Maternal Grandsire Verification and Detection without Imputation

***J.B.C.H.M. van Kaam*[1] *and B.J. Hayes*[2,3]**
[1]*Associazone Nazionale Allevatori Frisona Italiana, Cremona, Italy*
[2]*Biosciences Research Division, Department of Primary Industries, Bundoora, Victoria 3083, Australia*
[3]*Dairy Futures Cooperative Research Centre, Victoria, Australia*

Abstract

The rapidly increasing availability of SNP genotype information enables pedigree verification at a deeper level than before. Here a method for verification and detection of maternal grandsires (MGS), for animals without dam genotypes, is described which we name *long haplotype MGS detection*. This new method facilitates an easier workflow by avoiding phasing of maternal grandsires and imputation resulting in considerable time-savings. It is shown to be over 99% in agreement with maternal grandsires obtained using imputation even when most animals have low density genotypes and no imputation was used. Genomic information not only enables genomic selection but also helps in improving selection and traditional breeding value estimation and in avoiding inbreeding by enabling sampling and pedigree errors to be tracked and corrected.

Key words: SNP, grandsire verification, grandsire detection, pedigree, relationship

Introduction

Large numbers of dairy cattle have now been genotyped with SNP arrays (“chips”) ranging in density from 3K to 800K. In dairy cattle the number of male ancestors is much smaller than the number of female ancestors and a much larger percentage of them has been genotyped. Therefore genotyped animals often have genotyped sires and grandsires, however their dams are usually not genotyped. Pedigree verification and detection in the female line therefore is more complicated than in the male line. The SNP genotype information enables verification and detection not only of direct parents but also of grandparents, even without genotype of the intermediate animal.

Here four methods will be mentioned, each of which can be used for maternal grandsire (MGS) verification as well as detection. Each of these methods is aimed at the situation that dam genotypes are not available. Two methods were named duo and trio by VanRaden *et al.* (2013) and are based on counting conflicts of single markers. The duo method simply counts the number of SNPs which are opposite homozygotes in the MGS and the animal and uses this as test statistic. No sire or dam genotype is used. The trio method is an extension of the duo method, which in addition uses sire genotype information, thereby enabling the use of more SNPs. If the animal is heterozygous, and the sire is homozygous for a given allele, then the other allele of the animal was contributed by the dam and should be present in the genotype of a candidate MGS. If none of the MGS alleles corresponds with the dam allele then this SNP conflict will be counted in the test statistic. The inclusion of extra SNP information in the test statistic leads to more accurate results of the trio method compared to the duo method.

A third method for detecting a MGS is to count the number of haplotypes of a given length in common between the animal’s maternal haplotype and the haplotypes of each candidate MGS (VanRaden *et al.*, 2013) and use this as test statistic. Haplotype methods use the joint inheritance, identity-by-descent (IBD), of multiple markers and therefore have more information and should be more powerful than the duo and trio methods. The duo and trio methods can be performed as soon as a genotype is received because no phasing and imputation are required. If chips of different density are used then with the third method phasing and imputation will be needed. The need for phasing and imputation is a serious disadvantage because it requires more computation time and consequently interrupts the workflow. The fourth method, which we name *long haplotype MGS detection*, is presented here and can be seen as a variation on the third method. *Instead of using more markers in the same (short) segment obtained by imputation an alternative is to use much longer segments without phasing the MGSs and without imputation*. This will be discussed here and has been implemented at Anafi.

Material and Methods

All genotypes are converted to a standard SNP panel (currently the Illumina 50K version 1 chip), beforehand. SNPs which are not genotyped are set to missing. Therefore all samples have the same number of SNP genotype scores.

In practice our MGS method has two phases:

1. *Verification*: For animals with a genotyped MGS in the pedigree, verify if their pedigree MGS meets the criteria (number of matching haplotypes) to be trusted.
2. *Detection*: For those animals where the pedigree nominated MGS does not meet the criteria or without genotyped pedigree MGS create an ordered list of most likely candidate MGSs.

For each animal for which the MGS has to be verified or detected, the *long haplotype MGS* algorithm has the following steps:

1. Derive the maternal haplotype for the animal of interest, *for loci where this can be done unambiguously*. That is, where the animal is homozygous, or where the animal is heterozygous and the sire is homozygous like in the trio method.
2. Next loop through all suitable candidate MGSs (defined by the rules below), and count the number of matching non-overlapping haplotype segments of length *x* (defined by the number of SNPs) on all autosomes. This is used as the test statistic.

