The Clark Phase-able Sample Size Problem: Long-range Phasing and Loss of Heterozygosity in GWAS*

Bjarni V. Halldórsson[1,3,4,†], Derek Aguiar[1,2,†,‡], Ryan Tarpine[1,2,†,‡], and Sorin Istrail[1,2,‡]

[1] Center for Computational Molecular Biology, Brown University
[2] Department of Computer Science, Brown University
[3] School of Science and Engineering, Reykjavik University
[4] deCODE genetics

Abstract. A phase transition is taking place today. The amount of data generated by genome resequencing technologies is so large that in some cases it is now less expensive to repeat the experiment than to store the information generated by the experiment. In the next few years it is quite possible that millions of Americans will have been genotyped. The question then arises of how to make the best use of this information and jointly estimate the haplotypes of all these individuals. The premise of the paper is that long shared genomic regions (or tracts) are unlikely unless the haplotypes are identical by descent (IBD), in contrast to short shared tracts which may be identical by state (IBS). Here we estimate for populations, using the US as a model, what sample size of genotyped individuals would be necessary to have sufficiently long shared haplotype regions (tracts) that are identical by descent (IBD), at a statistically significant level. These tracts can then be used as input for a Clark-like phasing method to obtain a complete phasing solution of the sample. We estimate in this paper that for a population like the US and about 1% of the people genotyped (approximately 2 million), tracts of about 200 SNPs long are shared between pairs of individuals IBD with high probability which assures the Clark method phasing success. We show on simulated data that the algorithm will get an almost perfect solution if the number of individuals being SNP arrayed is large enough and the correctness of the algorithm grows with the number of individuals being genotyped.

We also study a related problem that connects copy number variation with phasing algorithm success. A loss of heterozygosity (LOH) event is when, by the laws of Mendelian inheritance, an individual should be heterozygote but, due to a deletion polymorphism, is not. Such polymorphisms are difficult to detect using existing algorithms, but play an important role in the genetics of disease and will confuse haplotype phasing

* corresponding authors are Bjarni V. Halldórsson bjarnivh@ru.is and Sorin Istrail sorin@cs.brown.edu
† contributed equally to this work
‡ member of the International Multiple Sclerosis Genetics Consortium GWAS Analysis team

algorithms if not accounted for. We will present an algorithm for detecting LOH regions across the genomes of thousands of individuals. The design of the long-range phasing algorithm and the Loss of Heterozygosity inference algorithms was inspired by analyzing of the Multiple Sclerosis (MS) GWAS dataset of the International Multiple Sclerosis Consortium and we present in this paper similar results with those obtained from the MS data.

1 Introduction

Genome-wide association studies (GWAS) proceed by identifying a number of individuals carrying a disease or trait and comparing these individuals to those that do not or are not known to carry the disease/trait. Both sets of individuals are then genotyped for a large number of Single Nucleotide Polymorphism (SNP) genetic variants which are then tested for association to the disease/trait. GWAS have been able to successfully identify a very large number of polymorphism associated to disease ([19, 4, 1] etc.) and the amount of SNP data from these studies is growing rapidly. Studies using tens of thousands of individuals are becoming commonplace and are increasingly the norm in the association of genetic variants to disease [5, 19, 13]. These studies generally proceed by pooling together large amounts of genome-wide data from multiple studies, for a combined total of tens of thousands of individuals in a single meta-analysis study. It can be expected that if the number of individuals being genotyped continues to grow, hundreds of thousands, if not millions, of individuals will soon be studied for association to a single disease or trait.

SNPs are the most abundant form of variation between two individuals. However, other forms of variation exist such as copy number variation – large scale chromosomal deletions, insertions, and duplications (CNV). These variations, which have shown to be increasingly important and an influential factor in many diseases [17], are not probed using SNP arrays. A further limitation of SNP arrays is that they are designed to probe only previously discovered, common variants. Rare variants, belonging perhaps only to a small set of carriers of a particular disease and hence potentially more deleterious, will not be detected using SNP arrays.

To reach their full potential, the future direction of genetic association studies are mainly twofold: the testing of more individuals using genome-wide association arrays and the resequencing of a small number of individuals with the goal of detecting more types of genetic variations, both rare SNPs and structural variation [16]. Testing multiple individuals for the same variants using standard genome-wide association arrays is becoming increasingly common and can be done at a cost of approximately $100 per individual. In the next couple of years it is plausible that several million individuals in the US population will have had their genome SNP arrayed. In contrast, whole genome resequencing is currently in its infancy. A few people have had their genome resequenced and the cost of sequencing a single individual is still estimated in the hundreds of thousands of

dollars. However, whole genome sequencing is preferable for association studies as it allows for the detection of all genomic variation and not only SNP variation.

