

Little loss of information due to unknown phase: redux

Thomas Mailund

June 29, 2007

I describe an extension to the HapCluster algorithm from Waldron et al.⁶ that extends the algorithm to deal with unphased data by integrating over unknown phase similarly to the approach taken in Morris et al.⁴.

Introduction

The HapCluster algorithm searches for the location of a variant affecting disease status by trying to identify flexible clusters of case haplotypes in the vicinity of a mutant allele. Rather than explicitly modelling the genealogy of the sample, the algorithm uses a locus-dependent distance measure between haplotype-sequences to identify a cluster of sequences close to the “ancestor sequence” where the mutation occurred. The “ancestor sequence” and distance threshold are nuisance parameters integrated out in the MCMC. Sequences within a certain distance of the ancestor sequence are considered to carry the risk affecting allele, all other sequences are considered non-carriers.

This approach is inherently working with phased sequences, in the sense that the sequence distance used for the clustering needs to compare phased sequences in order to assign genotypes to individuals. However, phased genotypes are not readily available from current genotyping technology, and must be inferred. While state-of-the-art statistical methods for inferring phase are very accurate, they are typically also very time consuming. Even heuristics that sacrifice some accuracy for increased runtime performance, such as the fastPHASE method⁵ will use hours or even days to infer the phase of even moderately sized datasets, making the phase inference the bottleneck in applying the HapCluster method.

In HapCluster, the accurate phase is only necessary for carrier chromosomes, and only locally around the disease affecting locus: the phase is only needed to calculate a local sequence distance around a potential disease locus. If we can integrate the mapping algorithm with phase inference, then, ideally, it would only be necessary to infer the phase of chromosomes in the cluster (and their paired chromosome, of course, but leaving homozygotic wildtypes free to assume any phase without affecting the likelihood) and only in a small region around the disease affecting locus.

In this note I describe a simple extension to the MCMC from Waldron et al.⁶ that uses the same approach as in Morris et al.⁴ to integrate over unknown phase within the MCMC algorithm.

Method

The input to the method is a list of sequences, $\mathcal{G} = \{g_i\}, i = 1, \dots, n$, each of length m , together with phenotypes $p_i, i = 1, \dots, n, p_i \in \{0, 1\}$, describing the genotypes and disease phenotype of individuals 1 to n . The g_i sequences represent the genotypes of consecutive markers over a genomic region of interest, and are sequences over the alphabet $\{0, 1, 2, ?\}$, such that $g_{ij} = 0$ denotes that individual i is homozygote 0 on marker j , $g_{ij} = 1$ denotes that individual i is homozygote 1 on marker j , $g_{ij} = 2$ denotes that individual i is heterozygote on marker j and $g_{ij} = ?$ denotes a missing value for individual i on marker j .

We define a *phasing compatible with \mathcal{G}* as a list of sequences $\mathcal{H} = \{h_i^k\}, i = 1, \dots, n, k = 1, 2$ such that $h_{ij}^1 \neq h_{ij}^2$ if $s_{ji} = 2$ and otherwise $h_{ij}^1 = h_{ij}^2 = s_{ij}$. The algorithm in Waldron et al.⁶ operated on a single inferred phase compatible with such a dataset; the approach we will take here is to integrate over all phasings compatible with our input with the MCMC algorithm.

To do this, we only need to specify a prior distribution over phasings of \mathcal{G} : $p(\mathcal{H} | \mathcal{G})$ and a change proposal distribution $p(\mathcal{H}' | \mathcal{H}, \Theta)$ (where Θ denotes all other parameters in the model). The distributions and proposals from the original algorithm can then be considered conditional on \mathcal{H} —as they implicitly were, as \mathcal{H} was the input of the algorithm in the original MCMC—and by extending the appropriate equations in Waldron et al.⁶ with these two distributions, we get an MCMC operating on \mathcal{G} instead of a fixed \mathcal{H} .

In the simplest setup, we can choose a uniform prior for $p(\mathcal{H} | \mathcal{G})$ and a reversible change proposal distribution independent of all other model parameters except for \mathcal{G} : $p(\mathcal{H}' | \mathcal{H}, \Theta) = p(\mathcal{H}' | \mathcal{H}, \mathcal{G}) = p(\mathcal{H} | \mathcal{H}', \mathcal{G})$. These terms will then cancel out in the acceptance probability in the MCMC, and new states of the chain will be accepted using the same expression as in Waldron et al.⁶ (but will of course still depend on the current and the proposed phasing since the expressions are all now conditional on \mathcal{H} and \mathcal{H}' respectively).

