Long Haplotype Method for Detection of Maternal Grandsire and Sire of Maternal Granddam

J.B.C.H.M. van Kaam and R. Finocchiaro

ANAFI, Cremona, Italy.

ABSTRACT: The availability of dense SNP genotypes on many animals enables the verification and detection of ancestors. Due to their large impact male ancestors are more often genotyped than female ancestors. It is possible to verify registered pedigree information for a few generations or even to detect candidate ancestors, when registered ancestors appear to be wrong. A method based on counting long haplotype segments in common between male ancestors and the animal of interest has been developed. This method enables the verification of registered paternal grandsires and sires of paternal granddams. In case of doubt a screening amongst available genotyped males can be done to detect the most likely candidate paternal grandsires or sires of paternal granddams. Pedigree, sampling and lab errors can often be corrected and data entering in and resulting from genetic and genomic evaluations is more accurate.

Keywords: haplotype; grandsire detection; greatgrandsire detection

Introduction

In recent years, large numbers of dairy cattle have been genotyped with SNP arrays (“chips”) ranging in density from 3K to 800K. Due to artificial insemination the number of male ancestors is much smaller than the number of female ancestors and a much larger fraction of them has genotypes available. So animals often have genotyped sires and grandsires, however their dams and granddams are usually not genotyped. Therefore in the female line pedigree verification and detection are more complicated than in the male line. SNP genotypes enable verification and detection not only of direct parents but also of grandparents and greatgrandparents, even when intermediate animals are without genotype.

Four methods will be mentioned here, each of which can be used for verification as well as detection of a maternal grandsire (MGS) and/or sire of the maternal granddam (SOMGD). These methods are all aimed at the case where the genotype of the dam (and maternal granddam) is not available. VanRaden et al. (2013) presented two methods named duo and trio which are based on conflicts of single markers. The duo method simply computes the fraction of markers which are opposite homozygotes in the MGS and the animal and uses this as test statistic. Sire and dam genotypes are not used. The trio method further extends the duo method, by adding sire genotype information, which results in more markers that are informative. If a sire is homozygous, while the animal is heterozygous, then the other allele of the animal originated from its dam. If none of the MGS alleles matches with the maternal allele of the animal then this will be counted in the test statistic as a genotype conflict. The trio method includes more genotype information in the test statistic compared to the duo method and therefore is more powerful.

A third method presented by VanRaden et al. (2013) for verifying or detecting a MGS is to compute the fraction of matching haplotypes of a given number of markers in common between the maternal haplotype of the animal and both haplotypes of a putative MGS and use this as test statistic. Methods based on haplotypes take advantage of the joint inheritance, and hence identity-by-descent (IBD), of multiple markers. Consequently they use more information and should be more powerful than methods ignoring linkage such as duo and trio methods. Given that duo and trio methods do not require phasing and imputation, they can be performed as soon as a genotype is received. When SNP chips of different density are used, then phasing (of the ancestors) and imputation will be necessary for the third method. Phasing and imputation requires substantially more computation time and consequently interrupts the workflow which is a serious disadvantage. A fourth method was introduced by Van Kaam and Hayes (2013) and can be seen as a variation of the third method. The method was named long haplotype ancestor detection. The need for phasing and imputation was resolved by using very long segments without phasing and without imputation. Instead of using more markers in the same (short) segment obtained by imputation their alternative is to use more markers by using much longer segments. This method has now been further extended to include not only verification/detection of MGSs, but also of the SOMGDs. This will be discussed here and has been implemented at Anafi.

Materials and Methods

Routinely, all genotypes from different SNP chips are converted to a standard SNP set (currently the Illumina 54K version 1 chip), before verifications take place. SNPs in the standard set which are not genotyped are set to missing. Therefore all samples have the same number of SNP genotype scores (currently 54,001).

The long haplotype method had two phases, which now is extended by a third phase:

1. Verification of MGS: For animals, whose pedigree MGS has a genotype, verify if their pedigree MGS meets all criteria to be credible.
2. Detection of MGS: For those animals where the pedigree MGS does not meet the criteria or has no genotype available create an ordered list of most likely candidate MGSs.
3. Detection of SOMGD: For those animals where the pedigree MGS does not meet the criteria or has no
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 of the animal of interest, for markers where this can be done unambiguously. That is, where the animal is homozygous, or where the animal is heterozygous and its sire is homozygous (as in the trio method).
2. Pass 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 markers) on all autosomes. This is used as test statistic 1 to select the most likely candidates.