Within a segment, the animals' maternal haplotype is considered to match the haplotype of the MGS if the MGS does have very few genotypes conflicting with the animals' maternal allele. If animal and MGS were both genotyped at high density then up to 2 conflicting SNPs were allowed before a segment was considered unequal in order to account for possible genotype errors. Any matching non-overlapping segment of sufficient length was counted in the test statistic. If more than the permitted number of conflicting SNPs were found in a segment, the SNP counting was reset to the SNP after the first conflicting SNP and the counting was continued from the point where it had arrived. Segments counted have the same length for all SNP chips, because all SNP genotypes were converted to the same SNP panel beforehand.

In order to implement the method, decisions must be made on the length of segments, and the number of matching segments that are required to decide whether the MGS is wrong or in doubt. In the second phase the test statistic is used to order the most likely candidate MGSs. Simulation of crossovers during gamete creation showed that parents have 25.6 matching segments of half an autosome each, grandparents 20.0 and great-grandparents 14.8. Haplotype IBD segments get shorter when more meioses have taken place, hence each generation the IBD segments get shorter. Actual data with the MGSs verified showed that pedigree MGS on average had 20.8 matching segments. In practice MGS with ≤ 14 matches were considered in doubt and with ≤ 6 as likely wrong. In regular analyses both these categories are included in candidate MGS detection to find possibly better MGS. Candidate MGSs need to have > 12 matches to be considered for the most likely candidate list.

The selection of suitable candidate MGS uses some rules:

- Only males can be MGS
- Animals cannot be their own MGS
- Pedigree sire of the MGS is excluded as candidate MGS
- Pedigree sire of the maternal granddam (MGD) is excluded as candidate MGS
- Pedigree maternal granddam's paternal sibs are excluded as candidate MGS
- Pedigree maternal granddam's maternal sibs are excluded as candidate MGS
- Pedigree maternal granddam's sons are excluded as candidate MGS

- Minimum 1-generation interval > 600 days between animal and dam and between dam and MGS
- Minimum 2-generation interval > 1200 days between animal and MGS in case the dam's birthdate is unknown

Our application includes various computational aspects in order to make the software fast:

- Fortran 2008 code
- Parallel processing using OpenMP
- Compilation optimization
- Use the inner-loop index as first index in the genotype array, because fortran uses column-major storage. This way SNPs are processed in the order in which they are stored.
- Avoid recounting (parts of) any haplotype segment.
- Stop comparing an autosome as soon as there is no longer enough remaining length to obtain a matching segment.
- Use half an autosome plus 1 SNP (in scenario 1 of Table 1) as the required segment length to count, because in this manner only 1 matching segment per autosome can be found, after which the SNP comparison can be stopped as indicated in the previous point.
- Stop comparing a candidate MGS if there is not enough space left on the remaining autosomes to reach the cutoff of > 12 matching segments for signaling a candidate MGS.
- Work in reverse order from the higher number autosomes to the lower number autosomes in order to first do smaller autosomes. This together with the previous point means that often the largest autosomes are not needed to compare. In practice, in scenario 1 of Table 1, if none of the first 17 autosomes has a matching segment then the remaining 12 autosomes do not need to be compared because they cannot result in exceeding the cutoff of > 12 matching segments.

Table 1. Effect of haplotype length on MGS detection.

	Haplotype segment length IBD[a]	Pedigree MGS as 1st candidate	Pedigree MGS in 4 candidate list	Elapsed time (in min)
1	0.50 chromosome+1 SNP, $\bar{x}$ = 716 SNPs	99.55%	100.00%	134
2	0.50 chromosome, $\bar{x}$ = 715 SNPs	99.64%	99.98%	194
3	0.33 chromosome, $\bar{x}$ = 477 SNPs	99.55%	99.73%	216
4	0.25 chromosome, $\bar{x}$ = 357 SNPs	99.61%	99.71%	224
5	0.10 chromosome, $\bar{x}$ = 143 SNPs	99.38%	99.68%	251
6	500 SNPs	99.50%	99.73%	242
7	75 SNPs	99.04%	99.59%	293
8	1 SNP (Trio method)	85.64%	89.20%	554

[a] Haplotype length expressed in number of SNPs on the Illumina 50K version 1 chip.