The update on phase, $\mathcal{H} \rightarrow \mathcal{H}'$, we propose¹ is the following: With a fixed probability, p , we leave the phase unchanged ($\mathcal{H}' = \mathcal{H}$). Otherwise, with probability $1 - p$, we choose a random individual, $i \sim U(1, n)$, and a random marker $j \sim U(1, m - 1)$, and flip the phase of all markers to the right of j between the two chromosomes of individual i : $\forall k > j : h_{ik}^1 \leftrightarrow h_{ik}^2$. This change clearly satisfy $p(\mathcal{H}' | \mathcal{H}, \mathcal{G}) = p(\mathcal{H} | \mathcal{H}', \mathcal{G})$, and assuming \mathcal{H} is compatible with \mathcal{G} , then so will \mathcal{H}' be.

In our implementation we arbitrarily choose $p = 0.95$ meaning we update the phase of a single individual in about 5% of the iterations.

¹We do not need to ever calculate the value $p(\mathcal{H}' | \mathcal{H}, \mathcal{G})$ but we do need to propose the state change to explore the parameter space.

Results

To assess the mapping accuracy when integrating over unknown phase as described above, we compared the method running on unknown phase with the original MCMC running on the known true phase. For population risk 10% and an additive disease model with genetic relative risk (GRR) 1.5, 2.0 and 4.0 for the causative marker, we simulated 10 datasets with 1000 cases and 1000 controls and 100 markers on a region of length $\rho = 400$, using the CoaSim simulator¹. Results are shown on figures 1–3.

The point estimates for GRR 1.5 are generally good, but the confidence interval is very wide, in several cases spanning almost the entire region, and while the point estimates seem to be equally good for phased and unphased data, the confidence intervals seem to be wider for unphased data. For the higher GRR, there seems to be little difference between the phased and unphased datasets. However, the confidence intervals contain the true position less than 95% of the time, indicating that both methods are somewhat overconfident.

Discussion and Conclusion

By adding a very simple phase inference to the MCMC from Waldron et al.⁶ we extend the algorithm to analyse unphased genotype data, removing the time-consuming phase inference preprocessing. Our preliminary experiments show, that integrating over unknown phase in this manner, does not significantly—if at all—reduce the localisation accuracy of the method, even when comparing the unknown phase analysis with results for the true known phase.

The extension to integrate over unknown phase is the same as used in Morris et al.⁴, but where the model in that paper—see Morris et al.³ for details—uses the inferred phase in an explicit modelling of the genealogy of cases, the inferred phase in our new method only affects the model in how chromosomes cluster. This makes for a much simpler approach, with great savings in runtime as a consequence.

The method also resembles the approach taken in Molitor et al.². Here, spacial clustering is also used to approximate the local genealogy of cases, and the unknown phase is dealt with in the MCMC integration as well. Unlike their method, however, we focus on a single contiguous region around a potential disease locus and attempt to phase that region sufficiently well to obtain the right clustering, while they consider arbitrary subsets (of fixed size) of markers.

References

1. T. Mailund, M. Schierup, C. Pedersen, P. Mechlenborg, J. Madsen, and L. Schauer. CoaSim: A flexible environment for simulating genetic data under coalescent models. *BMC Bioinformatics*, 6(252), 2005. doi: 10.1186/1471-2105-6-252. URL <http://dx.doi.org/10.1186/1471-2105-6-252>.

