Breakdown of methods for phasing and imputation in the presence of double genotype sharing

Carl Nettelblad[1*]

1 Carl Nettelblad Division of Scientific Computing, Department of Information Technology, Uppsala University, Uppsala, Sweden

∗ **E-mail: carl.nettelblad@it.uu.se**

Abstract

In genome-wide association studies, results have been improved through imputation of a denser marker set based on reference haplotypes and phasing of the genotype data. To better handle very large sets of reference haplotypes, pre-phasing with only study individuals has been suggested. We present a possible problem which is aggravated when pre-phasing strategies are used, and suggest a modification avoiding these issues with application to the MaCH tool.

We evaluate the effectiveness of our remedy to a subset of Hapmap data, comparing the original version of MaCH and our modified approach. Improvements are demonstrated on the original data (phase switch error rate decresasing by 10%), but the differences are more pronounced in cases where the data is augmented to represent the presence of closely related individuals, especially when siblings are present (30% reduction in switch error rate in the presence of children, 47% reduction in the presence of siblings). When introducing siblings, the switch error rate in results from the unmodified version of MaCH increases significantly compared to the original data.

The main conclusions of this investigation is that existing statistical methods for phasing and imputation of unrelated individuals might give subpar quality results if a subset of study individuals nonetheless are related. As the populations collected for general genome-wide association studies grow in size, including relatives might become more common. If a general GWAS framework for unrelated individuals would be employed on datasets where sub-populations originally collected as familial case-control sets are included, caution should also be taken regarding the quality of haplotypes.

Our modification to MaCH is available on request and straightforward to implement. We hope that this mode, if found to be of use, could be integrated as an option in future standard distributions of MaCH.

Introduction

Genome-wide association studies have shown great success in unravelling the genetic variation underlying many important traits and disease complexes in natural human populations [1, 2]. Imputation of marker data has been suggested, both as a way to augment missing or sparse genotype data based on reference haplotypes from sequenced reference haplotypes [3], and in order to reconcile study cohorts assembled from genotyping efforts using different SNP panels [4]. The process of imputation consists of inferring the genotype phase for all markers, and then finding the best corresponding genotypes in the reference population, for those markers that are missing in experimental data. The underlying assumption is that short haplotype blocks are most likely preserved over the course of many generations. Thus, a suitable panel of reference haplotypes can be highly informative for genotypes not observed directly, and increase detection power.

Panel sizes are constantly growing, from the tens or hundreds in original Hapmap populations [5], into the current goal of a total of 1 000 high-quality human genomes from the 1000 Genomes Project [6]. However, some popular algorithms for genotype imputation scale as $O(n^2)$ [7, 8] per individual analyzed, where n is the total number of haplotypes (reference and study). An increase in panel size by a factor of

100 might increase runtime by a factor of 10 000, exhausting computational resources. Other approaches exist [9], but reduce computational complexity by making additional approximations.

Due to the rapid increase in the computational complexity of Markov model phasing with increasing reference population size, it has been suggested to infer the phases using only the study population (or a subset thereof), followed by imputing genotypes into this fixed (pre-phased) haplotype set [10]. This operation reduces the computational complexity, allowing much larger reference panel sizes. However, as no known fixed haplotypes are available during pre-phasing, the Markov chain approaches used in the most popular pre-phasing schemes become more sensitive to the problem of chain trajectories getting stuck in local minima. In this paper we describe a specific scenario causing the model getting stuck, show the extent of the problem with experimental data, and suggest a possible modification of the MaCH [7] algorithm successfully circumventing the issues.

Materials and Methods

Most hidden Markov model approaches for phasing of genotype data lacking a pedigree share several characteristics [11]. A state in the model consists of a haplotype pair, meaning that an observed unordered genotype pair in one individual corresponds to a pair of haplotypes from other individuals. With a proper selection of transition probabilities, blocks of the genome will be attributed to identical states, reflecting identical ancestry. The posterior probabilities for the state distribution can be found at each position, and putative haplotype candidates can be determined by sampling from that distribution. By iterating over all individuals, the undetermined (sampled) haplotypes can be successively improved.

Consider that such a successive improvement is underway, and that the next step is to sample new haplotypes from the posterior distribution for individual A. Also assume that individuals A and B are completely identical, over a major stretch of a chromosome. If they are, a problem will arise. It is sufficient that the individuals are ordinary full siblings for this to occur. Approximately $1/4$ of the total genome for a pair of siblings will consist of such very long regions, as crossover events are relatively far apart relative to the marker density in modern maps. The posterior probability when individual A is analyzed will be completely dominated by the haplotypes for B in such a region. However, this dominating effect will only be justified if the haplotypes for B are truly correct. As the genotypes match in every position, *any* haplotype resolution for B will have a dominating influence on A. Correspondingly, any haplotype resolution for A will have a dominating influence on B. In an iterative optimization scheme starting out from randomly initialized haplotypes without an external reference, the pair of A and B will be locked in a local minimum very close to the starting point.

