Proc. Assoc. Advmt. Anim. Breed. Genet. 20:546-549

EFFECT OF GENOTYPE AND PEDIGREE ERROR ON DETECTION OF RECOMBINATION EVENTS, SIRE IMPUTATION AND HAPLOTYPE INFERENCE USING THE HSPHASE ALGORITHM

Mohammad H. Ferdosi[1], Brian P. Kinghorn[1], Julius H.J. van der Werf[1, 2] and Cedric Gondro[1]

[1]School of Environmental and Rural Science, University of New England, Armidale, Australia
[2]Cooperative Research Centre for Sheep Industry Innovation, Armidale, NSW 2351, Australia.

SUMMARY

HSPhase is a fast and accurate algorithm for detection of recombination events, sire imputation and haplotype inference of half-sib families. It can be used on data for half-sib families with as few as 4 individuals in a family. The robustness of this algorithm in relation to genotype and pedigree errors was evaluated. If there were more than 20 half-sibs in a family, the performance of the algorithm with 5% pedigree or genotyping errors was still reliable with the accuracy of phasing and imputation above 0.87. These error rates are above those commonly observed in industry data which indicates the algorithm is sufficiently robust for deployment in real world settings. An R package implementing the method is freely available and includes a function to generate diagnostic plots which are very useful to rapidly identify problems in the dataset.

INTRODUCTION

The availability of genotype information on large numbers of dense molecular markers (usually Single Nucleotide Polymorphisms - SNPs) or even full genome sequences has provided interesting challenges with regard to the best use of all available information. One way to start addressing this is by haplotype reconstruction. Although with current technology it is possible to generate phased data directly, it is still expensive and not suitable for routine usage (Browning *et al.* 2011). Alternatively, computational methods can be used to reconstruct haplotypes from genotypes. The most common approaches make use of population wide data and use a hidden Markov model, e.g. as implemented in BEAGLE (Browning *et al.* 2011). These methods and algorithms were mainly developed for the human population structure and few algorithms have been developed specifically for livestock populations that consist of complex pedigree and for which large half-sib families are usually available (Hickey *et al.* 2011*;* Boettcher *et al.* 2004*).*

We developed the HSPhase algorithm to create block structures of haplotype relationships, which are then used to impute/phase sire and phase genotype data specifically for half-sib family groups (Ferdosi *et al.* 2013). In the real world, data on pedigree and genotypes contain errors which could well affect the performance of phasing methods. In this paper we evaluate the robustness of HSPhase to genotype and pedigree errors.

METHODS

The HSPhase algorithm uses opposing homozygotes to create a block structure and finds a parental origin for each SNP allele; therefore, for each individual the haplotype of the strand that belongs to the sire becomes evident (Figure 1-A, 1-B and 1-C). As the parental origin at multiple SNPs becomes evident, sire haplotypes can be imputed by using the block structure and by calculating the average of SNPs that belong to the first or second sire haplotype (Figure 1-D). The haplotype can simply be reconstructed by replacing the haplotype of the sire with the corresponding block in the half-sib structure (Ferdosi *et al.* 2013).

