A RECURSIVE ALGORITHM FOR LONG RANGE PHASING OF SNP GENOTYPES

B. P. Kinghorn, J.M. Hickey and J.H.J. van der Werf

The Institute of Genetics and Bioinformatics,
University of New England, Armidale, Australia, NSW 2351

SUMMARY

High density genotyping of individuals does not by itself yield information on phase: linkage of alleles between loci in the contributing gametes. This additional information is important, because it helps to determine which segments of chromosome are identical by descent across the population, and thereby improve inference about segregation of quantitative trait loci (QTL). This paper presents a recursive algorithm to give efficient implementation of long range phasing, a recently-published strategy that infers IBD between distant relatives from long strings of SNP genotypes that show no opposing homozygotes (i.e. no evidence of lack of IBS). These relatives can be both distant and unknown to the analysis. Successful implementation of this strategy gives opportunity to use haplotype and/or combined linkage-linkage disequilibrium analysis for QTL mapping or genomic selection.

INTRODUCTION

Information on phase gives added power to map and/or exploit QTL, and to make inference about causal mutations. Without full linkage disequilibrium, alleles at a QTL will vary within the group of gametes that contain a given allele at the most informative marker locus (eg. for all gametes carrying marker allele M, some will carry QTL allele q, and some will carry Q). This reduces power to detect or exploit that QTL. Full phasing can act to place a notional subscript to each marker allele, denoting the identity by descent (IBD) segment of chromosome to which it belongs, and giving a tight relationship with the QTL allele carried.

Phasing at an individual locus can often be carried out with knowledge of the genotypes of the proband (the individual of interest) and its two parents. Only when all three are heterozygous at a biallelic marker locus is it not possible to declare which progeny allele has been inherited from which parent. In such cases, we can replace at least one of the heterozygous parents with a "surrogate parent" that is homozygous, and is known with high probability to carry the parental allele that the real parent transmitted to the proband. This is a novel approach presented by Kong *et al.* (2008). Its key feature is recognition that each parent of a proband can be represented by surrogate parents – individuals that carry the same parental haplotype at the region concerned, and can thus act in place of the parent when deducing which allele has been transmitted from parent to offspring. The method is termed "Long Range Phasing", which reflects the fact that information used for phasing the proband can come from individuals that are distantly connected by pedigree, at long range, with many degrees of separation involved.

If neither the parent nor any of its surrogates are homozygous, then there is still a chance inferring the solution: if any one of the heterozygous surrogates can have its genotype phased, then its homozygous parent or surrogate must be IBD at the opposite haplotype, and, being a homozygote, cannot be a surrogate of the original proband, which must thus have inherited the opposite allele. This second layer of surrogates are at an Erdös number of 2 from the proband (See Kong *et al.*, 2008 for more detail). There is no prior limit to the Erdös number of surrogates that can be used, such that phasing can be carried out using information from distantly related individuals, and non-relatives such as "in-laws" that are connected in the pedigree.

Surrogacy is bidirectional: if A is a surrogate of B, then B is a surrogate of A. Surrogacy is declared upon sufficient evidence that two individuals share a haplotype at the prevailing region. This in turn is inferred by detecting that no opposing homozygotes exist between the individuals over a long distance involving many genetic markers. With sufficient distance and marker density, there is a sufficiently small chance that no opposing homozygotes would occur without IBD due to a common haplotype.

This paper presents algorithms for long range phasing of biallelic markers in a population, given that surrogacy relationships have already been inferred. These algorithms use surrogates available or required at all levels of Erdös number for each individual/SNP combination that is phased. For simple presentation, incompatibilities and errors are not dealt with in this paper, but are addressed by Hickey *et al.* (2009). Kong *et al.* (2008) used an iterative algorithm with steps to deal with incompatibilities and errors. The recursive algorithm presented here has the apparent advantage of using more information, because all surrogates beyond Erdös number 2 can be used, not just those that can be assigned to the paternal or maternal side of the pedigree of the individuals for which they are direct surrogates.

METHODS

The simplest component of the overall task is to infer the source of inheritance (paternal or maternal) of each allele in a heterozygous individual. If genotyping and pedigree errors are absent, we only need to do this for one of the two alleles. For simple presentation we will refer to this process as phasing, even though it is done for a single locus at a time. The individual to be phased is referred to as the proband, and it is taken to be heterozygous, as the case for homozygosity is trivial: the same allele is inherited from each parent.

Each marker locus is handled in turn. Because of this, there is a need to separate maternal and paternal surrogates for the proband itself, so that direction of inheritance for each phased marker can be aligned with those for the linked markers.

The recursive algorithm. The first call is to a non-recursive function (steps 3, 4, 5 below) to give special treatment to the proband. If phasing is not made using the direct maternal and paternal surrogates of the proband, this function makes calls to the recursive function (steps 6, 7, 8 below) which does not distinguish between maternal and parental surrogates.

The dummy argument for the recursive function is referred to as the "current individual" (CI), and the function result is either an allele (as described at step 6), or a code denoting no success in phasing the CI. As no CI is the proband, it is sufficient to identify a homozygous surrogate from either side of its pedigree, as this defines the allele that is in the opposite haplotype to the one that is IBD with that in the calling individual, one Erdös level below. The recursive function has a generally bigger overall pool of surrogates, because it includes those surrogates that cannot be allocated to one (or occasionally both) side(s) of the CI's pedigree.