Due to the fact whole genome SNP arrays are becoming increasingly abundant and whole genome resequencing is still quite expensive, the question has been raised whether it would suffice to whole genome sequence a small number of individuals and then impute [7] other genotypes using SNP arrays and the shared inheritance of these two sets of individuals. It has been shown – in the Icelandic population with a rich pedigree structure known – that this could be done most efficiently using the haplotypes shared by descent between the individuals that are SNP arrayed and those that have been resequenced [10]. Haplotype sharing by descent occurs most frequently between closely related individuals, but also occurs with low probability between individuals that are more distantly related. In small closely related populations, as in the Icelandic population, only a moderately sized sample size is therefore needed in order for each individual to have, with high probability, an individual that is closely related to it. In larger populations, such as the US population, a larger sample size will be needed for there to be a significant probability of an individual sharing a haplotype by descent within the population. We say that an individual is "Clark phaseable" with respect to a population sample if the sample contains an individual that shares a haplotype with this individual by descent. In this paper we explore what the required sample size is so that most individuals within the sample are Clark phaseable, when the sample is drawn from a large heterogeneous population, such as the US population.

Problem 1. Current technologies, suitable for large-scale polymorphism screening, only yield the genotype information at each SNP site. The actual haplotypes in the typed region can only be obtained at a considerably high experimental cost or computationally by haplotype phasing. Due to the importance of haplotype information for inferring population history and for disease association, the development of algorithms for detecting haplotypes from genotype data has been an active research area for several years [3, 15, 18, 14, 10, 6]. However, algorithms for determining haplotype phase are still in their infancy after about 15 years of development (e.g. [3, 18, 9]). Of particular worry is the fact that the learning rate of the algorithm, i.e. the rate that the algorithms are able to infer more correct haplotypes, grows quite slowly with the number of individuals being SNP arrayed.

Solution 1. In this paper we present an algorithm for the phasing of a large number of individuals. We show that the algorithm will get an almost perfect solution if the number of individuals being SNP arrayed is large enough and the correctness of the algorithm grows with the number of individuals being genotyped. We will consider the problem of haplotype phasing from long shared genomic regions (that we call tracts). Long shared tracts are unlikely unless the haplotypes are identical by descent (IBD), in contrast to short shared tracts which may be identical by state (IBS). We show how we can use these long shared tracts for haplotype phasing.

Problem 2. We further consider the problem of detecting copy number variations from whole genome SNP arrays. A loss of heterozygosity (LOH) event is when, by the laws of Mendelian inheritance, an individual should be heterozygote but due to a deletion polymorphism, is not. Such polymorphisms are difficult to detect using existing algorithms, but play an important role in the genetics of disease [17] and will confuse haplotype phasing algorithms if not accounted for.

Solution 2. We provide an exact exponential algorithm and a greedy heuristic for detecting LOH regions.

For this paper, we run empirical tests and benchmark the algorithms on a simulated GWAS datasets [8] resembling the structure of the International Multiple Sclerosis Genetics Consortium [4] data. To determine LOH events we assume the data is given in trios, i.e. the genotypes of a child and both its parents are known.

2 Long Range Phasing and Haplotype Tracts

The haplotype phasing problem asks to computationally determine the set of haplotypes given a set of individual's genotypes. We define a *haplotype tract* (or *tract* for short) denoted $[i, j]$ as a sequence of SNPs that is shared between at least two individuals starting at the same position i in all individuals and ending at the same position j in all individuals. We show that if we have a long enough tract then the probability that the sharing is IBD is close to 1. Multiple sharing of long tracts further increases the probability that the sharing corresponds to the true phasing.

2.1 Probability of Observing a Long Tract

We show that as the length of the tract increases the probability that the tract is shared IBD increases. Let t be some shared tract between two individual's haplotypes and l be the length of that shared tract. We can then approximate the probability this shared tract is identical by state (IBS) $p_{IBS}(l)$. Let $f_{M,i}$ be the major allele frequency of the SNP in position i in the shared tract t. Assuming the Infinite Sites model and each locus is independent,

$$p_{IBS}(l) = \prod_{i=1}^{l} \left((f_{M,i})(f_{M,i}) + (1 - f_{M,i})\,(1 - f_{M,i}) \right)$$

We can approximate $p_{IBS}(l)$ by noticing $f_{M,i} * f_{M,i}$ dominates $(1-f_{M,i})(1-f_{M,i})$ as $f_{M,i} \to 1$, $p_{IBS}(l) \approx \prod_{i=1}^{l} (f_{M,i})^2$. Let f_{avg} be $\frac{1}{l} f_{M,i}$ $\forall i \in t$. Then $p_{IBS}(l) \approx (f_{avg})^{2l}$. Given $f_{M,i}$ is some high frequency, say 95%, then a sharing of 100 consecutive alleles is very unlikely, $p_{IBS}(100) \approx 0.95^{200} = 10^{-5}$. For very large datasets we will need to select the length of the tract being considered to be large enough so that the probability that the sharing is identical by state is small.

The probability two individuals separated by $2(k+1)$ meiosis (kth-degree cousins) share a locus IBD is 2^{-2k} [10]. As k increases, the probability kth-degree cousins share a particular locus IBD decreases exponentially. However, if two individuals share a locus IBD then they are expected to share about $\frac{200}{2k+2}$ cM [10]. Relating $P(IBD)$ to length of tract l,

$$P(IBD|\textit{sharing of length } l) = \frac{2^{-2n}}{2^{-2n} + \left((f_{M,i})^{2l} + (1-f_{M,i})^{2l}\right)}$$

which is shown in Fig. 1.

