JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 18, Number 3, 2011

Pp. 323–333
DOI: 10.1089/cmb.2010.0288

The Clark Phaseable Sample Size Problem: Long-Range Phasing and Loss of Heterozygosity in GWAS

*BJARNI V. HALLDÓRSSON,[1] *DEREK AGUIAR,[2,3,4]
*RYAN TARPINE,[2,3,4] and SORIN ISTRAIL[2,3,4]

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 this article is that long shared genomic regions (or tracts) are unlikely unless the haplotypes are identical by descent. These tracts can be used as input for a Clark-like phasing method to obtain a phasing solution of the sample. We show on simulated data that the algorithm will get an almost perfect solution if the number of individuals being genotyped 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 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 our analysis of the Multiple Sclerosis (MS) GWAS dataset of the International Multiple Sclerosis Genetics Consortium. We present similar results to those obtained from the MS data.

Key words: algorithms, computational molecular biology, haplotype inference, haplotype phasing, haplotypes, long-range phasing, loss of heterozygosity.

[1]School of Science and Engineering, Reykjavik University, Reykjavik, Iceland.
[2]Center for Computational Molecular Biology and [3]Department of Computer Science, Brown University, Providence, Rhode Island.
[4]Member of the International Multiple Sclerosis Genetics Consortium GWAS Analysis team.
*These three authors contributed equally to this work.

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 polymorphisms associated to disease (Altshuler et al., 2008; Styrkarsdottir et al., 2008; The International Multiple Sclerosis Genetics Consortium, 2007), 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 (Gudbjartsson et al., 2008; Rivadeneira et al., 2009; Styrkarsdottir et al., 2008). 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. These variations, which have shown to be increasingly important and an influential factor in many diseases (Stefansson et al., 2008), 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 (The 1000 Genomes Project Consortium, 2010). 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. Over the next couple of years, it is plausible that several million individuals in the U.S. population will have been genotyped. 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 that 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 (Howie et al., 2009) 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 genotyped and those that have been resequenced (Kong et al., 2008). 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 them. In larger and more genetically diverse populations, such as the U.S. 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 another individual that shares a haplotype with this individual by descent. In this article, 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 U.S. 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 associations, the development of algorithms for detecting haplotypes from genotype data has been an active research area for several years (Clark, 1990; Halldórsson et al., 2004; Kong et al., 2008; Scheet and Stephens, 2006; Sharan et al., 2006;

Stephens et al., 2001). However, algorithms for determining haplotype phase are still in their infancy after about 15 years of development. Of particular worry is the fact that the learning rate of the algorithms (i.e., the rate that the algorithms are able to infer more correct haplotypes) grows quite slowly with the number of individuals being genotyped.

Solution 1. In this article, 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 genotyped 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 deletion polymorphisms 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 (Stefansson et al., 2008) 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 article, we run empirical tests and benchmark the algorithms on simulated GWAS datasets (Hudson, 2002) resembling the structure of the International Multiple Sclerosis Genetics Consortium data (The International Multiple Sclerosis Genetics Consortium, 2007). For our haplotype phasing and LOH algorithms, 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. The probability a tract is shared IBD increases further when multiple individuals share a tract.

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 the shared tract. We can then approximate the probability that this shared tract is 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,

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

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 $(\prod_{i=1}^{l}(f_{M,i}))^{1/l}$. Then $p_{IBS}(l) \approx (f_{avg})^{2l}$. Given f_{avg} 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, the length of the tract being considered must be large enough so the probability of IBS sharing is small.

The probability that two individuals separated by $2(n+1)$ meiosis (nth-degree cousins) share a locus IBD is 2^{-2n} (Kong et al., 2008). As n increases, the probability nth-degree cousins share a particular locus IBD decreases exponentially. However, if two individuals share a locus IBD, then they are expected to share about $\frac{200}{2n+2}$ cM (Kong et al., 2008). Relating $P(IBD)$ to the length l of a tract which starts at SNP s,