Results

In Table 1 results are shown for 5 600 non-Italian genotyped animals born in 2013 with sire and MGS genotypes and without dam genotype available. The MGSs of these animals had been verified by consortium partners before. Scenarios with segments of different lengths are included. The results differ from the numbers presented at the Interbull meeting due to further improvements.

Of these 5 600 animals 72% had a genotype with less than 10 000 genotyped SNPs in common with the standard SNP panel (currently the Illumina 50K version 1 chip), whereas the other 28% had 50K genotypes. SNP comparisons are only done for SNPs actually genotyped, so while most candidate MGSs had 50K genotypes, the actual comparison is determined by the genotyped SNPs in common with the animal’s genotype which is often low density. The MGS method proposed up to 4 most likely candidate MGSs.

Table 1 shows how often the pedigree MGS was found as most likely candidate MGS and how often amongst the at most 4 most likely candidates. The candidate MGSs were selected from on average 24 539 old enough bulls.

Computations in scenario 1, which was the fastest, required 134 minutes elapsed time on a server containing Intel Xeon X5560 quad core processors @ 2.8 Ghz with hyperthreading. The CPU load was 338% when using 4 threads (i.e. 2 cores with hyperthreading). Memory use is largely determined by the genotypes. With 87 874 genotyped animals 3.4 GB of RAM was used. Even though the required segment length between scenario 1 and 2 differed only with 1 SNP, a substantial time saving is obtained, because the number of potential matching segments per autosome reduced to 1 in scenario 1. The trio analyses was the slowest analyses requiring 554 minutes with a CPU load of 370%. The main reason for this difference is that the last 4 computational optimizations above only apply to the *long haplotype MGS detection* method. In contrast to the analyses done in Table 1 for demonstration purposes, in routine processing only animals with pedigree MGS in doubt (≤ 14 matches) and which are having an Italian registration number or have been genotyped by Italy are processed. Animals belonging to consortium partners are already processed by them before exchanging them. Therefore actual computation time in routine processing is far less.

Discussion

The alternative *long haplotype MGS* method of using very long haplotype segments for MGS verification and detection leads to the following advantages:

- Much more powerful than duo and trio methods
- No imputation and hence no imputation errors
- No need for a quick initial test at genotype arrival plus another final test after imputation as in VanRaden *et al.* (2013)
- Works across different chips, including low density
- No need for different thresholds per chip because of the conversion to a standard SNP panel beforehand.

Disadvantages:

- Some more chance of recombinations
- SNP chips should have a reasonable number of overlapping SNPs to allow for sufficient genotype comparisons without imputation. With the currently available SNP chips this is no problem.

For the best scenario (1) the pedigree MGS was rated 1st candidate MGS in 99.55% of the cases when including 24 539 candidate MGSs on average and genotypes of various densities. Furthermore the pedigree MGS was always within the list of 4 most likely candidate MGSs. This shows a strong concordance with the data which was already processed by USDA. In practice the pedigree MGS will only be rejected if it is not present in the candidate list. Therefore none of the MGS nominated by USDA would be rejected. VanRaden *et al.* (2013) mentioned an accuracy of 97% in Holstein with the imputed haplotype method when including 12 152 candidates. Between our method long haplotype MGS detection and their haplotype method with imputation results are very similar. These results are achieved while most combinations of animal and candidate MGS genotypes had less than 10 000 genotyped SNPs in common.

Conclusions

Long haplotype MGS detection using very long haplotypes enables MGS verification and detection without imputation and without phasing the pedigree or candidate MGS, resulting in an easier workflow. Even with low density genotypes good results are obtained. The rapid availability of a list of the most likely candidate MGSs can assist in finding and resolving switched samples and resolving pedigree errors. This is turn helps in improving selection, traditional and genomic breeding value estimation and in avoiding inbreeding.

Acknowledgements

The authors thank Maurizio Marusi and Tabatha Cooper for useful discussion and Paul VanRaden for his paper.

References

VanRaden, P.M., Cooper, T.A., Wiggans, G.R., O'Connell, J.R. & Bacheller, L.R. 2013. Confirmation and discovery of maternal grandsires and great-grandsires in dairy cattle. *J. Dairy Sci. 96,* 1874–1879.