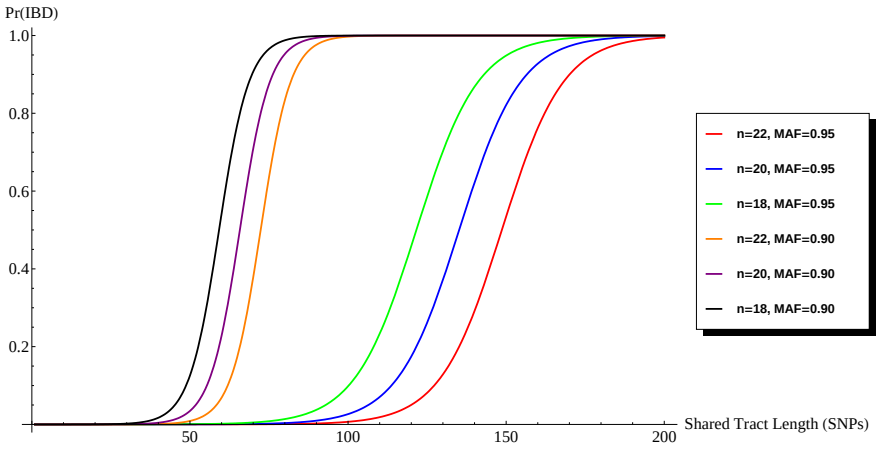


Fig. 1. Probability of IBD as a function of shared tract length (measured in SNPs) and plotted for several n and major allele frequencies (MAF). n is the number of meiosis between the two individuals. The smaller the MAF or n the faster P(IBD) converges to 1.

2.2 The Clark Phase-able Sample Size Problem

Given the large tract sharing, we can construct the *Clark consistency graph* having individuals as vertices and an edge between two individuals if they share a tract [15]. Figure 2 shows the Clark consistency graph for different *minimum significant tract lengths* (or window sizes) in the MS dataset. At what minimum significant tract lengths will the graph become dense enough so that phasing can be done properly? What percentage of the population needs to be genotyped so that the Clark consistency graph becomes essentially a single connected component? We call this "The Clark sample estimate: the size for which the Clark consistency graph is connected, C."

We computed the average number of edges in the haplotype consistency graph as a function of window size to get a sense when the Clark consistency graph of the MS data becomes connected. Based on Fig. 3 and $P(IBD)$ we can propose an algorithmic problem formulation from the Clark consistency graph. Preferably we would like to solve either one of the below problems.

Problem 3. Remove the minimum number of the edges from the Clark consistency graph so that the resulting graph gives a consistent phasing of the haplotypes.

Problem 4. Maximize the joint probability of all the haplotypes given the observed haplotype sharing.

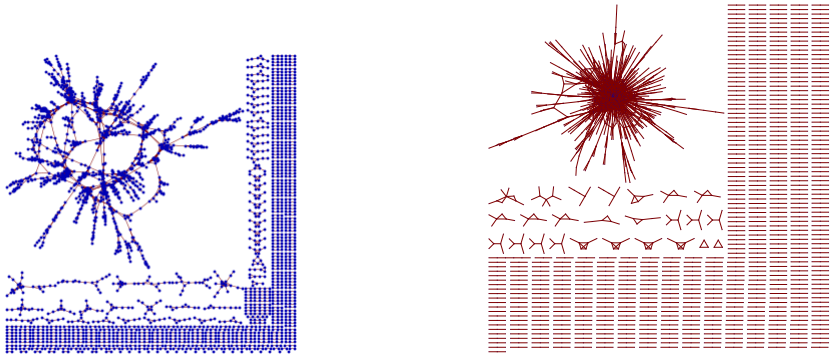


Fig. 2. Left: The Clark consistency graph for region [1400,1600). A large fraction of individuals share consistent haplotypes of length 200 suggesting many are IBD. Right: The Clark consistency graph for a smaller window size of 180 base pairs. We observe a more dense connected component in part due to the smaller windows size but also because of the specific genomic region.

We believe that both of these problem formulations are NP-hard and instead propose to solve these problems using a heuristic. Our benchmarking on simulated data shows that this heuristic works quite well.

2.3 Phasing the Individuals That Are Part of the Largest Component

We now proceed with an iterative algorithm working on the connected components in the Clark haplotype consistency graph. First we construct the graph according to some length of haplotype consistency (Fig. 3 and $P(IBD)$ help define this length). We iterate through each site of each individual to find the

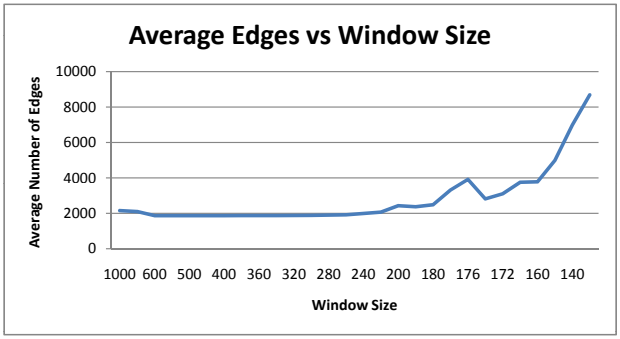


Fig. 3. The average number of edges per window size stays relatively constant until a window size of about 180. The graph becomes more connected at this point likely because the window size is small enough to not be largely affected by recombination (but still large enough for the shared tracts to not likely be IBS).

tracts. After we find a site with some long shared region, we look at its neighbors in the connected component and apply a voting scheme to decide what the value is for each heterozygous allele. After each individual has been processed we iterate with having resolved sites in the original matrix.