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

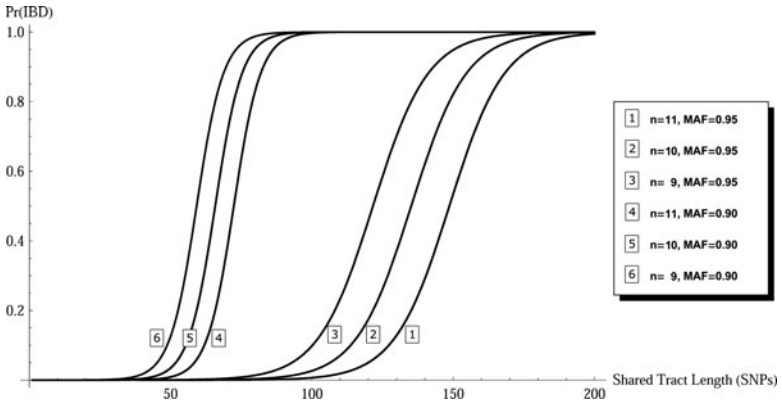


FIG. 1. Probability of identical by descent (IBD) as a function of shared tract length (measured in single nucleotide polymorphisms [SNPs]) and plotted for several n and major allele frequencies (MAF). Lower values for the MAF or n require less SNPs in a tract to commit to an IBD relationship.

which is shown in Figure 1. Figure 1 shows the probability of IBD haplotype sharing given a tract of length l. We developed our phasing algorithm based on genotype sharing which exhibits a similar trend as Figure 1, but shifted to the right (that is, we require more SNPs to commit to an IBD relationship).

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 (Sharan et al., 2006). 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.''

We computed the average number of edges in the haplotype consistency graph as a function of window size to get a sense of when the Clark consistency graph of the MS data becomes connected. Based on Figure 3 and $P(IBD)$, we can propose an algorithmic problem formulation from the Clark consistency graph. Preferably, we would like to solve either Problem 3 or 4.

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.

We believe that both of these problem formulations are NP-hard and instead propose to solve them 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 minimum length of haplotype

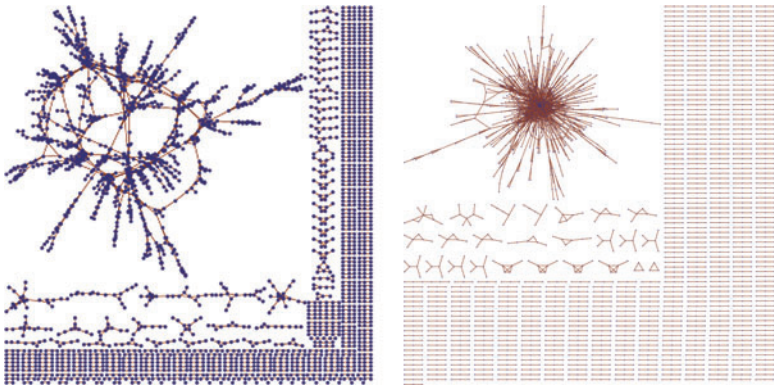


FIG. 2. **(Left)** The Clark consistency graph for SNP region (1400,1600). A large fraction of individuals share consistent haplotypes of length 200, suggesting many are IBD. **(Right)** The Clark consistency graph for a smaller window size of 180 base pairs.

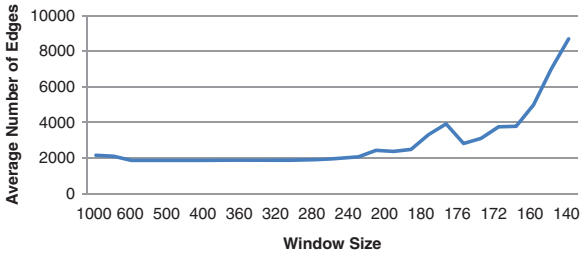


FIG. 3. The average number of edges per window size stays relatively constant until a window size of about 180. The graph becomes more connected at this point likely because the window size is small enough to not be largely affected by recombination (but still large enough for the shared tracts to not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

consistency (Fig. 3 and $P(IBD)$ aid in defining this length). We iterate through each site of each individual to find the tracts. After finding 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 in a window, all individuals can be phased at sites where there is at least one homozygote.*

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 IBS sharing to either one or both of an individuals haplotypes. The IBS sharing may be a result of the haplotype having undergone recombination, possibly a part of the haplotype is shared IBD and a part is IBS.

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:

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

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.

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.

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 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 manner similar to that of Minichiello and Durbin (2006).

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 (Browning and Browning, 2009) using a simulated dataset. Using the Hudson Simulator (Hudson, 2002), 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 two, three, and four 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 the length of sharing IBD rather than a fixed set of votes per edge, we achieve more accurate phasings.

TABLE 1. WE CREATED THREE POPULATIONS USING A BASE POOL OF 3000 SIMULATED HAPLOTYPES USING THE HUDSON SIMULATOR

	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

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.