If transition probabilities are also iteratively updated based on observed data, the problem is further compounded, as the single very favorable state also makes transition events rare. This will make transitions even more infrequent in later iterations, further decreasing the probability of sampling another haplotype.

The effect is not necessarily confined to two individuals. If a larger set of individuals share a comparatively long region on both strands, i.e. carry identical-by-state genotypes for all markers in a long region, the same kind of lock-in effect will appear. The state distribution will consist of a mix of states, but it will be almost totally occupied by different combinations of haplotypes from the set of similar individuals, and the sampling at each marker in each iteration will almost always be drawn from this set, thus only reflecting the initial randomization of phase.

Our proposed remedy to this is to keep the current model formulation, but improving the mixing properties of the sampling process. The sampling process in MaCH [7] starts from the last marker, iteratively going backwards, sampling based on the forward probabilities given the state at the previous marker sampled. Specifically, there is a vector for all unique pairs of n haplotypes. What should be filtered out is those cases where the pair taking one haplotype from B and one from C ($B0C0$) is just as likely as taking the other haplotype from both individuals ($B1C1$). When that is the case, any haplotype

resolution would match, as per the reasoning above. Thus, the match can be uninformative, causing a local (incorrect) minimum to be maintained. The actual sampling probability used for $P'(B0C0)$ is, with our modification, instead $P(B0C0) - P(B1C1)$ (assuming the result is positive), where P is the forward probability. In the case where $B = C$, the result is that sampling the "copy another individual" pair $B0B1$ is precluded, as $P'(B0B1) = P(B0B1) - P(B0B1) = 0$. By only modifying the sampling probability, our approach does not affect the overall structure of the model.

Experiments using Hapmap population data

In order to verify the extent of the problem when phasing a small set of realistic dense human data, we used the 60 first chromosome 21 haplotypes (30 parents with 19,306 markers) of the phased Hapmap3 release 2 Utah residents with Northern and Western European ancestry from the CEPH collection (CEU) trios [12]. The full set of identified SNPs available in the phased data were used. Half of them were cleared to be used as a test set for imputation. As the data only consisted of the parental generation in the dataset, the individuals were not supposed to be closely related.

In order to introduce a high degree of double genotype sharing, the problem condition we are interested in, we also created modified datasets based on this original data. These included adding back the child in each trio set (45 individuals in total); resampling the haplotypes into new virtual individuals with a degree of sharing (1/4) consistent with the presence of siblings to the existing parents; and finally introducing a mono-zygotic twin to each parent (60 individuals in total).

Phasing in MaCH was run for 2 000 iterations. While benefits of more than 200 iterations were limited, we were interested in discovering whether the near-asymptotic behavior of the original MaCH and our modified version were identical. It could be argued that the improved mixing of our modification would only speed up convergence, but not affect the results with a high number of iterations.

After phasing was completed, the number of switch errors in the phase sequences were counted compared to the original phased data. The switches, or flips, were only counted for the 30 original parents for all datasets, in order to make the numbers directly comparable. The phased data as well as estimated genotype error and recombination rates were then fed to `minimac` for imputation using the remaining 57 parents in the trio dataset as a reference panel.

Results

We have implemented the modification outlined in the Methods section in MaCH. The change could easily be added to the main source tree as an extra option. Instructions on how to make the corresponding changes to the source are available on request. The performance of our modified approach is demonstrated in Table 1, with comparisons relative to an unmodified version of MaCH 1.0.17. Clear improvements are demonstrated for the number of switches needed to represent the true haplotypes (as reported by the Hapmap consortium), as well as in imputation accuracy, even for a dataset consisting of supposedly unrelated individuals. When artificial siblings were added, compounding the problem, the effects are far more drastic.

Our modified version results in modest improvements in error rate for flips as well as imputed alleles for the unmodified dataset, despite the fact that no long regions of double genotype sharing would generally be expected in unrelated individuals. For the other cases, the differences detected are drastic. The error rate in flips at most rises by 30% for our modified version (in the case of simulated MZ twins). The original MaCH phasing breaks down in this scenario, with an almost eight-fold increase in the flip error rate and doubling of imputation errors.

In the more plausible scenario of siblings rather than twins being present, the original MaCH error rate still increases by over 70%.