Observation 1. *If the Clark consistency graph is fully connected all edges are due to IBD sharing and all individuals can be perfectly phased up to the point were all individuals are heterozygote at a particular site.*

Therefore, phasing individuals in a connected component of the graph should be easy, but in practice there will be some inconsistencies for a number of reasons. If a node in the Clark consistency graph has a high degree then the phasing of that node will be ambiguous if its neighbors are not consistent. At some times this may be due to genotyping error and at times this may be due to identical by state sharing to either one or both of an individuals haplotypes. The identical by state sharing may because the haplotype has undergone recombination, possibly a part of the haplotype is shared identical by descent and a part is identical by state.

Our alphabet for genotype data is $\Sigma = \{0, 1, 2, 3\}$. 0s and 1s represent the homozygote for the two alleles of a SNP. A 2 represents a heterozygous site and a 3 represents missing data. Given a set of n-long genotype strings $G = \{g_1, g_2, \ldots, g_{|G|}\}$ where $g_i \in \Sigma^n$, we represent this in a matrix M with $m = 2\,|G|$ rows and n columns:

$$M = \begin{bmatrix} M_{1,1} & M_{1,2} & \cdots & M_{1,n} \\ M_{2,1} & M_{2,2} & \cdots & M_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ M_{m,1} & M_{m,2} & \cdots & M_{m,n} \end{bmatrix}$$

Each genotype g_i is represented by the two rows $2i-1$ and $2i$. Initially, $M_{2i-1,j} = M_{2i,j} = g_i[j]$.

We define allele consistency to be:

$$c(a,\, b) = \begin{cases} 1 & \text{if } a = b \text{ or } a \in \{2,\, 3\} \text{ or } b \in \{2,\, 3\} \\ 0 & otherwise \end{cases}$$

Rows r and s of M are consistent along a tract $[i,\, j]$ (i.e. have a shared tract) is written

$$C_{[i,\, j]}(r,\, s) = \prod_{k \in [i,\, j]} c\,(M_{r,k},\, M_{s,k})$$

The length of a tract is written $|[i,\, j]| = j - i + 1$.

A shared tract $[i,\, j]$ between rows r and s is *maximal shared tract* if it cannot be extended to the left or right; i.e., $i = 1$ or $c(M_{r,i-1},\, M_{s,i-1}) = 0$ and $j = n$ or $c(M_{r,j+1},\, M_{s,j+1}) = 0$. The maximal shared tract between rows r and s at position i is written $S_i^{r,s}$. It is unique. Note that if $S_i^{r,s} = [j,\, k]$ then $\forall_{l \in [j,\, k]} S_l^{r,s} = S_i^{r,s}$.

2.4 Tract Finding and Phasing Algorithm

Given that there are some loci for which individuals share IBD and that these sharings are expected to be large, we developed an algorithm to detect and use these sharings to resolve the phase at heterozygous sites. Each site is resolved by determining if there are any other individuals that likely share a haplotype by descent. SNPs that do not have their phase determined during any given iteration will be processed in succeeding iterations. If there are enough long IBD loci, this algorithm should unambiguously determine the phase of each individual.

If we know that the data contains trios, a child and both of its parents, we start by phasing the trios using Mendelian laws of inheritance. This replaces many of the heterozygote sites (whenever at least one member of a family is homozygous) and even a few of the sites having missing data (i.e., when the parents are both homozygous and the child's genotype is missing).

To phase using long shared tracts, we start by fixing a minimum significant tract length L. We run several iterations, each of which generate a modified matrix M' from M, which is then used as the basis for the next iteration.

First, we set $M' := M$.

For each row r we examine position i. If $M_{r,i} \in \{0,\, 1\}$ then we move to the next i. Otherwise $M_{r,i} \in \{2,\, 3\}$, and we count "votes" for whether the actual allele is a 0 or 1.

$$V_0^r = |\{s \,|\, s \neq r \text{ and } |S_i^{r,s}| \geq L \text{ and } M_{s,i} = 0\}|$$

V_1^r is defined analogously (the difference being the condition $M_{s,i} = 1$). If $V_0^r > V_1^r$ then we set $M'_{r,i} := 0$. Similarly for $V_1^r > V_0^r$. If $V_0^r = V_1^r$ then we do nothing.