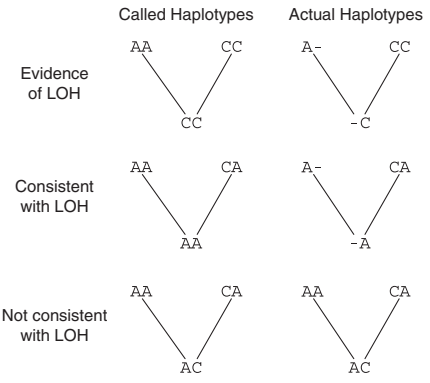
3. LOSS OF HETEROZYGOSITY REGIONS

In some situations, individuals that are heterozygous at a particular locus can possess one normal allele and one deleterious allele. We call the loss of the normal allele a loss of heterozygosity (LOH), which may be a genetic determinant in the development of disease (McCarroll et al., 2008; Stefansson et al., 2008). The detection of copy number variation, 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.

Some LOH can be inferred using data from SNP arrays. SNP calling algorithms for SNP arrays usually do not 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. Let M denote the matrix of trio genotypes. Let M_i denote the i^{th} trio of M (individuals $3i$, $3i+1$, and



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 genome-wide association studies (GWAS) data in the context of loss of heterozygosity (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. Let E represent an ELOH, C represent a CLOH, and N represent an NCLOH. The superscript defines the parent the putative deletion haplotype is associated to. We define the superscript F to be consistent with a deletion haplotype inherited from the father, M for mother, and P for both parents.

$3i + 2$). At any site j, the trio M_i may have 4^3 possible genotype combinations. To define these patterns, we use a model similar to that of Conrad et al. (2006). A trio can either be *consistent with LOH* (CLOH), be *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 a 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 a *putative deletion*. We call two putative deletions, p_i and p_j, overlapping if they share at least one 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 that 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 either of the following two conditions hold:

- h_i and h_j are members of the same putative deletion (i.e. $h_i \in [s_i, e_i]$ and $h_j \in [s_i, e_i]$)
- h_i is contained in the interval defining p_j and h_j is contained in the interval defining p_i (i.e. $h_i \in [s_j, e_j]$ and $h_j \in [s_i, e_i]$)

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 that weighs the relative costs of calling genotyping errors or deletions.

3.3. Prior work

Corona et al. (2007) developed a probabilistic algorithm that estimates frequencies of haplotypes in a region both with and without a deletion polymorphism. Conrad et al. (2006) produced a model—which is most similar to ours—for detecting deletion polymorphism. They identify inherited deletions by finding regions where particular Mendelian inheritance errors are found in close proximity. McCarroll et al. (2005) developed a clustering algorithm on bit vectors containing the signature of specific ''failure profiles'' in genotype data. All algorithms were tested on several hundred HapMap genotypes.

3.4. 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 that corresponds to using the table defined in Figure 4 (right) to translate each SNP site. This new matrix is denoted $N^{\left(\frac{|m|}{3} \times n\right)}$. To identify the genotyping errors and putative deletions, we define two operations on the ELOH sites of 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. A deletion haplotype call will identify a putative deletion as an inherited deletion haplotype. We infer inherited deletion haplotypes using the objective function

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

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 call a putative deletion with at least four 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. In the case of the multiple

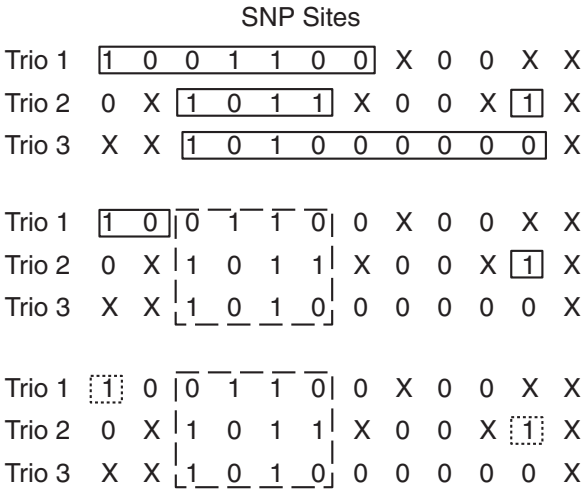


FIG. 5. A visual depiction of the greedy algorithm for finding putative deletions (consistencies with particular parents are omitted for simplicity). The solid rectangles denote trio SNP sites that have not been called yet. The dashed rectangle denotes a called inherited deletion haplotype. A dotted rectangle denotes a genotype error call. First, the algorithm finds the submatrix (a clique in G(V,E)) with the maximum trio sharing: SNP sites 3–6. Using the objective function, the algorithm either calls the set of SNPs an inherited deletion or a set of genotyping errors (in this case, the former). The intervals are updated by removing vertices and edges from the overlap graph, and the algorithm continues. Both remaining subgraphs consisting of SNP sites 1 and 11 are both cliques of size one. These will most likely be called genotyping errors.

