A Combined Strategy To Infer High-Density SNP Haplotypes In Large Pedigrees

D. Habier*, *R. L. Fernando** and D. J. Garrick*

Introduction

Ordered genotypes, haplotypes and grand-parental allele origins of high-density single nucleotide polymorphisms (SNPs) may become important sources of information in animal and plant breeding for genomic selection (GS). So far, unordered SNP genotypes have been used, but haplotypes can provide additional information to model linkage disequilibrium (LD), while grand-parental origins can be used to model cosegregation. Furthermore, the low-density GS approach described by Habier et al. (2009) requires ordered high-density SNP genotypes of ancestors to track alleles at ungenotyped loci through the pedigree using information from evenly spaced low-density SNPs.

Pedigrees with a large number of individuals genotyped for high-density SNPs are becoming available, but existing software packages such as *Phase*, *fastPhase* or *Merlin* are not suitable for pedigrees with several thousand individuals and do not utilize information from both LD and cosegregation. Peeling and reverse peeling (Heath, 1997) using a Gibbs sampler with overlapping blocks (Thomas et al., 2000) is also computationally demanding because of the large cutset sizes. However, this can be avoided by reducing the state space of ordered genotypes and grand-parental allele origins prior to sampling. Rule-based methods can be used for this purpose such as simple genotype and origin elimination as well as the rules described by Wang et al. (2007) or Daetwyler (2009).

The objective of this study was to present and test a strategy for estimating probabilities of ordered genotypes and grand-parental allele origins of high-density SNPs in large complex pedigrees by combining a rule-based method with a Gibbs sampler with overlapping blocks.

Material and methods

Rule-based approach. The rule-based approach requires at least three generations of genotyped individuals and can be explained with the following example. Let's assume that high-density SNP genotypes are available for an individual, its father and grandfather. Then at every locus at which the father is heterozygous and the other two individuals are homozygous, the grand-parental origin of the paternal allele of the individual can be inferred as follows: As the grandfather is homozygous, the observed heterozygous genotype of the father can be ordered, i.e. it is known which one of the two possible allele states was received from the grandfather.

*Department of Animal Science and Center for Integrated Animal Genomics, Iowa State University, Ames, IA 50011, USA

Furthermore, the homozygous genotype of the individual is already ordered, because both alleles have the same state. Thus, if the individual received the very same allele state as the father from the grandfather, then the origin of the paternal allele in the individual is grand-paternal. The grand-parental origin can be resolved for many SNPs along the genome leaving chromosomal segments that are smaller than 1 cM and flanked by SNPs with known grand-parental origin. In addition, genotypes of both the father and the individual can be ordered at many other SNPs either based on the genotype pairs of grandfather-father and father-individual, respectively, or because they are homozygous. The state spaces of ordered genotypes and grand-parental origins can be further reduced by applying the following rules: If the paternal alleles of the SNPs flanking a chromosomal segment have the same grand-parental origin and these two SNPs are so close that a crossover between them is very unlikely, then the paternal alleles of all SNPs on that chromosomal segment obtain the grand-parental origin state of the two flanking SNP alleles. With this information, the genotype order can be resolved for the father if the genotype order is known for the individual, and vice versa. This completes one iteration of the algorithm, and a new one starts by inferring unknown grand-parental origins using the currently resolved ordered genotypes determined by applying the rules in the previous iteration. The algorithm stops as soon as no additional grand-parental origins can be determined. The same principles hold when mothers are also genotyped. The maximum length of a chromosomal segment to be resolved is determined by the maximum probability of a double crossover on that segment, which was 0.0001 in this study.

Gibbs sampling with overlapping blocks. The rule-based method cannot reduce the state space of heterozygous founder genotypes nor of grand-parental origins of the offspring of founders. Moreover, the rule-step is not applied to chromosomal segments for which the grand-parental origins at flanking SNPs are different or which are too long (e.g. > 2 cM) such that a crossover becomes more likely. Therefore, the remaining states are sampled by a Gibbs sampler with overlapping blocks (Thomas et al., 2000) using information from both LD and cosegregation. The strategy with overlapping blocks refers to pedigree and SNPs and is to reduce cutset sizes during peeling and reverse peeling (Heath, 1997). Pedigree blocking is only used for the first two generations because of the unknown order of heterozygous founder genotypes and the unknown grand-parental origins of the second generation. Thus, each pedigree block contains only a male founder, its mates and offspring. In later generations, only a few chromosomal segments with undetermined grand-parental origins and even fewer unresolved ordered genotypes are left for each individual. Those segments are independent from the rest of the genome given the grand-parental origins at the flanking SNPs, and independent from the genotypes of the ancestors if the genotypes of the parents are ordered. As a result, cutset sizes, mixing problems and computing time are substantially reduced. Furthermore, the strategy with overlapping blocks allows parallel computing with multiple processors. The implementation used here distributes locus blocks to different processors with overlapping loci across processors.

