the different quasi-species infecting an individual will help in diagnosis and treatment, and understanding virus evolution. There has been some work in finding the quasi-species of a virus using sequencing data [33, 34] however, the computational problem still remains relatively unexplored.

References

[1] J. C. Venter *et al.* The sequence of the human genome. *Science*, 291:1304–1351, Feb 2001.

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

[3] M. W. Nachman and S. L. Crowell. Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156:297–304, Sep 2000.

[4] S. Levy, G. Sutton, P. C. Ng, L. Feuk, A. L. Halpern, B. P. Walenz, N. Axelrod, J. Huang, E. F. Kirkness, G. Denisov, Y. Lin, J. R. MacDonald, A. W. Pang, M. Shago, T. B. Stockwell, A. Tsiamouri, V. Bafna, V. Bansal, S. A. Kravitz, D. A. Busam, K. Y. Beeson, T. C. McIntosh, K. A. Remington, J. F. Abril, J. Gill, J. Borman, Y. H. Rogers, M. E. Frazier, S. W. Scherer, R. L. Strausberg, and J. C. Venter. The diploid genome sequence of an individual human. *PLoS Biol.*, 5:e254, Sep 2007.

[5] T. Shiina, K. Hosomichi, H. Inoko, and J. K. Kulski. The HLA genomic loci map: expression, interaction, diversity and disease. *J. Hum. Genet.*, 54:15–39, Jan 2009.

[6] E. K. Suk, G. K. McEwen, J. Duitama, K. Nowick, S. Schulz, S. Palczewski, S. Schreiber, D. T. Holloway, S. McLaughlin, H. Peckham, C. Lee, T. Huebsch, and M. R. Hoehe. A comprehensively molecular haplotype-resolved genome of a European individual. *Genome Res.*, 21:1672–1685, Oct 2011.

[7] Affymetrix. Affymetrix Genome-Wide Human SNP Array 6.0 data sheet. 2007.

[8] Illumina. Hi-Seq 2000 Sequencing System. 2010.

[9] The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073, Oct 2010.

[10] D. Gusfield. Inference of haplotypes from samples of diploid populations: complexity and algorithms. *J. Comput. Biol.*, 8:305–323, 2001.

[11] A. G. Clark. Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol. Biol. Evol.*, 7:111–122, Mar 1990.

[12] C. M. Drysdale, D. W. McGraw, C. B. Stack, J. C. Stephens, R. S. Judson, K. Nandabalan, K. Arnold, G. Ruano, and S. B. Liggett. Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. *Proc. Natl. Acad. Sci. U.S.A.*, 97:10483–10488, Sep 2000.

[13] N. Rosenberg, M. Murata, Y. Ikeda, O. Opare-Sem, A. Zivelin, E. Geffen, and U. Seligsohn. The frequent 5,10-methylenetetrahydrofolate reductase C677T polymorphism is associated with a common haplotype in whites, Japanese, and Africans. *Am. J. Hum. Genet.*, 70:758–762, Mar 2002.

[14] N. Li and M. Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165:2213–2233, Dec 2003.

[15] M. Stephens, N. J. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.*, 68:978–989, Apr 2001.

[16] M. Stephens and P. Scheet. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am. J. Hum. Genet.*, 76:449–462, Mar 2005.

[17] A. Auton, K. Bryc, A. R. Boyko, K. E. Lohmueller, J. Novembre, A. Reynolds, A. Indap, M. H. Wright, J. D. Degenhardt, R. N. Gutenkunst, K. S. King, M. R. Nelson, and C. D. Bustamante. Global distribution of genomic diversity underscores rich complex history of continental human populations. *Genome Res.*, 19:795–803, May 2009.

