# PHASING OF SNP DATA BY COMBINED RECURSIVE LONG RANGE PHASING AND LONG RANGE HAPLOTYPE IMPUTATION

**J.M. Hickey[1,2], B.P. Kinghorn[2], B. Tier[3], J.H.J. van der Werf [1,2]**

[1]Cooperative Research Centre for Sheep Industry Innovation, Armidale, NSW 2351
[2]School of Environmental and Rural Science, University of New England, Armidale, NSW, 2351
[3]Animal Breeding and Genetics Unit, University of New England, Armidale, NSW, 2351

## SUMMARY

A Long Range Haplotype Imputation algorithm was developed and combined with a Recursive Long Range Phasing algorithm. These were tested on simulated SNP data sets which had structures possibly similar to those found in sheep and dairy cattle. Performance, measured by accuracy of phasing (>97.6%) and computing time, was competitive in comparison to industry standard software.

## INTRODUCTION

High density SNP arrays can provide detailed information on aspects of the genetic make-up of individuals, enabling whole genome SNP association studies, with possible applications in prediction of genetic value/predisposition in medicine, forensics, and animal and plant breeding. Data from SNP arrays are unphased i.e. paternal and maternal alleles are not determined. However, knowledge of the phase could be advantageous. The high density of SNP arrays allows inference about phase, and relatively efficient algorithms can be found to determine phase.

Local phasing methods like fastPHASE (Scheet and Stephens 2006) exploit strong correlations within LD blocks but they are computationally intensive, and SNPs separated by many LD blocks are not reliably phased (Kong *et al.* 2008). Long range phasing (Kong *et al.* 2008), which is faster and more powerful, was formulated in a recursive algorithm (RLRP) by Kinghorn *et al.* (2009). RLRP gives high computational speed and can lead to inference of phase with little error, for large proportions SNPs. However like LRP it may be unable to phase proportions of SNPs in a data set.

The first objective of this research was to develop a long range haplotype imputation (LRHI) algorithm to impute phase for the proportions of SNP alleles unphased by an RLRP algorithm. The second objective was to test the performance of a combined RLRP-LRHI algorithm for accuracy and computational efficiency under different scenarios possibly existing in some livestock populations.

## MATERIALS AND METHODS

The RLRP-LRHI algorithm outlined here uses a modified version of the RLRP algorithm outlined by Kinghorn *et al.* (2009). It uses cores and spans to move along a chromosome. A core is the string of consecutive SNPs being phased. A span is a string of consecutive SNPs which include the core but may extend outside it. Spans are used to identify surrogates (defined below) at cores. Animals that have phase imputed by LRHI have their cores aligned at the end by re-running the algorithm with the core shifted by *n* SNPs to create overlapping cores for each location. Core and span length as well as tuning parameters, *threshold % of SNPs* and *n SNPs* (these feature below), which help manage genotyping errors present in real data, can be optimised for a given population and data structure.

### Recursive Long Range Phasing Algorithm.

*Step 1 – Identification and partitioning of surrogates.* Surrogates of each animal are defined as animals for which >threshold % of consecutive SNP loci are not of opposite homozygote genotypes, and are partitioned into surrogates of the maternal and paternal haplotypes using a dummy dam and sire. A dummy dam (or sire) is identified by searching elements of the numerator relationship matrix pertaining to the dam (sire) and the sire (dam) of the animal for a surrogate that is related to its dam (sire) but not its sire (dam). Surrogates of the maternal (paternal) haplotype are then identified as the surrogates that are surrogates of the animal and its dummy dam (sire).

*Step 2 – Initial phasing of each SNP.* Each SNP in the maternal (paternal) gamete is attempted to be phased (see Kinghorn et al., 2009 for details). The stopping criteria are (a) a number of surrogates found suggesting that phase is one of the categories and zero conflicting surrogates, (b) if *a* is not satisfied the recursion continues until the number of surrogates suggesting that phase is one of the directions is statistically significantly, and (c) if neither *a* or *b* are satisfied before a maximum Erdös number is reached phase is undeclared. These criteria can be tuned to suit the data.

*Step 3 – Pruning surrogates and re-phasing of each SNP.* The lists of surrogates for each animal are then pruned using the pruning step of Kong et al. (2008). Step 2 is then repeated using the pruned lists of surrogates with slight modification. If phase of the paternal (maternal) allele is unambiguous, phase of the maternal (paternal) allele is imputed as the genotype minus the paternal (maternal) allele.