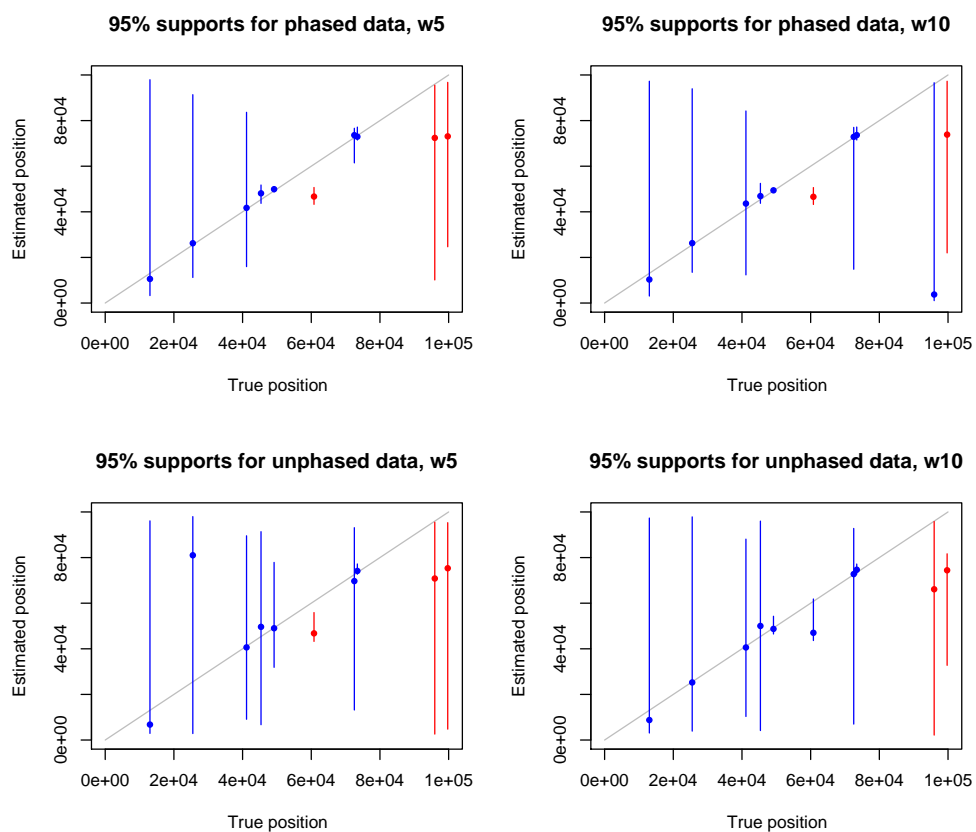


Fig. 1: Localisation accuracy for genetic relative risk (GRR) 1.5. Top row shows accuracy for phased input (with the true phase) and the bottom row shows accuracy for unphased input. The right-most plots show analysis where the max window size was set to 5 markers on either side of the locus, the left-most plots show analysis where the max window size was set to 10 markers on either side of the locus. Dots show posterior modes, while the vertical segments show the 95% support (HDR) for the posterior. Blue is used for datasets where the true locus is contained in the HDR, red for datasets where the true locus is not contained in the HDR.

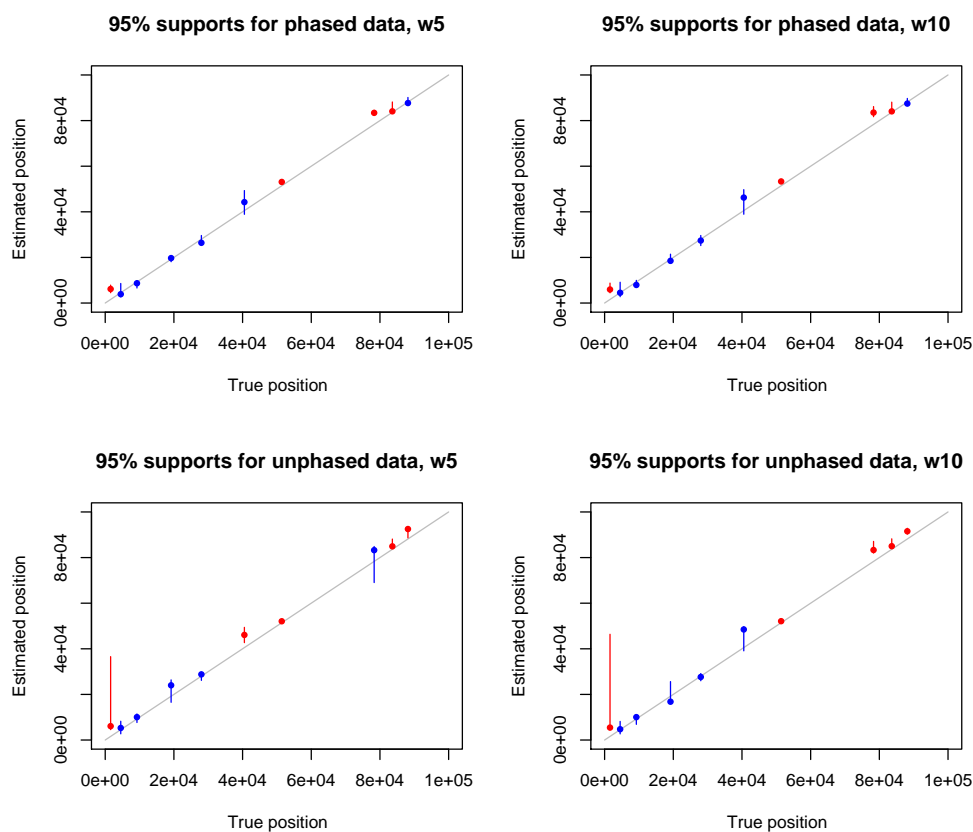


Fig. 2: Localisation accuracy for genetic relative risk (GRR) 2. Top row shows accuracy for phased input (with the true phase) and the bottom row shows accuracy for unphased input. The right-most plots show analysis where the max window size was set to 5 markers on either side of the locus, the left-most plots show analysis where the max window size was set to 10 markers on either side of the locus. Dots show posterior modes, while the vertical segments show the 95% support (HDR) for the posterior. Blue is used for datasets where the true locus is contained in the HDR, red for datasets where the true locus is not contained in the HDR.