A more complex case is when $M_{r,i} = 2$. We make sure the complementary haplotypes are given different alleles by setting the values of both haplotypes

simultaneously. This does not cause a dependency on which haplotype is visited first because we have extra information we can take advantage of. We count votes for the complementary haplotype and treat them oppositely. That is, votes for the complementary haplotype having a 1 can be treated as votes for the current haplotype having a 0 (and vice versa). So letting r' be the row index for the complementary haplotype, we actually compare $V_0^r + V_1^{r'}$ and $V_1^r + V_0^{r'}$. This is helpful when SNPs near position i (which therefore will fall within shared tracts involving i) have already been phased (by trio pre-phasing or previous iterations). It also helps in making the best decision when both haplotypes receive a majority of votes for the same allele, e.g., both have a majority of votes for 0. In this case, taking into account votes for the two haplotypes simultaneously will result in whichever has *more* votes getting assigned the actual value 0. If they each receive the exact same number of votes, then no allele will be assigned. This also avoids the above-mentioned dependency on the order in which the haplotypes are visited – the outcome is the same since votes for both are taken into account.

In this manner, M' is calculated at each position. If $M' = M$ (i.e. no changes were made) then the algorithm terminates. Otherwise, $M := M'$ (M is replaced by M') and another iteration is run.

2.5 Phasing the Individuals That Are Not a Part of the Largest Component

Individuals that are part of small connected components will have a number of ambiguous sites once they have been phased using the edges in their connected component. For these individuals, we compute a minimum number of recombinations and mutations from their haplotypes to others that have better phasing (belong to larger components). We then assign these haplotypes phase based on minimizing the number of mutations plus recombinations in a similar manner as the approach of Minichiello Durbin [12].

Alternatively this could be done in a sampling framework, where we sample the haplotype with a probability that is a function of the number of mutations and recombinations.

2.6 Experimental Results on Simulated Data

We compared the correctness and learning rate of our algorithm against BEAGLE [2] using a simulated dataset. Using the Hudson Simulator [8], we generated 3000 haplotypes each consisting of 3434 SNPs from chromosomes of length 10^5. We estimated a population size of 10^6 with a neutral mutation rate of 10^{-9}. To generate genotypes, we randomly sampled from the distribution of simulated haplotypes with replacement such that each haplotype was sampled on average 2, 3, and 4 times. We applied our algorithm and BEAGLE to the simulated data after combining haplotypes to create parent-offspring trio data (inspired by our analysis of the MS dataset). Both algorithms effectively phase the simulated dataset largely due to the initial trio phasing (Table 1). Our algorithm learns

the true phasing at an increasing rate as the expectation of haplotypes sampled increases. The most clear example of this trend is in the Brown Long Range Phasing miscall rate. By weighing edges proportional to probability of sharing IBD rather than a fixed set of votes per edge, we should achieve more accurate phasings (subject of future work).

Table 1. We created three populations using a base pool of 3000 simulated haplotypes using the Hudson simulator. Populations 1, 2, and 3 were created by sampling each haplotype according to a geometric distribution with expectation 2, 3, and 4 respectively. Haplotypes were then randomly paired to create genotypes. The miscall rate is the ratio of 2's miscalled to total 2's (after trio phasing). Error-free phasings denote the number of haplotype phasings with zero miscalled 2's.

	Population 1	Population 2	Population 3
BEAGLE miscall rate	0.0685%	0.0160%	0.00951%
Brown Long Range Phasing miscall rate	0.0501%	0.0148%	0.00503%
BEAGLE Error-free phasings	4467	6819	8898
Brown Long Range Phasing Error-free phasings	4459	6840	8923
Total haplotypes	4524	6870	8940

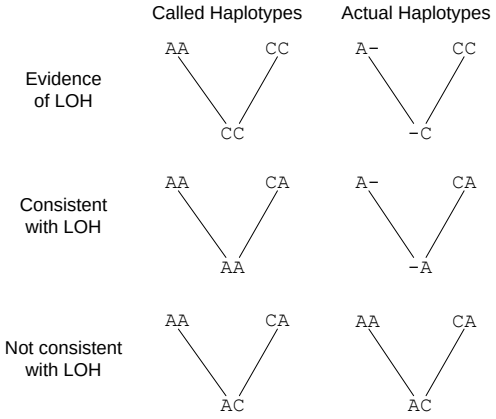
3 Loss of Heterozygosity Regions

We call the loss of the normal allele a Loss of Heterozygosity (LOH) which may be a genetic determinant in the development of disease [11, 17]. In some situations, individuals that are heterozygous at a particular locus can possess one normal allele and one deleterious allele. The detection of CNVs, such as deletions, is an important aspect of GWAS to find LOH events, and yet, it is commonly overlooked due to technological and computational limitations.

LOH can be inferred using data from SNP arrays. The SNP calling algorithm for SNP arrays cannot distinguish between an individual who is homozygous for some allele a and an individual who has a deletion haplotype and the allele a (Fig. 4, Left). LOH events can then be inferred by finding such genotypic events throughout the dataset. We will present two algorithms for computing putative LOH regions across GWAS datasets.