[18] D. A. Tregouet, I. R. Konig, J. Erdmann, A. Munteanu, P. S. Braund, A. S. Hall, A. Grosshennig, P. Linsel-Nitschke, C. Perret, M. DeSuremain, T. Meitinger, B. J. Wright, M. Preuss, A. J. Balmforth, S. G. Ball, C. Meisinger, C. Germain, A. Evans, D. Arveiler, G. Luc, J. B. Ruidavets, C. Morrison, P. van der Harst, S. Schreiber, K. Neureuther, A. Schafer, P. Bugert, N. E. El Mokhtari, J. Schrezenmeir, K. Stark, D. Rubin, H. E. Wichmann, C. Hengstenberg, W. Ouwehand, A. Ziegler, L. Tiret, J. R. Thompson, F. Cambien, H. Schunkert, and N. J. Samani. Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. *Nat. Genet.*, 41:283–285, Mar 2009.

[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. Altshuler. The structure of haplotype blocks in the human genome. *Science*, 296:2225–2229, Jun 2002.

[20] Y. Li, C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.*, 34:816–834, Dec 2010.

[21] B. N. Howie, P. Donnelly, and J. Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.*, 5:e1000529, Jun 2009.

[22] B. L. Browning and S. R. Browning. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.*, 84:210–223, Feb 2009.

[23] J. C. Roach, G. Glusman, A. F. Smit, C. D. Huff, R. Hubley, P. T. Shannon, L. Rowen, K. P. Pant, N. Goodman, M. Bamshad, J. Shendure, R. Drmanac, L. B. Jorde, L. Hood, and D. J. Galas. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, 328:636–639, Apr 2010.

[24] R. Cilibrasi, L. van Iersel, K. Steven, and J. Tromp. On the complexity of several haplotyping problems. *Proceedings of the 8th International Workshop on Algorithms in Bioinformations. Lecture Notes in Computer Science*, 3692:128–139, 2005.

[25] A. Panconesi and M. Sorzio. Fast Hare: a fast heuristic for single individual SNP haplotype reconstruction. *Science*, 328:636–639, Apr 2010.

[26] V. Bansal, A. L. Halpern, N. Axelrod, and V. Bafna. An MCMC algorithm for haplotype assembly from whole-genome sequence data. *Genome Res.*, 18:1336–1346, Aug 2008.

[27] V. Bansal and V. Bafna. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics*, 24:i153–159, Aug 2008.

[28] C. Lo, A. Bashir, V. Bansal, and V. Bafna. Strobe sequence design for haplotype assembly. *BMC Bioinformatics*, 12 Suppl 1:S24, 2011.

[29] H. C. Fan, J. Wang, A. Potanina, and S. R. Quake. Whole-genome molecular haplotyping of single cells. *Nat. Biotechnol.*, 29:51–57, Jan 2011.

[30] K. Zhang. Haplotype phasing with single genome amplification. *Unpublished*.

[31] J. O. Kitzman, A. P. Mackenzie, A. Adey, J. B. Hiatt, R. P. Patwardhan, P. H. Sudmant, S. B. Ng, C. Alkan, R. Qiu, E. E. Eichler, and J. Shendure. Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat. Biotechnol.*, 29:59–63, Jan 2011.

[32] J. Marchini, D. Cutler, N. Patterson, M. Stephens, E. Eskin, E. Halperin, S. Lin, Z. S. Qin, H. M. Munro, G. R. Abecasis, and P. Donnelly. A comparison of phasing algorithms for trios and unrelated individuals. *Am. J. Hum. Genet.*, 78:437–450, Mar 2006.

[33] N. Eriksson, L. Pachter, Y. Mitsuya, S. Y. Rhee, C. Wang, B. Gharizadeh, M. Ronaghi, R. W. Shafer, and N. Beerenwinkel. Viral population estimation using pyrosequencing. *PLoS Comput. Biol.*, 4:e1000074, Apr 2008.

[34] I. Astrovskaya, B. Tork, S. Mangul, K. Westbrooks, I. M?ndoiu, P. Balfe, and A. Zelikovsky. Inferring viral quasispecies spectra from 454 pyrosequencing reads. *BMC Bioinformatics*, 12 Suppl 6:S1, 2011.