Although the differences in results are modest in some cases, we have also seen the original MaCH method to be much more sensitive. Including all markers, rather than leaving every second out for imputation purposes in our experiment, will increase the switch error rate dramatically for the original MaCH, but only results in a modest increase in our modified version. The total number of switch errors for the 30 CEU parents tested in our experiments, when no markers are masked, are 6 597 for our modified version and 14 770 for the original MaCH. Figure 1 shows the location of individual flip errors for the unmasked parent dataset, indicating that the original MaCH version will sometimes create long regions of repeated phasing errors that also coincide between multiple individuals, as predicted.

Discussion

We think that our results regarding the extent of deterioration in haplotype quality when some types of related individuals are included in the data should be of interest to all situations where imputation or phasing based on Markov model methods are used, but especially so in the case where pre-phasing is used followed by imputation with e.g. `minimac`, or when it is known that some of the individuals to be phased might be closely related. It is also relevant to point out that even in a dataset with supposedly unrelated individuals, our remedied version reduces the switch error rate by 50% when no markers were masked.

It should be noted that the degree of relationship required for the issue of double genotype sharing to be present does not have to be as close as full siblings. Rather, the relevant condition is whether there is some probability that two individuals share both homologues of a certain region identical by descent. This could be the case for e.g. double cousins, but the condition could also hold for far shorter regions (but still on the range of multiple Mbps) in relatively isolated populations with little historical exchange of genetic material.

The effects seen in switch error rate are not fully reflected in the imputation error rate. We suggest that this is due to the insufficient size of the very limited reference panel used in this specific experiment. The quality of the pre-phasing only influences imputation quality when the reference sets contain matches to the true haplotypes.

If one is reluctant to use our remedy or other modifications of existing haplotype inference approaches, we still suggest investigating the quality of phasing, both in pre-phasing schemes and more traditional schemes where reference haplotypes are present in all iterations. One way to do so is to perform cross-validation of the phasing of the study population, creating different subsets where e.g. 20% of individuals are left out, counting the number of flips when comparing the resulting haplotypes for individuals common between subsets. Regions of individual genomes where the number of flips are high indicate that the resulting haplotypes are influenced by the information from only a few other individuals in the population, possibly indicating the issue of insufficient chain mixing noted here.

Acknowledgments

The computations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at Chalmers Centre for Computational Science and Engineering (C3SE). Gonçalo Abecasis and Yun Li are thanked for providing additional details on MaCH and giving valuable comments.

References

1. Burton P, Clayton D, Cardon L, Craddock N, Deloukas P, et al. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661–678.

2. Turnbull C, Ahmed S, Morrison J, Pernet D, Renwick A, et al. (2010) Genome-wide association study identifies five new breast cancer susceptibility loci. Nature Genetics 42: 504–507.

3. Li Y, Willer C, Sanna S, Abecasis G (2009) Genotype imputation. Annual Review of Genomics and Human Genetics 10: 387-406.

4. Druet T, Schrooten C, de Roos A (2010) Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. Journal of Dairy Science 93: 5443 - 5454.

5. International HapMap Consortium (2003) The international HapMap project. Nature 426: 789-796.

6. 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073.

7. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genetic Epidemiology 34: 816–834.

8. Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genetics 5: e1000529.

9. Delaneau O, Coulonges C, Zagury JF (2008) Shape-IT: new rapid and accurate algorithm for haplotype inference. BMC Bioinformatics 9: 540.

10. Abecasis G, Fuchsberger C (2012). Minimac. http://genome.sph.umich.edu/wiki/Minimac.

11. Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. The American Journal of Human Genetics 78: 629–644.

12. Altshuler D, Gibbs R, Peltonen L, Dermitzakis E, Schaffner S, et al. (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467: 52-58.