The selection of suitable candidate MGSs 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
- Pedigree maternal granddam's sons are excluded as candidate MGS
- Pedigree paternal granddam's sons 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

The MGSs were detected assuming the most likely candidate MGS is the correct one. For each animal for which the SOMGD has to be verified or detected, the long haplotype SOMGD algorithm has the following steps:

1. Derive the grandmaternal haplotype of the animal of interest, for markers where this can be done unambiguously.
2. Pass through all suitable candidate SOMGDs (defined by the rules below), and count the number of matching non-overlapping haplotype segments of length $y$ (defined by the number of markers) on all autosomes. This is used as test statistic 2 to select the most likely candidates.

The selection of suitable candidate SOMGDs uses some rules:
- Only males can be SOMGD
- Animals cannot be their own SOMGD
- Pedigree sire of the SOMGD is excluded as candidate SOMGD
- Pedigree sire of the dam of the maternal granddam (DOMGD) is excluded as candidate SOMGD
- Pedigree DOMGD's paternal sibs are excluded as candidate SOMGD
- Pedigree DOMGD's maternal sibs are excluded as candidate SOMGD
- Pedigree DOMGD’s sons are excluded as candidate SOMGD
- Pedigree DOMGD’s sons are excluded as candidate SOMGD
- Minimum 1-generation interval > 600 days between dam and MGD and between MGD and DOMGD
- Minimum 2-generation interval > 1200 days between dam and SOMGD in case the MGD's birthdate is unknown

Within a segment, haplotypes are considered to match if the ancestor does not have any conflicts. With the animals’ allele received from the oldest intermediate ancestor. (Previously (Van Kaam and Hayes, 2013) 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.) All matching non-overlapping segments with a sufficient number of markers anywhere on any autosome are counted in the test statistics. If conflicting markers were found in a segment, the marker counting was reset to the first marker after the conflicting marker and the counting was continued from the point where it had arrived. Segments counted in the test statistics have the same number of markers for all SNP chips, because all SNP genotypes were converted to the same SNP set beforehand.

Implementation of the long haplotype method requires decisions regarding the length of haplotype segments and the required number of matching segments to conclude that the MGS or SOMGD is wrong or uncertain. In the second phase and third phase the test statistics are used to rank the most likely candidate MGSs and SOMGDs respectively. Haplotype IBD segments get shorter when more meioses have taken place, hence each generation the IBD segments get shorter. In practice MGS with $\leq 11$ matches were considered uncertain and with $\leq 6$ as likely wrong. Both these groups are included in regular candidate MGS detection to find possibly better candidates. Candidate MGSs need to have $\geq 9$ matches to be considered for the most likely candidate list. For SOMGD analysis the same threshold values were used, however the required length of a matching segment between a SOMGD and a MGD was a full autosome. Longer segments were used to have more informative markers.

Various computational aspects were applied in our software in order to make it fast:
- Fortran 2008 code
- Use of OpenMP for parallel processing
- Compilation optimization
- In the genotype array the inner-loop index was used as first index to take advantage of column-major storage as used in Fortran. In this manner marker processing corresponds with the storage order.
- Recounting (parts of) any haplotype segment was averted.
- When the remaining length on an autosome is no longer enough to obtain a matching segment the comparison is stopped.
- Use half an autosome plus 1 SNP (in the first scenario 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 terminated as indicated in the previous point.

- Stop comparing a candidate MGS or SOMGD if there is not enough space left on the remaining autosomes to reach the cutoff of > 9 matching segments for signaling a candidate MGS or SOMGD.
- Process autosomes in reverse order from higher number to the lower number autosomes in order to first do smaller autosomes. Together with the previous point this means that often comparison of the largest autosomes can be omitted. In practice, in the first scenario of Table 1, if none of the first 20 autosomes has a matching segment then the remaining 9 autosomes do not need to be compared because they cannot result in exceeding the cutoff of > 9 matching segments.