Simulations. Two types of simulations were conducted. Scenario 1 was simulated to quantify the fraction of ordered genotypes and grand-parental origins that can be determined by the rule-based approach depending on whether dams were genotyped in addition to sires and their offspring. Three discrete generations were simulated each having 100 sires that were mated to either 1, 2, 3, 4 or 10 dams, where each dam was mated to only one male and each

mating produced 1 male and 1 female offspring. Sires and dams of the next generation were randomly selected from the offspring. Pedigrees with and without females were analyzed in order to investigate situations in which only the sires are genotyped such as in dairy cattle. The simulated genome consisted of 2,000 SNPs that were randomly distributed on a chromosome of length 1 M. Initially, the SNPs were in linkage equilibrium and Hardy-Weinberg equilibrium and the allele frequency was 0.5, which also applies to the following scenario. Scenario 2 was simulated to demonstrate the feasibility of Gibbs sampling with overlapping blocks in large pedigrees. Genotypes of both 8,239 North-American Holstein bulls that are genotyped for high-density SNPs in practice and their ancestors born after 1950 were simulated, but only the genotypes of the 8,239 bulls were used in the analysis to represent the current situation in dairy cattle. The simulated genome consisted of 100 SNPs on a chromosome of 5 cM. Four 2.4 GHz AMD 280 Opteron processors were used for sampling, where the block size was 8 loci with 3 overlapping loci. The sampler was run for 1,000 iterations with a burn in of 100 iterations.

Results and discussion

Scenario 1. Table 1 shows the proportion of ordered genotypes determined by the rule-based approach for individuals in the second and third generation in scenario 1. When both parents of an individual in generation 3 were genotyped, 99% of both the sire's and the offspring's genotypes could be ordered from only 1 genotyped offspring per sire. With two offspring per sire that proportion reached 99.9%. When only sires were genotyped, the increase with number of offspring per sire was larger from 87% with 1 offspring per sire to 99.1% with 10 offspring per sire. In general, the proportion of ordered genotypes was slightly higher for sires than for offspring when more than 1 offspring were genotyped.

Table 1: Proportion of ordered genotypes of sires and their offspring in a three generation pedigree (founder, parent, offspring) resolved by the rule-based approach according to the number of offspring per sire and whether dams were genotyped

No. of	Sire genotypes		Offspring genotypes	
offspring	Sire & Dam	Sire only	Sire & Dam	Sire only
1	99.0	87.0	99.0	87.0
2	99.9	93.4	99.9	92.6
3	99.9	96.7	99.9	96.4
4	99.9	98.3	99.9	98.0
10	99.9	99.7	99.9	99.1

Table 2 shows the proportion of grand-parental origins of the third generation determined by the rule-based method. That proportion was 99% when both parents were genotyped and 96% when only sires were genotyped. This difference is due to the fact that the chromosomal segments with known grand-parental origins at the flanking SNPs are smaller when both parents are genotyped compared to sires only. The reason is that more grand-parental origins can be determined from the trio genotypes of sire, dam and offspring. The proportion of resolved

grand-parental origins was not sensitive to the number of offspring per sire, because the remaining grand-parental origins were on chromosomal segments which were either flanked by alleles with different grand-parental origins or were longer than 2 cM. The proportion of a false determination of genotype order and grand-parental origin was below 10^{-5} and 5×10^{-5}, respectively. Those errors occurred when there was a double crossover on a chromosomal segment. The error rate can be reduced by shorten the length of the chromosomal segments that are resolved, but with the result that fewer ordered genotypes and grand parental origins can be determined. The increasing SNP density in the future, however, will shorten the chromosomal segments that are flanked by SNPs with known grand-parental origins, and thus more variables can be determined by applying the rules with fewer errors.

Table 2: Proportion of grand-parental allele origins of individuals in the third generation of scenario 1 resolved by the rule-based method according to the number of offspring per sire and whether dams were genotyped

No. of offspring	Sire & Dam	Sire only
1	98.9	96.3
2	99.8	95.6
3	99.8	96.4
4	99.9	96.5
10	99.8	95.5

Scenario 2. The proportion of ordered genotypes and grand-parental origins of the 7,573 bulls with genotyped paternal grandfathers was 97% and 87%, respectively, while 98% of the genotypes of their fathers were ordered. The computing time for sampling was 1 hour and 55 minutes, where 1 hour and 13 minutes was required to setup the cutsets needed for peeling and reverse peeling. The 1,000 iterations of sampling took 42 minutes.

Conclusion

The rule-based approach reduces the state space of ordered genotypes and grand-parental origins such that sampling of the remaining ambiguous states is feasible, while the fraction of allele states and origins that are falsely determined by the rules is limited. This will also be true with increasing SNP density, because more states can be inferred by the rule-step as crossovers are less likely to occur on shorter chromosomal segments.

References

Daetwyler, H. D. (2009). Ph.D. thesis.

Habier, D., Fernando, R. L., and Dekkers, J. C. M. (2009). *Genetics*, 182(1):343–353.

Heath, S. C. (1997). *Am. J. Hum. Genet.*, 61:748–760.

Thomas, A., Gutin, A., Abkevich, V., et al. (2000). *Stat. and Comp.*, 10:259–269.

Wang, C., Wang, Z., Qiu, X., et al. (2007). *Chi. Sci. Bull.*, 52:471–476.