1. Erdös number, stored globally, is set to one.
2. The non-recursive function (steps 3, 4, 5) is called for the proband.
3. A pass is made across all the individuals that are a surrogate of both the proband and its father, looking for a surrogate that is homozygous. If a homozygote is found, the allele of the homozygote is the allele inherited by the proband from its father, and the non-recursive function is exited to Step 9 with a value denoting the paternal allele.
4. As for step 3, but replace 'father' with 'mother'. However, the 'paternal' allele is still the allele reported to Step 9.
5. If no homozygote is found, the non-recursive function continues, the Erdös number is incremented by one, and ***the recursive function is called*** (starting at Step 6) in turn for

each of the surrogates described in step 3, and then in step 4 with an allele swap to report the paternal allele. If a function call returns a positive paternal allele result, the non-recursive function is exited to Step 9 with a value denoting the paternal allele. If no positive paternal allele result arises, Erdös number is decreased by one and the non-recursive function is exited to Step 9 with a value denoting lack of a phasing call.

6. ***The recursive function is entered*** with the CI as the dummy argument and the alternative allele status of the CI as the result. A pass is made across all the individuals that are a surrogate of the CI, looking for a surrogate that is homozygous. If a homozygote is found, the allele of the homozygote is the allele inherited by the CI at the alternative site to that which the CI has transmitted to the calling individual (for which the CI is a surrogate), and the current instance of the function is exited into Step 5 or Step 7, from where it was called, with a value denoting the alternative allele.
7. If no homozygote is found, the prevailing Erdös number is incremented by one, and ***the recursive function is called*** in turn for each of the surrogates described in step 6. As soon as a function call returns a positive alternative allele result, the current instance of the function is exited into Step 5 or Step 7, from where it was called, with a value denoting the alternative allele.
8. If the current instance of the recursive function has not been exited, the Erdös number is decreased by one and the function is exited into Step 5 or 7, from where it was called, with a value denoting lack of success in phasing the CI.
9. From Step 5 alone. If a positive result returns for the proband, the Erdös number in which homozygosity was found has to be taken into account: If mod(ErdösNumber,2)=0 then the allele of the result is swapped so that it pertains to the proband. [EG. when ErdösNumber =2, a heterozygous surrogate of the proband has had one of its alleles phased, and it is the other allele that is IDB with the proband.] After this manipulation, the reported allele is the paternally inherited allele.

The algorithm can be seen to be acting as a ratchet, increasing Erdös number when no homozygote has been found in the current Erdös 'layer', for the current part of the pedigree. In the end, if no homozygote is found, the ratchet slips back through shells of the recursive function to return a negative result for the proband. However, given the recursive nature of the algorithm, multiple pathways are searched.

An iterative algorithm.

1. For each unphased individual:
    a. Cycle through all paternal surrogates. If a homozygote is found, record that allele as the paternal allele, flag the individual as phased and loop to the next individual.
    b. Cycle through all maternal surrogates. If a homozygote is found, record the alternate allele as the paternal allele, flag the individual as phased and loop to the next individual.
    c. Cycle through all paternal surrogates. If a surrogate is found to have been phased, record the alternative allele as the paternal allele in the individual, flag the individual as phased and loop to the next individual.
    d. Cycle through all maternal surrogates. If a surrogate is found to have been phased, record that allele as the paternal allele in the individual, flag the individual as phased and loop to the next individual.
2. If Step 1 resulted in at least one new phasing, go to Step 1, else stop

This iterative algorithm is simpler, but it is less powerful because at any stage it uses surrogates to make inference about phase in a current individual. In contrast, the recursive algorithm, at all levels above Erdös 1, only needs to make inference about which allele in the surrogate belongs to the haplotype that is not IBD with the proband. It does not matter whether it was paternally or maternally inherited. This permits inclusion of all surrogates, including those that are not connected by pedigree and those that are surrogates to individuals with unknown parents.

Notes on implementation. Improvements can be made to both of these algorithms as presented here. Conditions are deemed satisfactory when the first homozygous surrogate is found, and this is risky because of pedigree errors, genotyping errors, and errors in inference of IDB. In practice the algorithms should be allowed to continue to collect further information about homozygous surrogates, and to handle ambiguous situations. Both algorithms as presented require at least some surrogates to be shared with at least one parent, to align direction of inheritance across linked loci. However, strategies to group surrogates according to homozygosity status across linked loci provides a route to releasing the need for pedigree to carry out long range phasing. In both cases, individuals that are surrogate on both sides of a proband's pedigree should be detectable by identical genotypes across the region concerned, and these should be eliminated as uninformative.

Hickey *et al.* (2009) provide preliminary results from use of a developed version of the recursive algorithm, based on simulated and real data. They report promising results, both for phasing and for diagnosing map and genotyping errors in high-density SNP data.

REFERENCES

Hickey, J.M., Kinghorn, B. P. and van der Werf, J.H.J. (2009) *Proc. Assoc. Advmt. Anim. Breed. Genet.* **18**:72.

Kong, A., *et al.* (2008) *Nature Genetics* **40**:1068