**Long range haplotype imputation algorithm**

*Step 1 – Building of haplotype library.* Upon completion of the RLRP step a library of long haplotypes for the region being phased is created from completely phased haplotypes of animals.

*Step 2 – Identification of candidate haplotypes.* Candidate haplotypes for an animal with unphased SNPs are identified by comparing its phased SNPs to haplotypes in the library. A candidate haplotype for an animal's gamete is allocated if > threshold % of its SNPs match the gamete's phased SNPs.

*Step 3 – Resolving phase.* If just one candidate haplotype agrees with > threshold % of the animals SNP genotypes, phase is taken to be that haplotype. The alternative gamete has its unphased SNPs imputed as the genotype minus the candidate haplotype and this new haplotype is added to the library. If only two candidate haplotypes are identified and their sum agrees with > threshold % of the animal's SNP genotypes, phase is taken to be these haplotypes. If more than two candidate haplotypes are identified they are paired and if a sum of a single pair agrees with > threshold % of the animal's SNP genotypes, phase is taken to be these haplotypes.

*Step 4 – Incompatibility check.* Each SNP genotype for each animal is compared to the imputed phase. If these did not agree phase is undeclared at this location.

*Step 5 – Iteration.* Steps 1 to 4 were repeated until no new haplotype was added to the library.

**Simulations.** Sheep and dairy cattle populations in equilibrium between mutation, drift, and recombination were simulated using a gene-drop involving a burn in of 9,000 discrete generations of random mating of 500 (50) males and 500 (50) females for sheep (dairy cattle). Each individual's genome consisted of a pair of chromosomes, each 0.1 Morgan long. Mating involved sampling a parent twice without replacement, in such a manner that each animal in the parental generation had two offspring and parents were randomly paired at each mating. For each parent, a gamete was formed by sampling its chromosomes using a recombination rate (0.1 per gamete) and crossover points sampled from a uniform distribution respectively. Base animals had all alleles set to 0. Each gamete had one locus randomly selected as a candidate mutation locus. Mutational events occurred at candidate loci if no segregation at these loci occurred in the population in that generation. 400,000 loci were simulated. In both the sheep and cattle populations two data structures were simulated each with parental and offspring generations of genotyped animals. The offspring generation comprised 1,000 animals. The parental generation comprised 500 dams while one structure had 10 sires and the other had 100 sires. Genotypes of animals were created by continuing the gene drop into these pedigrees. At the end of each simulation 200 and 1,000 SNPs that had a minor allele frequency of >0.1 were randomly chosen for placement on two SNP arrays, reflective of 50,000, and 300,000 SNP array density, and possibly the uneven distribution of SNPs, with regard to physical location, on these arrays.

**Validation methods.** Three replicates were simulated and analysed for each scenario. There were two groups of animals to be phased, parents, who themselves had unknown and ungenotyped parents, and offspring, who had known and genotyped parents. Performance was measured for these groups as % of SNPs correctly phased, % of SNPs incorrectly phased, and % of SNPs not attempted to be phased. For comparison the widely used phasing software, fastPHASE, was used to phase one replicate of the simulated data.

## RESULTS

Mean performance, across the three replicates, for sheep and cattle with sire family size of 10 is given in Table 1. Results for sire family size of 100 were very similar. Variation in performance across replicates was small. Performance for RLRP ranged from 86.88% to 99.73% correctly phased for all SNPs in the offspring. The higher performance was only obtained for dense SNPs in the sheep population. With the exception of where RLRP performed very well, the SNPs that were not correctly phased were mostly not attempted to be phased rather than incorrectly phased. The highest incorrectly phased % was for the sheep population when low SNP density was used (1.40% - 1.92%). For the other scenarios the incorrectly phased % for RLRP was less than 0.24%. The performance of the combined RLRP-LRHI algorithm was encouraging (>97.6%) resulting in major reductions in the not phased % with, at worst, only tiny increases in the incorrectly phased %. A strong relationship between mistakes generated by the RLRP and the LRHI for the offspring animals existed. Performance of RLRP and LRHI increased with increasing SNP density, slightly increased with reducing population size used during the burn in, but was unaffected by family size. Computational time was many times greater for fastPHASE (circa 18 hours) than for RLRP-LRHI (circa 0.33 hours). In the sheep scenario phasing performance of RLRP-LRHI was at worst only very slightly poorer than fastPHASE but in some cases was slightly or considerably better (Table 2). Differences were smaller for cattle as inbreeding was higher.

