Recursive Long Range Phasing and Long Haplotype Library Imputation: Building a Global Haplotype Library for Holstein cattle.

J.M. Hickey[12], B.P. Kinghorn[1], M.A. Cleveland[3], B. Tier[4], and J.H.J. van der Werf[1,2]

Introduction

Long range phasing (**LRP**) is a fast and accurate rule based method which uses information from both related and unrelated individuals by invoking the concepts of surrogate parents and Erdös numbers (Kong et al., 2008). Recursive long range phasing and long haplotype imputation (**RLRPLHI;** Hickey et al., 2009) is an extended LRP algorithm with increased robustness partially due to the extra long haplotype imputation step (**LHI**) which is based on a the construction of a library of long haplotypes (e.g. 10cM) for a dataset. The LHI part of the algorithm is computationally much less intensive and less error prone compared to the LRP part, and it can easily incorporate prior information from other phasing methods, both laboratory and *in silico*. After building the initial haplotype library it may be possible to phase other genotyped individuals from the same population simply via the LHI step. The application of such an approach may be useful where small numbers of extra individuals are regularly added to a data set of previously phased individuals, especially where very many animals are genotyped making even a fast phasing method feasible for only small subsets of the data. It may also be useful through incorporation into the LRP part of RLRPLHI in order to improve the accuracy and speed of RLRPLHI. Another application of a long haplotype library could be to use it in combination with segregation analysis (Kerr and Kinghorn, 1996) to impute dense genotype or sequence data in ungenotyped individuals in a sparsely genotyped pedigree. The objectives of this research were to evaluate LHI for phasing genotyped individuals, and segregation analysis and long haplotype imputation (**SALHI**) for imputing genotypes in ungenotyped individuals, using simulated genotypic data in a national and global Holstein pedigree.

Material and methods

RLRPLHI and LHI. RLRPLHI uses surrogate parents at all Erdös levels in the place of true parents to determine phase of a proband (Kong et al., 2008; Hickey et al., 2009). Upon completion of the RLRP step the LHI step builds a library of all completely phased haplotypes for the genome region being phased and then attempts to phase the remaining

[1] School of Environmental and Rural Science, University of New England, Armidale, NSW, 2351, Australia.

[2] Cooperative Research Centre for Sheep Industry Innovation, Armidale, NSW, 2351, Australia.

[3] Genus plc., 100 Bluegrass Commons Blvd., Suite 2200, Hendersonville, TN, 37075, USA.

[4] Animal Breeding and Genetics Unit, University of New England, Armidale NSW 2351, Australia (AGBU is a joint venture between UNE and I&I NSW).

unphased individuals by comparing their genotypes to each haplotype in the library to identify pairs of candidate haplotypes not conflicting with the genotype. When a single pair of candidate haplotypes does not conflict with the genotype these are assumed to be phase. New haplotypes can be added to the library if they can be derived as the complement of the genotype via a single compatible haplotype and the process continues until convergence.

SALHI. SALHI uses segregation analysis (Kerr and Kinghorn, 1996) to determine genotype probabilities for ungenotyped individuals in a pedigree, and RLRPLHI to phase all genotyped individuals in a pedigree and build a haplotype library. Genotypes of ungenotyped individuals are imputed using the genotype probabilities to identify the most probable pair of haplotypes for the individual and if this probability is above a certain threshold this pair of haplotypes is assumed to be the true haplotypes that the individual carries.

Simulations and analysis. A sample of base haplotypes representing a 100 cM region (100,000,000 base pairs) of sequence data was simulated using a per site mutation rate of 1 $\times 10^{-8}$ with MaCS (Chen et al., 2008). The present and historical effective population size was based on the results of Villa-Angulo et al. (2009) for Holstein. Briefly the current Ne was 100, the Ne 1,000 years ago was 1,200, the Ne 10,000 years ago was 4,500, and the Ne 800,000 years ago was 80,000. The simulated haplotypes were dropped through the INTERBULL Holstein pedigree and the pedigree of the genotyped Australian Holstein individuals (**DairyAus**, http://www.adhis.com.au/). The INTERBULL pedigree contained 389,026 individuals of which 4,210 were sires with at least 5 offspring in the pedigree, while the DairyAus pedigree comprised 20,792 indiviudals, and the last 2,000 individuals in the pedigree were taken to represent the Holsteins genotyped in Australia as of June 2010.

To test the performance of LHI three data structure scenarios were created for the INTERBULL pedigree and two for the DairyAus pedigree. For the INTERBULL pedigree each scenario assumed that the 4,210 sires were initially genotyped and phased using RLRPLHI. Scenario 1, 2, and 3 assumed that a random 2,000, a random 10,000, and all 384,816, respectively, of the remaining individuals were subsequently genotyped and phased using LHI based on the RLRPLHI haplotype library for the sires. For the DairyAus pedigree each scenario assumed that the last 2,000 indiviudals were initially genotyped and phased using RLRPLHI. Scenario 1 and 2 assumed that a random 1,000 and all 18,792, respectively, of the remaining individuals were subsequently genotyped and phased using LHI based on the RLRPLHI haplotype library for the last 2,000 individuals. Eight replicates were carried out for two SNP densities (60,000 (**60k**) and 300,000 (**300k**)). Across replicate average % correctly phased and % incorrectly phased were calculated.

To test the performance of SALHI a single replicate of the DairyAus pedigree for 60k density was used. The last 2,000 individuals in the pedigree and 50% of the first 18,792 indiviudals were assumed to be genotyped resulting in 9,396 ungenotyped indivuals. Segregation analysis was then used to obtain genotype probabilites for the first 142 SNPs (7.1cM) on the chromosome. Haplotype libraries of the first 10, 50, and 142 SNP on the chromosome were built based on the true phase of the genotyped individuals and the most probable candidate genotypes were identified. The % correctly imputed and % incorrectly imputed SNPs were then cacluated for the 9,396 ungenotyped individuals using 5 different minumum probability thresholds below which the genotype was undeclared.