3.1 Definitions

A *trio* consists of three individual's genotypes and is defined by the inheritance pattern of parents to child. As before, let M denote the matrix of genotypes but we now assume M consists of trios. Let M_i denote the i^{th} trio of M (individuals i, $i+1$, and $i+2$). At any site j the trio M_i may have 4^3 possible genotype



Father Genotype

Mother Genotype	0	1	2	3	Child Genotype
0	C^P	E^F	C^M	C^P	0
	N	E^M	E^M	E^M	1
	N	N	N	N	2
	C^P	C^P	C^M	C^P	3
1	E^M	N	E^M	E^M	0
	E^F	C^P	C^M	C^P	1
	N	N	N	N	2
	C^P	C^P	C^M	C^P	3
2	C^F	E^F	N	C^F	0
	E^F	C^F	N	C^F	1
	N	N	N	N	2
	C^F	C^F	N	C^F	3
3	C^P	E^F	C^M	C^P	0
	E^F	C^P	C^M	C^P	1
	N	N	N	N	2
	C^P	C^P	C^M	C^P	3

Fig. 4. Left: Three examples of inheritance patterns in GWAS data in the context of LOH. The Evidence of LOH (ELOH) pattern shows strong correlation between LOH and a SNP site because the only possible explanation involves introducing a deletion haplotype. An inheritance pattern is called consistent with LOH (CLOH) if it does not contradict the presence of a deletion haplotype and can be explained with normal inheritance patterns. An inheritance pattern not consistent with LOH (NCLOH) occurs when a deletion haplotype cannot be introduced to explain the trio inheritance pattern. Right: The correlation between inheritance pattern and ELOH, CLOH, and NCLOH. We define E to be ELOH, C to be CLOH, and N to be NCLOH. The superscript defines for which parent the putative deletion haplotype is associated. We define the superscript F to be consistent with a deletion haplotype inherited from the father, M for mother, and P for both parents.

combinations for which the trio can either be *consistent with LOH* (CLOH), *not consistent with LOH* (NCLOH), or show *evidence of LOH* (ELOH) (Fig. 4, Left). A trio at site i shows ELOH if the inheritance pattern can only be explained with the use of a deletion haplotype (or a genotyping error). A trio at site i is NCLOH if the inheritance pattern cannot be explained with the use of a deletion haplotype, and CLOH if it *may* be explained with the use of a deletion haplotype.

3.2 The LOH Inference Problem

We are given a set of n SNPs and a set of m trios genotyped at those SNPs. For each SNP/trio pair the SNP can have one of three labels:

– X – The marker is inconsistent with having a loss of heterozygosity (Fig. 4, Left: Not Consistent with LOH).
– 0 – The marker is consistent with having a loss of heterozygosity (Fig. 4, Left: Consistent with LOH).
– 1 – The SNP shows evidence of loss of heterozygosity, (Fig. 4, Left: Evidence of LOH).

For any trio M_i, a contiguous sequence of at least one 1 and an unbounded number of 0 sites is called an *putative deletion*. We call two putative deletions, p_i and p_j, overlapping if they share at least 1 common index. Let h_i and h_j be two ELOH and let p_i and p_j contain h_i and h_j respectively. Each putative deletion is associated with an interval which is defined by their start and end indices: $[s_i, e_i]$ and $[s_j, e_j]$ respectively. h_i and h_j are called compatible (or overlapping) if h_i and h_j are members of the same putative deletion (i.e. $h_i \in [s_i, e_i]$ and $h_j \in [s_i, e_i]$) or h_i is contained in the interval defining p_j and h_j is contained in the interval defining p_i. All CLOH and ELOH sites within a putative deletion must share the same parent (Fig. 4, Right). The task is to call all 1's $\in M$ either a deletion or a genotyping error according to some objective function which weighs the relative costs of calling genotyping errors or deletions.

3.3 LOH Inference Algorithms

We present an exponential algorithm and a greedy heuristic for computing putative deletions. Both algorithms begin by parsing M and removing SNPs in which the Mendelian error rate is above 5% to remove artifacts from genotyping. We then calculate the LOH site vector for each trio in the dataset which corresponds to using the table defined in Fig. 4 (Right) to translate each SNP site. This new matrix is denoted $N^{\left(\frac{|M|}{3} \times l\right)}$. To identify the genotyping errors and putative deletions, we define two operations on N: error correction call and deletion haplotype call. An error correction call will categorize an ELOH site as a genotyping error effectively removing it from any particular deletion haplotype. An deletion haplotype call will identify a putative deletion as an inherited deletion haplotype. We infer inherited deletion haplotypes using the objective function

$$min_N \left(k_1 * (\text{genotype error corrections calls}) + k_2 * (\text{deletion haplotypes calls})\right)$$

where k_1 and k_2 are weighing factors. k_1 and k_2 can be simple constant factors or a more complex distribution. For example, setting k_1 to 2 and k_2 to 7, we will prefer calling a putative deletion with at least 4 pairwise compatible ELOH sites an inherited deletion. For a more complex objective function, we could define k_2 to be k_3(number of conserved individuals) + k_4(length of overlapping region) + k_5((number of ELOH)/(number of CLOH)). The parameters must be tuned to the input data. For example, association tests will tune the parameter to favor putative deletions with many conserved individuals. We suspect that this problem is NP-complete for general N. In the case of the Multiple Sclerosis dataset,

the matrix N contains small overlapping putative deletions and over 95% of N is non-putative deletions, that is, N is very sparse.