Figure Legends

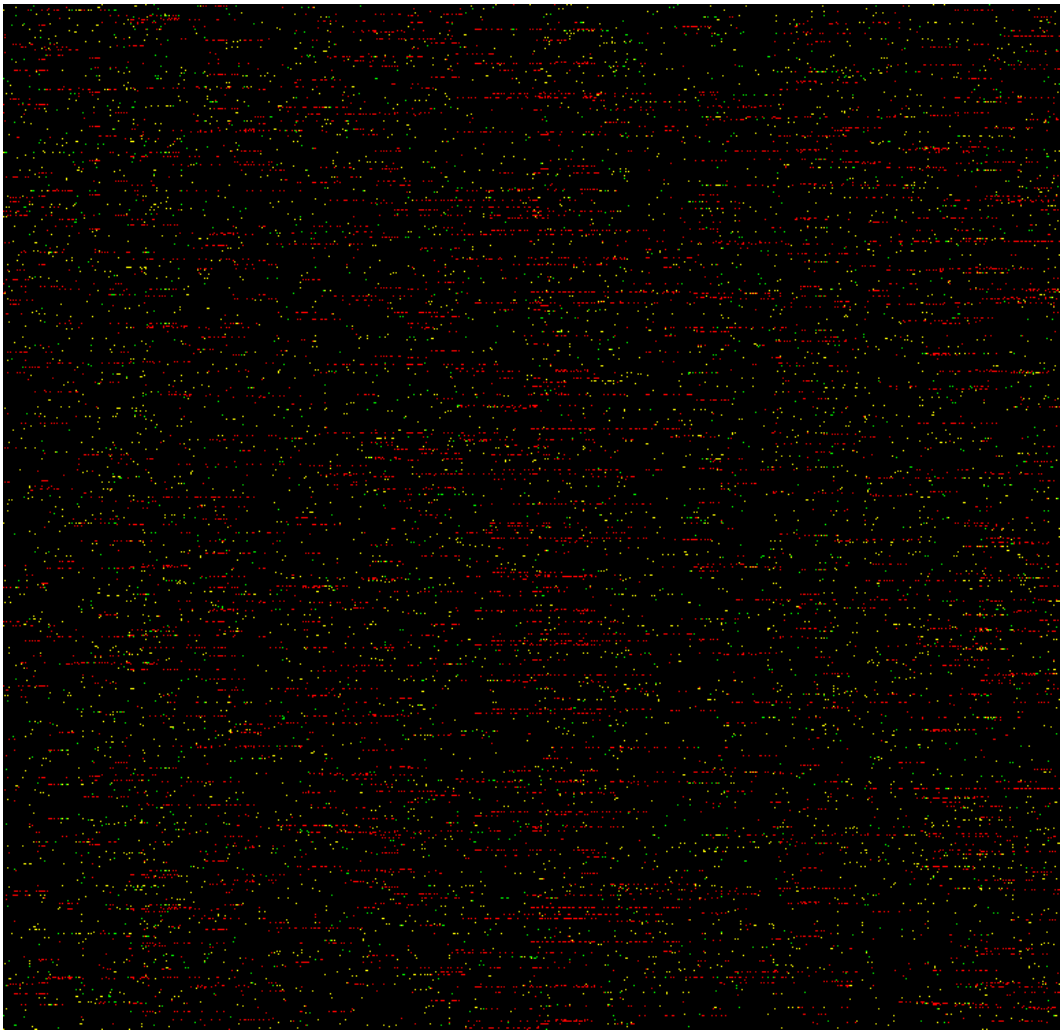

Figure 1. Comparison of switch error locations. The figure depicts all 19 306 markers for the 30 first CEU trio parents. Individuals and markers are stacked vertically, with 773 markers per individual row. Switch errors in original MaCH are coded red, errors from our modified version green, while coinciding errors are yellow. As the total error count (compare Table 1 and the Results section) is lower with our version, the number of unique green dots is limited. More importantly, green and yellow dots are mostly scattered, while red dots also occur in long horizontal stretches, indicating that phase is switched at almost every other marker within these regions. In other words the resulting phase in these places is no better than random, which is what one would predict according to the lock-in scenario in the Markov chain sampling process that we describe. This figure also shows that even if overall haplotype quality in terms of error rate would be acceptable, some regions can still be heavily affected, and paradoxically those regions are the ones where multiple individuals share both haplotypes identical by descent.

Tables

Table 1. Comparison between original and modified MaCH

Dataset	Total # flips	# incorrectly imputed marker alleles
No children, original MaCH	5 408	3 730
No children, modified MaCH	4 915	3 566
With children, original MaCH	1 907	3 261
With children, modified MaCH	1 350	3 217
Normal siblings, original MaCH	9 657	4 611
Normal siblings, modified MaCH	5 096	3 616
Mono-zygotic twins, original MaCH *	42 074	8 787
Mono-zygotic twins, modified MaCH	6 309	4 016

Comparison between original MaCH and a modified version with our remedy, showing both the total number of flips and the number of incorrectly imputed alleles. The comparison is based on the 30 first phased Hapmap3 release 2 CEU trio parents [12]. Four versions are used: 1. the original dataset (only parents), 2. including their children, as well as 3. simulating siblings to parents, 4. simulating twins to parents. When children are excluded and no virtual siblings are present, no known relationships exist between the individuals in the dataset. Imputation performance was verified by reconstructing the half (9, 653) of markers left out using `minimac` [10], with the remainder of the phased CEU trio data (57 individuals) employed as reference panel. All MaCH runs were executed for 2, 000 iterations, with 20 rounds for `minimac`. Metrics are reported for only the 30 original individuals, in order to aid comparisons.

* The `minimac` run starting from the recombination frequencies determined by MaCH in this case failed to converge at all, with errors for all markers. The results for this row in the table are based on the pre-phased haplotypes from original MaCH, but starting out with the recombination frequencies from the modified version, in order to allow the `minimac` imputation to complete at all.