Results and discussion

RLRPLHI and LHI. RLRPLHI gave excellent results (>98.88% correctly; <0.23% incorrectly phased) in both pedigrees for both SNP densities (Table 1). LHI was lower in terms of % correctly phased (80.76% to 94.11%) but also gave lower errors (<0.08%). The lower yield of LHI compared to RLRPLHI was consistent with expectation as LHI only uses information on an individual's homozygous loci to identify candidate haplotypes whereas RLRPLHI uses homozygous loci as well as heterozygous loci that it can phase. Performance of LHI was better: for the 300k data compared to the 60k data; in the DairyAus pedigree compared to the INTERBULL pedigree; and for the subsets of each pedigree compared to when all individuals in the pedigree were to be phased because combinatorial power is higher for each of these scenarios. LHI fails when it finds more than one pair of compatible haplotypes and thus cannot determine phase. This is less likely with the 300k data and in the subsets where there is less chance for meiotic events to create large numbers of different yet similar haplotypes. In addition both the DairyAus and the INTERBULL pedigrees were given the same base Ne in the simulation despite the fact that the INTERBULL pedigree was deeper. In reality the base Ne in the INTERBULL pedigree is likely to be larger than 100. Reducing the numbers of candidate haplotypes would help to prevent LHI finding more than one pair of candidate haplotypes and could be achieved using genealogy, segregation analysis, information from partially overlapping haplotypes, or denser genotype data. Because of their high accuracy the results for LHI are useful even when yield is low. Even a relatively fast phasing algorithm such as RLRPLHI is infeasible for large data sets (e.g. 300,000 individuals with 1 million SNPs). In RLRPLHI the identification and partitioning of surrogates into paternal/maternal surrogates is a computational bottleneck and because it only uses homozygous loci for this it is underpowered in comparison to a method which uses heterozygous loci as well. By first using LHI on a data set the computational requirements to identify and partition surrogates reduces to a single test (based on haplotypes) in the case of two individuals which have been phased by LHI and makes use of information at heterozygous loci where either one of a pair of individuals has been phased by LHI.

Table 1. Performance of RLRPLHI and LHI, in terms of % correctly phased (%Cor) and % incorrectly phased (%Incor), for the different data scenarios and SNP densities

		60k		300k	
		%Cor	%Incor	%Cor	%Incor
INTERBULL	4,210 sires phased by RLRPLHI	99.24	0.18	98.88	0.23
	2,000 random phased by LHI	87.10	0.04	90.74	0.02
	10,000 random phased by LHI	87.50	0.05	90.57	0.02
	All 384,816 other phased by LHI	80.76	0.08	85.75	0.03
DairyAus	Last 2,000 phased by RLRPLHI	99.11	0.21	98.99	0.19
	1,000 random phased by LHI	91.42	0.06	94.11	0.06
	All 18,792 other phased y LHI	91.30	0.05	94.09	0.02

Table 2. Performance of SALHI[1], in terms of % SNPs correctly inferred (%Cor) and % SNPs incorrectly inferred (%Incor), when different lengths of haplotypes and different minimum probability thresholds (Min. ∑Log(Pr)) are used

	142 SNP		50 SNP		10 SNP	
Min. ∑Log(Pr)	%Cor	%Incor	%Cor	%Incor	%Cor	%Incor
Infinity	74.39	25.61	80.23	19.77	85.11	14.89
-30	2.95	0.01	25.44	2.20	85.10	14.88
-25	1.98	0.00	9.15	0.20	76.55	9.55
-22	1.32	zero	4.59	0.02	50.99	2.44
-21	1.00	zero	3.19	0.00	19.29	0.07

[1] Pedigree of 20,792 individuals (11,396 were genotyped and 9,396 were ungenotyped)

SALHI. The results for SALHI (Table 2.) suggest that it is worth pursuing for genotype or sequence imputation. Depending on the minimum probability threshold used up to 85% of genotypes were correctly imputed. Where the minimum probability threshold was restricted so that <0.1% SNPs were imputed with error, 19.29% of SNPs were correctly imputed in the 9,396 ungenotyped individuals. Shorter haplotypes performed better because they gave fewer candidate genotypes. This is a first attempt at SALHI and there are several ways in which it could be improved including application of an iterative approach where in repeated round of segregation analysis genotype information imputed in used that was imputed in previous iterations; optimization of genotyping strategies to identify key individuals to sparsely/densely genotype; and the elimination of candidate haplotypes using genealogy.

This research suggests that a global haplotype library for livestock populations may be useful. Combining it with LHI and SALHI could speed up and increase accuracy of the phasing of genotyped individuals, make phasing of very large data sets feasible, and aid imputation of genotype or sequence data. However all of this is predicated on there being a high quality physical map which persists in all families in the population.

Acknowledgments

Jette Jacobsen is thanked for providing the INTERBULL pedigree.

References

Chen, G.K., Marjoram, P., and Wall, J.D. (2009). *Genome Res.*, 18: 136 – 142.

Hickey, J.M., Kinghorn, B.P., Tier, B. *et al.* (2009). *Proc. Assoc. Advmt. Anim. Breed. Genet.*, 18: 72-75.

Kong, A., Masson, G., Frigge, M.L. *et al.* (2008). *Nature Genetics*, 40:1068 – 1075.

Kerr, R.D. and Kinghorn, B.P. (1996). *J. Anim. Breed. Genet.*, 113: 457-469.

Villa-Angulo, R., Matukumalli, L.K., Gill, C.A. *et al.* (2009). *BMC Genetics*, 10:19.