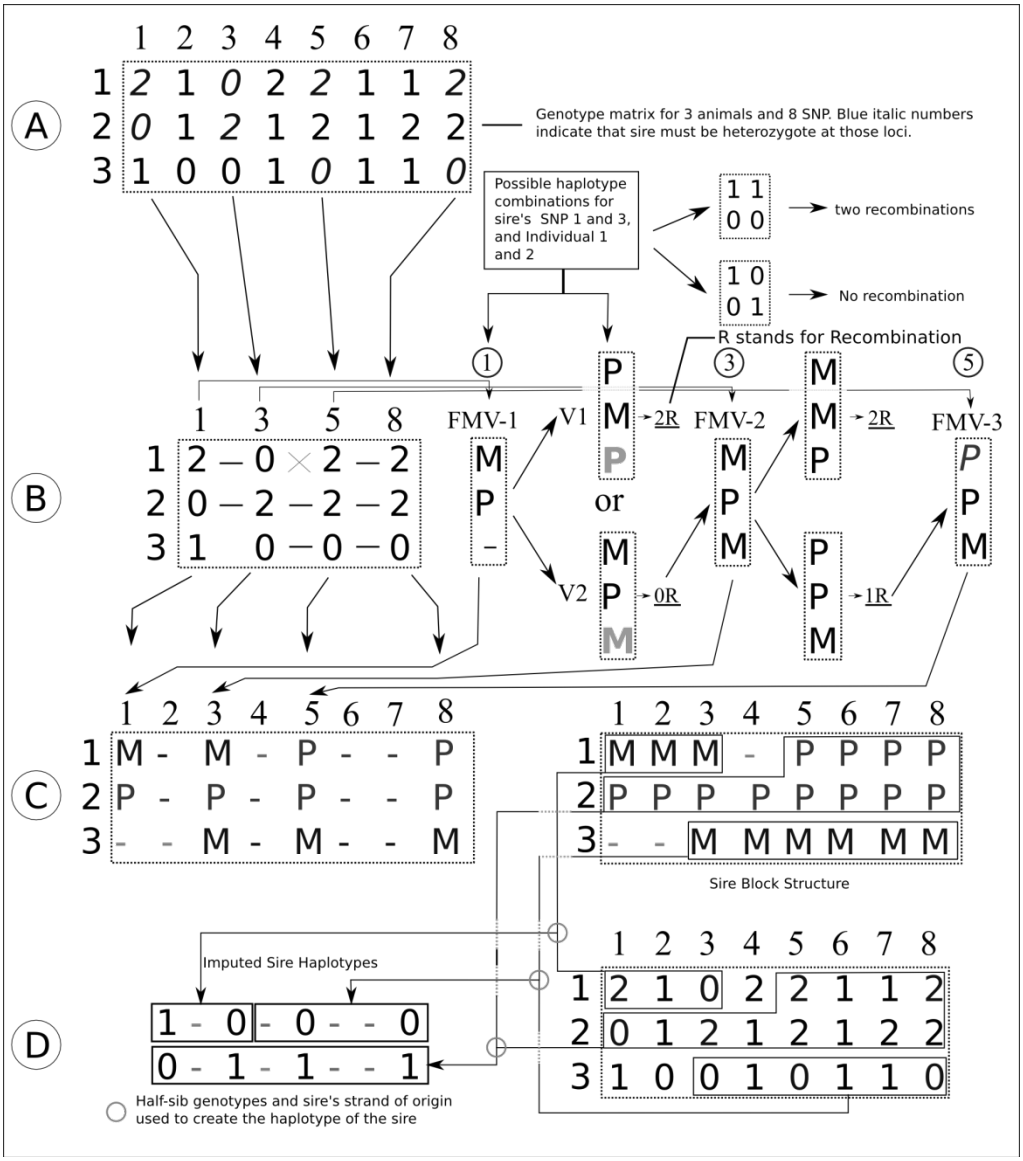

Figure 1. A: Selection of heterozygote loci in sire based on the genotype of the offspring, B: detection of the parental strand of origin of each offspring, C: Filling the gap between different heterozygote, if two loci in one individual have the same parental origin, the SNPs between them are deemed to come from the same haplotype, D: By using the strand origin and genotype the sire's haplotype becomes evident.

The QMsim (Sargolzaei *et al*. 2009) program, which simulates genotype data based on the population structure found in commercial livestock, was used to generate a dataset. A single chromosome of 500 cM in length and 10,000 markers was simulated. For each dataset 20 males were mated to 400 females each generation and genotypes for the last 7 generations were recorded. Each of the final datasets consisted of 120 half-sib groups with 40 offspring in each. Smaller

half-sib families (4, 6, 8, 10, 20 and 40) are sampled randomly from this population by using *sample* function in R (R Core Team 2013).