**Table 1: Effect of haplotype length on MGS detection.**

<table>
<thead>
<tr>
<th>Haplotype length IBD</th>
<th>Pedigree MGS as 1&lt;sup&gt;st&lt;/sup&gt;</th>
<th>Pedigree MGS in top 4</th>
<th>Elapsed time, min&lt;sup&gt;¥&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.50 chr+1 SNP</td>
<td>99.18%</td>
<td>99.91%</td>
<td>83</td>
</tr>
<tr>
<td>0.50 chr</td>
<td>99.82%</td>
<td>99.88%</td>
<td>105</td>
</tr>
<tr>
<td>0.33 chr</td>
<td>99.59%</td>
<td>99.72%</td>
<td>114</td>
</tr>
<tr>
<td>0.25 chr</td>
<td>99.63%</td>
<td>99.72%</td>
<td>118</td>
</tr>
<tr>
<td>0.10 chr</td>
<td>99.36%</td>
<td>99.70%</td>
<td>125</td>
</tr>
<tr>
<td>500 SNPs</td>
<td>99.52%</td>
<td>99.70%</td>
<td>123</td>
</tr>
<tr>
<td>75 SNPs</td>
<td>99.11%</td>
<td>99.61%</td>
<td>149</td>
</tr>
</tbody>
</table>

<sup>1</sup>Haplotype length for MGS detection.

<sup>2</sup>Haplotype length for SOMGD detection.

<sup>3</sup>Time includes MGS and SOMGD detection.

**Results and Discussion**

Results, for 5,600 non-Italian genotyped animals born in 2013 with sire and MGS genotypes and without available dam genotype, are given in Table 1 and 2. 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 by Van Kaam and Hayes (2013) due to further improvements. Of these 5,600 genotypes 72% had less than 10,000 genotyped SNPs in common with the standard SNP set (currently the Illumina 54K version 1 chip), while the rest had 54K genotypes. Only SNPs which were actually genotyped were included in SNP comparisons, so while most candidate males had 54K genotypes, the actual comparisons are determined by the genotyped SNPs in common with the animal’s genotype which is often low density. Our application proposed up to 4 most likely candidate MGSs and 3 most likely candidate SOMGDs assuming the most likely candidate MGS is the correct one. In Table 1 is shown how often the pedigree MGS was detected as most likely candidate MGS and how often amongst the at most 4 most likely candidates. In Table 2 is given how often the pedigree SOMGD was detected as most likely candidate SOMGD and how often amongst the at most 3 most likely candidates. The candidate MGSs and SOMGDs were selected from on average 24,539 and 14,869, respectively, old enough bulls.

**Table 2: Results of SOMGD detection.**

<table>
<thead>
<tr>
<th>Haplotype length IBD&lt;sup&gt;4&lt;/sup&gt;</th>
<th>Pedigree SOMGD as 1&lt;sup&gt;st&lt;/sup&gt;</th>
<th>Pedigree SOMGD in top 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00 chr</td>
<td>88.66%</td>
<td>92.85%</td>
</tr>
</tbody>
</table>

<sup>4</sup>Haplotype length for SOMGD detection.

Elapsed time in Table 1 refers to using 6 threads on a server containing 2 Intel Xeon X5560 quad core processors @ 2.8 Ghz with hyperthreading. In the fastest scenario computations required 83 minutes with a CPU load of 482%. Using 4 or 8 threads, elapsed time was 112/74 min and CPU load was 346%/551%. Memory use is chiefly determined by the genotypes. With 87,874 genotyped animals 3.4 GB of RAM was used. Even though the required segment length between the first two scenarios differed only with 1 SNP, a substantial time saving is obtained, because the number of potential matching segments per autosome reduced to 1 in the first scenario. Different to the demonstrations shown in Table 1 and 2, in routine processing only animals with pedigree MGS in doubt (≤ 11 matches) and which are having an Italian registration number or have been genotyped by Italy are processed. Genotypes shared by consortium partners are already processed by them before exchange. Therefore actual computation time in routine processing is far less.

**Conclusion**

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

**Literature Cited**