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.

We start by giving an exact exponential algorithm that minimizes the objective function. Let x_i denote a set of overlapping putative deletions.

Algorithm 1. *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.

For practical purposes, we've developed a greedy heuristic algorithm for cases where the exact exponential algorithm is infeasible (Fig. 5).

Algorithm 2. *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, regions with strong evidence for a deletion form 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. We 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*

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 Our LOH Algorithm Using the Six Tunable Parameters as Defined in Table 2

Parameter set	*Site error probability*	*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	9	3	0.75	11	1000	0	1000
4	0.0001	7	3	0.50	15	58	0	100
5	0.00333	9	3	0.75	15	100	38888	100

Each configuration was run with a coefficient of genotyping error of 2.

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.

3.5. Experimental results on simulated data

We tested the algorithm using the same simulated dataset used to test our phasing algorithm. 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 individuals 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 1–3). 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 may be incorrectly called a genotyping error, possibly missing an LOH (Table 3, parameter set 4). 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 5). Finding appropriate tuning mechanisms for the two coefficients to maximize algorithm specificity and sensitivity will be the subject of future work.

4. CONCLUSION

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 number of individuals in the dataset. We have also given an algorithm that removes 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 author.

ACKNOWLEDGMENTS

We thank the International Multiple Sclerosis Genetics Consortium for sharing the Multiple Sclerosis GWAS dataset.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Altshuler, D., Daly, M.J., and Lander, E.S. 2008. Genetic mapping in human disease. *Science* 322, 881–888.

Browning, B.L., and Browning, S.R. 2009. 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.

Clark, A. 1990. Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol. Biol. Evol.* 7, 111–122.

Conrad, D.F., Andrews, T.D., Carter, N.P., et al. 2006. A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.* 38, 75–81.

Corona, E., Raphael, B., and Eskin, E. 2007. Identification of deletion polymorphisms from haplotypes. *Proc. RECOMB 2007* 354–365.

Gudbjartsson, D.F., Walters, B.G., Thorleifsson, G., et al. 2008. Many sequence variants affecting diversity of adult human height. *Nat. Genet.* 40, 609–615.

Halldórsson, B., Bafna, V., Edwards, N., et al. 2004. A survey of computational methods for determining haplotypes *Lect. Notes Bioinformatics* 2983, 613–614.

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

Hudson, R.R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18, 337–338.

Kong, A., Masson, G., Frigge, M.L., et al. 2008. Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.* 40, 1068–1075.

McCarroll, S.A., Hadnott, T.N., Perry, G.H., et al. 2005. Common deletion polymorphisms in the human genome. *Nat. Genet.* 38, 86–92.

McCarroll, S.A., Kuruvilla, F.G., Korn, J.M., et al. 2008. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.* 40, 1166–1174.

Minichiello, M.J., and Durbin, R. 2006. Mapping trait loci by use of inferred ancestral recombination graphs. *Am. J. Hum. Genet.* 79, 910–922.

Rivadeneira, F., Styrkársdottir, U., Estrada, K., et al. 2009. Twenty bone-mineral-density loci identified by large-scale meta-analysis of genome-wide association studies. *Nat. Genet.* 41, 1199–1206.

Scheet, P., and Stephens, M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* 78, 629–644.

Sharan, R., Halldórsson, B.V., and Istrail, S. 2006. Islands of tractability for parsimony haplotyping. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 3, 303–311.

Stefansson, H., Rujescu, D., Cichon, S., et al. 2008. Large recurrent microdeletions associated with schizophrenia. *Nature* 455, 232–236.

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

Styrkarsdottir, U., Halldorsson, B.V., Gretarsdottir, S., et al. 2008. Multiple genetic loci for bone mineral density and fractures. *N. Engl. J. Med.* 358, 2355–2365.

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

The International Multiple Sclerosis Genetics Consortium. 2007. Risk alleles for multiple sclerosis identified by a genomewide study. *N. Engl. J. Med.* 357, 851–862.

Address correspondence to:
Dr. Bjarni V. Halldórsson
Department of Engineering
Reykjavik University
Kringlan 1
Reykjavik, 103 Iceland

E-mail: bjarnivh@ru.is