To evaluate the accuracy of the algorithm with varying rates of genotyping errors, a random proportion of SNP errors were added to the genotype by using the *sample* function. The accuracy of the algorithm was evaluated for 1, 5 and 10 percent genotyping errors. To evaluate the effect of pedigree errors, an individual in the half-sib family was replaced with a random genotype; therefore, we had a pedigree error for different size of half-sib groups and the critical number of offspring that require to handle one pedigree error was estimated.

The accuracy of the method was tested as the squared correlation (R^2) between true and inferred results using the *lm* function in R (R Core Team 2013) using different numbers of half sibs in each family group.

RESULTS AND DISCUSSION

R^2 between true and detected blocks and haploypes were calculated to evaluate the effect of pedigree and genotype errors on the HSPhase algorithm (Tables 1 and 2).

Pedigree errors had a negative effect on the accuracy of the method when less than 10 half-sibs were available. This was mainly driven by genotypes that did not belong to the half-sib group, incorrectly suggesting a heterozygous site in the sire. And also, when the number of half-sibs in the group is limited, wrong genotypes will have a more significant effect on the detection of recombination. As the number of individuals in the half-sib group increases the effect of an incorrect genotype decreases and with 10 half-sibs R^2 values were generally high. Results also show that generally with more individuals per family the accuracy increased (Table 1).

Table 1. Effect of pedigree errors on accuracy (R^2 +/- standard deviation between inferred and true results) and call rate (percentage of known results) for various HS family sizes

	4	6	8	10	20	40
BP	0.58±0.38	0.77±0.32	0.90±0.19	0.93±0.10	0.96±0.03	0.96±0.02
PB%	0.95±0.13	0.98±0.02	0.99±0.02	0.99±0.02	0.99±0.02	0.99±0.01
BPE	0.09±0.19	0.18±0.20	0.46±0.29	0.65±0.24	0.85±0.04	0.90±0.02
BPE %	0.99±0.02	0.98±0.01	0.99±0.01	0.98±0.02	0.99±0.02	0.99±0.01
SI	0.75±0.25	0.87±0.20	0.95±0.11	0.97±0.06	1.00±0.01	1.00±0.00
SI%	0.50±0.09	0.69±0.06	0.80±0.06	0.88±0.06	0.99±0.01	1.00±0.00
SIPE	0.43±0.16	0.52±0.18	0.71±0.21	0.85±0.17	0.98±0.02	1.00±0.00
SIPE%	0.54±0.04	0.71±0.05	0.82±0.04	0.89±0.03	0.99±0.01	1.00±0.00
HI	0.96±0.13	0.97±0.03	0.98±0.02	0.98±0.02	0.98±0.01	0.99±0.01
HI%	0.67±0.09	0.79±0.04	0.87±0.03	0.92±0.03	0.99±0.01	1.00±0.00
HIPE	0.45±0.06	0.55±0.04	0.65±0.03	0.72±0.02	0.88±0.01	0.94±0.01
HIPE%	0.75±0.03	0.84±0.02	0.90±0.02	0.94±0.02	0.99±0.00	1.00±0.00

BP: Block Partitioning, PB%: Percent of Known Block, BPE: Block Partitioning with 1 pedigree error in the half-sib family, BPE %: Percent of Known Block Partitioning with 1 pedigree error in the half-sib family, SI: Sire Imputation, SI%: Percent of Known Sire, SIPE: Sire Imputation with 1 pedigree error in the half-sib family, SIPE%: Percent of Known Sire with 1 pedigree error in the half-sib family, HI: Haplotype Inference, HI%: Percent of Known Haplotype, HIPE: Haplotype Inference with 1 pedigree error in the half-sib family, HIPE%: Percent of Known Haplotype with 1 pedigree error in the half-sib family.