## DISCUSSION

For this first test of an RLRP-LRHI algorithm performance were encouraging both in terms of computational speed and phasing accuracy. Given the wide range of options implemented in the algorithm greater performance is likely. Any reduction in the incorrectly phased % in RLRP would also improve LRHI. More stringent stopping criteria during the first recursion of RLRP could improve performance. Running the algorithm multiple times using different start and end points for cores and spans, or randomly deleting large percentages of surrogates in a quasi Bootstrap manner could identify SNPs that are prone to being incorrectly phased by RLRP. Currently RLRP and LRHI use the same core length, which is probably suboptimal. Longer cores may be more appropriate for LRHI. An implementation of the RLRP which does not use pedigree information is possible and should improve the performance of the LRHI. The current algorithm is rule based except for a small step where a statistical test is invoked. Incorporation of statistical inference may improve performance especially when genotyping errors are present in the data, particularly for LRHI. A large proportion of the errors from the RLRP-LRHI algorithm occurred where strings of consecutive heterozygous SNPs existed. These result from pairs of very common haplotypes that dominate SNP selection on allele frequency to the extent that loci that are IBS for these haplotypes are not chosen. When these haplotype form a zygote, this lack of IBS yields strings of heterozygotes. This appears to be uncommon in real datasets, but that may partly be due to incorrect map order. Performance differences between RLRP-LRHI and fastPHASE are expected to increase with increasing effective population size. The RLRP-LRHI algorithm outlined here could be used for imputing dense SNP or full sequence data through a pedigree, possibly increasing power of genome wide association/selection studies by reducing allelic heterogeneity and by increasing the numbers of phenotyped animals with genotypes imputed. Performance in human populations, which generally have smaller family sizes, larger effective population sizes, and different relationship structures compared to livestock populations, may be different and will be the subject of a further study.

**Table 1. Percentage of SNPs correctly (Cor.), incorrectly (N cor.), and not (N pha.) phased for parents and offspring in sheep and dairy cattle populations with sire family sizes of 10 offspring using SNP arrays of 50,000 and 300,000 density**

| | | | SNP Density | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | 50,000 | | | 300,000 | | |
| | | | Cor. | N cor. | N pha. | Cor. | N cor. | N pha. |
| Sheep | RLRP alone | Parents | 73.07 | 0.00 | 26.93 | 73.09 | 0.00 | 26.91 |
| | | Offspring | 90.54 | 1.92 | 7.55 | 96.67 | 0.20 | 3.13 |
| | RLRP + LRHI | Parents | 98.08 | 0.09 | 1.83 | 98.91 | 0.22 | 0.87 |
| | | Offspring | 98.34 | 0.97 | 0.68 | 99.50 | 0.32 | 0.18 |
| Cattle | RLRP alone | Parents | 67.93 | 0.00 | 32.07 | 67.85 | 0.00 | 32.15 |
| | | Offspring | 87.00 | 0.10 | 12.90 | 88.93 | 0.07 | 11.01 |
| | RLRP + LRHI | Parents | 97.59 | 1.35 | 1.06 | 98.29 | 0.93 | 0.78 |
| | | Offspring | 99.35 | 0.09 | 0.57 | 99.45 | 0.09 | 0.46 |

**Table 2. Performance of RLRP and LRHI and fastPHASE (Individual error minimised/Switch error minimised) for offspring in replicate 1 of the sheep (family size = 10) scenario.**

| | SNP Density | |
|---|---|---|
| | 50,000 | 300,000 |
| fastPHASE | 98.59 / 97.69 | 93.03 / 96.11 |
| RLRP+LRHI | 98.03 | 99.36 |

## REFERENCES

Kinghorn, B.P., Hickey, J.M., and van der Werf, J.H.J. (2009) *Proc. Assoc. Advmt. Anim. Breed. Genet.* **18**:76.

Scheet, P., and Stephens, M. (2006) *American Journal of Human Genetics.* **78***:*629

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