Algorithm 1. We start by giving an exact exponential algorithm which minimizes the objective function. Let x_i denote a set of overlapping putative deletions. For sparse N we can reduce the minimization function from min_N to $min_{x_1..x_s}$ where $x_1..x_s \in N$ and $\{x_1..x_s\} \subseteq N$. Since any particular putative deletion is defined by the ELOH sites, we can enumerate all feasible non-empty sets of ELOH sites for all x_i. Computing this for all putative deletions demands work proportional to $\sum_{i=1}^{s} B(e_i)$ where e_i is the number of ELOH sites in x_i and B is the Bell number. In practice, we found that e_i is bounded by a small constant but this complexity is still unreasonable for most e_i.

Algorithm 2. For practical purposes, we've developed a greedy algorithm for cases where the exact exponential algorithm in unreasonable (Fig. 5). For each $x_i \in N$, the algorithm selects the component with the maximum *trio sharing*, that is, the possibly overlapping putative deletions that include the most ELOH sites. Because every two ELOH sites in an inherited deletion must be pairwise compatible, this component is a clique. To find the maximum clique, we construct an overlap graph $G(V, E)$ where $h_i \in V$ if h_i is an ELOH in a putative deletion in this interval and $(h_i, h_j) \in E$ if h_i and h_j are compatible. Identifying the maximum clique in this graph is NP complete. We therefore find maximum cliques using a greedy approach that iterates over a queue containing the compatible vertices, selecting the highest degree node v_m and adding it to the potential clique set if and only there is an edge between v_m and each vertex in the clique. At the end of this process, the algorithm calls the site(s) a deletion haplotype or genotyping error according to the objective function, clears the set, and continues until all vertices in the queue are processed.

3.4 Experimental Results on Simulated Data

We tested the algorithm using the same simulated phasing dataset. To simulate and score an error-prone GWAS dataset containing an LOH, we define six parameters, two metrics, and generate only one deletion in the genotype matrix (Table 2). We randomly select a set of trios and an interval in the simulated haplotype matrix to contain the generated deletion. After the site is selected, we place ELOH sites on the SNPs according to some probability (assumed independent for each SNP in the interval).

Although our LOH model is quite simplistic, we do observe promising results. Our algorithm is sensitive to inherited deletions that are very short but shared among many people and also sensitive to inherited deletions that are longer and shared by few people.

In general, the algorithm is accurate when the coefficient of deletion call and genotype error call are tuned well (Table 3 – parameter sets 1-4). For a dataset with low genotyping error rate ($\sim$0.0001 site error probability), the coefficient of deletion call can be set low; if it is set too high, a true inherited deletion

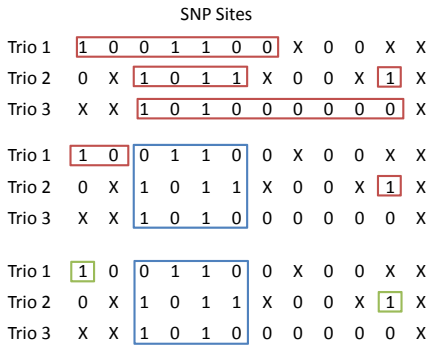


Fig. 5. A visual depiction of the greedy algorithm for finding putative deletions (consistencies with particular parents are omitted for simplicity). The red rectangles denote trio SNP sites which have not been called yet. The blue rectangle denotes a called inherited deletion haplotype. A green rectangle denotes a genotype error call. First, the algorithm finds the component (a clique in G(V,E)) with the maximum trio sharing: SNP sites 3-6. It checks if the score of this component and either calls it an inherited deletion or a set of genotyping errors (in this case the former). The intervals are updated by remove vertices and edges from the overlap graph and the algorithm continues. Both remaining components consisting of SNP sites 1 and 11 are both of size 1. These will most likely be called genotyping errors.

may be incorrectly called a genotyping error, possibly missing an associative LOH (Table 3 – parameter set 5). A similar caveat pertains to datasets with significant genotyping error rates (for instance, the MS dataset). A coefficient of deletion call that is too low can yield false positives (Table 3 – parameter set 6). Finding appropriate tuning mechanisms for the two coefficients to maximize algorithm specificity and sensitivity will be the subject of future work.

4 Conclusion and Future Work

We have shown that long range phasing using Clark consistency graphs is practical for very large datasets and the accuracy of the algorithm improves rapidly with the size of the dataset. We have also given an algorithm that removes most Mendelian inconsistencies and distinguishes between genotyping errors and deletion events which can be factored into the phasing algorithm when applied to GWAS data. Future work includes applying probabilistic models to both algorithms to score tract sharings and putative deletions more appropriately.

All algorithms are available via sending a request to the corresponding authors.

Table 2. Six tunable parameters and two scoring metrics for testing of the LOH algorithm.