With sufficient half-sibs, pedigree errors can be easily detected by counting the number of recombination events. Figure 2 shows the image plot for a half-sib group. The second individual does not belong to this family due to excessive recombinations. This is an easy way to identify pedigree errors. Since the algorithm can phase the sire of half-sib groups with high accuracy, by counting the number of opposing homozygotes between this individual and other sires, it is easy to

detect the half-sib family that this individual belongs to, provided that sufficient SNPs are available (Figure 2).

The haplotype accuracy of HSPhase did not markedly decrease with genotype errors of up to 5 percent, provided that more than 10 half-sibs were available. A higher error rate significantly decreased accuracy across all family sizes, but especially when less than 20 half-sibs were available (Table 2). These results show that the algorithm is robust to both genotyping and pedigree errors beyond the levels commonly observed in livestock data. This makes it suitable for routine adoption by applications that require phasing and/or imputation.

Table 2. Effect of genotype errors on accuracy (R^2 +/- standard deviation between inferred and true results) and call rate (percentage of known results) for various HS family sizes

	4	6	8	10	20	40
BPE1%	0.50±0.38	0.70±0.34	0.86±0.23	0.91±0.11	0.95±0.03	0.95±0.02
BPE5%	0.40±0.36	0.52±0.38	0.73±0.33	0.86±0.18	0.92±0.10	0.93±0.03
BPE10%	0.28±0.29	0.36±0.31	0.60±0.36	0.71±0.31	0.88±0.12	0.88±0.08
SIPE1%	0.71±0.23	0.82±0.21	0.93±0.13	0.96±0.06	0.99±0.01	1.00±0.00
SIPE5%	0.61±0.22	0.68±0.22	0.81±0.18	0.90±0.09	0.97±0.04	1.00±0.00
SIPE10%	0.50±0.17	0.55±0.19	0.70±0.20	0.79±0.16	0.95±0.06	0.99±0.05
HIPE1%	0.95±0.38	0.94±0.34	0.94±0.23	0.95±0.11	0.96±0.03	0.96±0.02
HIPE5%	0.85±0.36	0.84±0.38	0.84±0.33	0.85±0.18	0.87±0.10	0.88±0.03
HIPE10%	0.75±0.29	0.73±0.31	0.73±0.36	0.74±0.31	0.77±0.12	0.78±0.08

BPE1%, BPE5%, BPE10%: Block Partitioning with 1%, 5% and 10% Genotyping Errors, SIPE1%, SIPE5%, SIPE10%: Sire Imputation with 1%, 5% and 10% Genotyping Errors, HIPE1%, HIPE5%, HIPE10%: Haplotype Inference with 1%, 5% and 10% Genotyping Errors

Figure 2. Block structure in a half-sib family of chromosome 1 using real data from Hanwoo cattle with 11 individuals (dark and light gray are for the first and second haplotype of the sire; markers of unknown origin are shown in white). The second individual is not related to this sire given the large number of recombination events observed.

ACKNOWLEDGEMENTS

This study was funded by a grant from the Next-Generation BioGreen 21 Program (No. PJ008196), Rural Development Administration, Republic of Korea

REFERENCES

Boettcher P.J., Pagnacco G., and Stella A. (2004) *J. of Dairy Sci.* **87:** 4303.
Sargolzaei M. and Schenkel F.S. (2009) *Bioinformatics* **25**: 680.
Browning S.R. and Browning B.L. (2011) *Nature Reviews Genetics* **12**: 703.
Hickey J., Kinghorn B.P., Tier B., Wilson J.F., Dunstan N., and Van der Werf J.H.J (2011) *Genetics Selection Evolution* **43:** 12.
Ferdosi M.H., BP Kinghorn, JHJ van der Werf and C Gondro (2013) (under review)
R Core Team (2013). *R Foundation for Statistical Computing*, Vienna, Austria