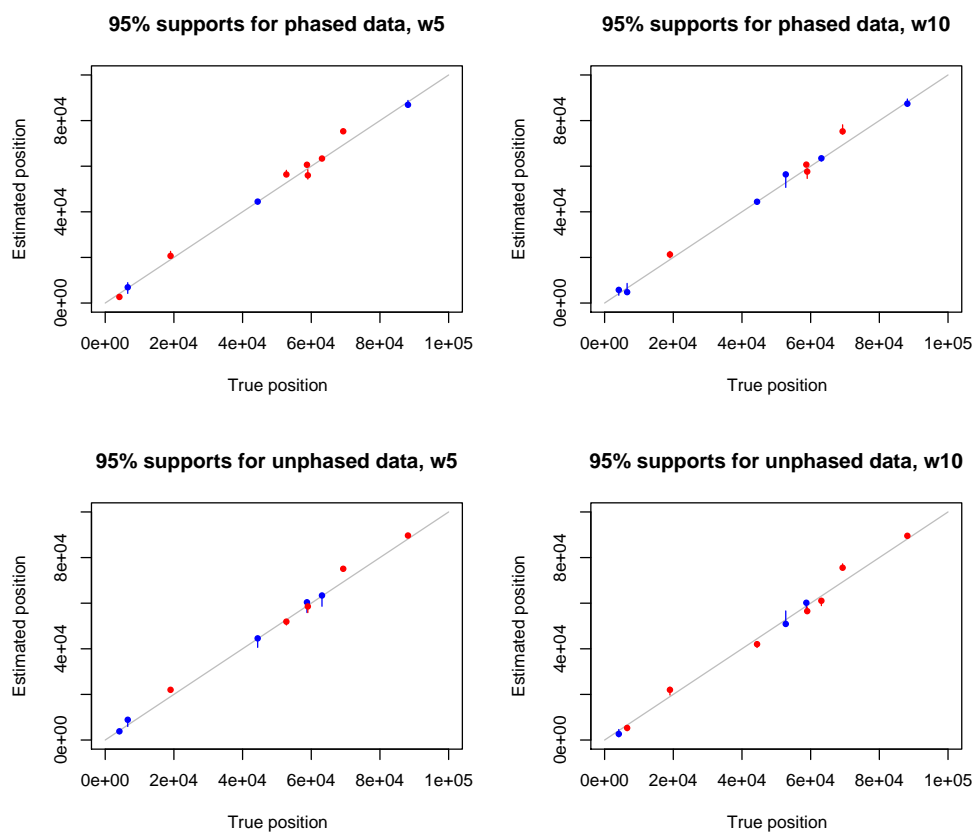


Fig. 3: Localisation accuracy for genetic relative risk (GRR) 4. Top row shows accuracy for phased input (with the true phase) and the bottom row shows accuracy for unphased input. The right-most plots show analysis where the max window size was set to 5 markers on either side of the locus, the left-most plots show analysis where the max window size was set to 10 markers on either side of the locus. Dots show posterior modes, while the vertical segments show the 95% support (HDR) for the posterior. Blue is used for datasets where the true locus is contained in the HDR, red for datasets where the true locus is not contained in the HDR.

2. J. Molitor, K. Zhao, and P. Marjoram. An integrated approach to fine mapping of unphased genotype data. Personal communication.
3. A. P. Morris, J. C. Whittaker, and D. J. Balding. Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. *Am J Hum Genet*, 70(3):686–707, 2002.
4. A. P. Morris, J. C. Whittaker, and D. J. Balding. Little loss of information due to unknown phase for fine-scale linkage-disequilibrium mapping with single-nucleotide-polymorphism genotype data. *Am J Hum Genet*, 74(5):945–953, 2004.
5. P. Scheet and M. Stephens. 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(4):629–44, 2006.
6. E. R. B. Waldron, J. C. Whittaker, and D. J. Balding. Fine mapping of disease genes via haplotype clustering. *Genet Epidemiol*, 30(2):170–179, 2006.