Probability of Error per Site	For all SNP-trio pairs, we add a Mendelian error according to this probability (assumed independent for each site).
Interval Length	The exact length of the generated deletion.
Trios in Deletion	The exact number of trios sharing the generated deletion.
Probability of ELOH in Interval	The probability a SNP is an ELOH site within the generated deletion interval.
Coefficient of Genotype Error Call	The objective function cost for calling an ELOH site a genotyping error (parameter k_1 in our objective function)
Coefficient of Inherited Deletion Call	The objective function cost for calling a set of ELOH sites an inherited deletion (parameter k_2 in our objective function)
True Positive	There is one interval that contains the inherited deletion, thus a true positive corresponds to correctly identifying an inherited deletion in this region.
False Positive	We have a false positive if we identify an inherited deletion in a region disjoint from the generated deletion's region.

Table 3. We tested out algorithm using the six tunable parameters as defined in Table 2. Each configuration was run with a coefficient of genotyping error of 2.

Param Set	Site Error Prob.	Interval Length	Trios in Deletion	Prob. of ELOH	Coeff. of Deletion	True Positive	False Positive	Runs
1	0.0001	5	5	0.75	11	1000	0	1000
2	0.0001	2	5	1	11	1000	0	1000
3	0.0001	2	5	1	11	1000	0	1000
4	0.0001	9	3	0.75	11	1000	0	1000
5	0.0001	7	3	0.50	15	58	0	100
6	0.00333	9	3	0.75	15	100	38888	100

5 Acknowledgments

Thanks to the International Multiple Sclerosis Genetics Consortium for sharing the Multiple Sclerosis GWAS dataset.

References

[1] David Altshuler, Mark J. Daly, and Eric S. Lander, *Genetic mapping in human disease.*, Science (New York, N.Y.) **322** (2008), no. 5903, 881–888.

[2] 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.*, American journal of human genetics **84** (2009), no. 2, 210–223.

[3] AG Clark, *Inference of haplotypes from PCR-amplified samples of diploid populations*, Mol Biol Evol **7** (1990), no. 2, 111–122.

[4] The International Multiple Sclerosis Genetics Consortium, *Risk alleles for multiple sclerosis identified by a genomewide study*, N Engl J Med **357** (2007), no. 9, 851–862.

[5] Daniel F. Gudbjartsson, G. Bragi Walters, Gudmar Thorleifsson, Hreinn Stefansson, Bjarni V. Halldorsson, et al., *Many sequence variants affecting diversity of adult human height*, Nat Genet **40** (2008), no. 5, 609–615.

[6] Bjarni V. Halldórsson, Vineet Bafna, Nathan Edwards, Shibu Yooseph, and Sorin Istrail, *A survey of computational methods for determining haplotypes*, (2004).

[7] Bryan N. Howie, Peter Donnelly, and Jonathan Marchini, *A flexible and accurate genotype imputation method for the next generation of genome-wide association studies*, PLoS Genet **5** (2009), no. 6, e1000529.

[8] Richard R. Hudson, *Generating samples under a wright-fisher neutral model of genetic variation*, Bioinformatics **18** (2002), no. 2, 337–338.

[9] Sorin Istrail, *The haplotype phasing problem*, Symposium in Honor of Mike Waterman's 60th Birthday, 2002.

[10] Augustine Kong, Gisli Masson, Michael L. Frigge, et al., *Detection of sharing by descent, long-range phasing and haplotype imputation*, Nat Genet **40** (2008), no. 9, 1068–1075.

[11] Steven A. McCarroll, Finny G. Kuruvilla, Joshua M. Korn, Simon Cawley, et al., *Integrated detection and population-genetic analysis of snps and copy number variation*, Nat Genet **40** (2008), no. 10, 1166–1174.

[12] Mark J. Minichiello and Richard Durbin, *Mapping trait loci by use of inferred ancestral recombination graphs*, **79** (2006), no. 5, 910–922.

[13] F. Rivadeneira, U. Styrkarsdottir, K. Estrada, B. Halldorsson, et al., *Bone*, vol. 44, ch. Twenty loci associated with bone mineral density identified by large-scale meta-analysis of genome-wide association datasets, pp. S230–S231, Elsevier Science, Jun 2009.

[14] Paul Scheet and Matthew Stephens, *A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase*, **78** (2006), no. 4, 629–644.

[15] Roded Sharan, Bjarni V. Halldórsson, and Sorin Istrail, *Islands of tractability for parsimony haplotyping*, IEEE/ACM Transactions on Computational Biology and Bioinformatics **3** (2006), no. 3, 303–311.

[16] Nayanah Siva, *1000 genomes project.*, Nature biotechnology **26** (2008), no. 3, 256.

[17] Hreinn Stefansson, Dan Rujescu, Sven Cichon, Olli P. H. Pietilainen, et al., *Large recurrent microdeletions associated with schizophrenia*, Nature **455** (2008), no. 7210, 232–236.

[18] Matthew Stephens, Nicholas J. Smith, and Peter Donnelly, *A new statistical method for haplotype reconstruction from population data*, **68** (2001), no. 4, 978–989.

[19] Unnur Styrkarsdottir, Bjarni V. Halldorsson, Solveig Gretarsdottir, Daniel F. Gudbjartsson, G. Bragi Walters, et al., *Multiple Genetic Loci for Bone Mineral Density and Fractures*, N Engl J Med **358** (2008), no. 22, 2355–2365.