Methods for Discovering and Validating Relationships among Genotyped Animals

G. R. Wiggans[1], P. M. VanRaden[2] and L. R. Bacheller[2]

[1] Council on Dairy Cattle Breeding, Bowie, MD 20716, USA

[2] Animal Genomics and Improvement Laboratory, Agricultural Research Service, U.S. Department of Agriculture, Beltsville, MD 20705-2350, USA

Summary

Genomic selection based on single-nucleotide polymorphisms (SNPs) has led to the collection of genotypes for over 2.2 million animals by the Council on Dairy Cattle Breeding in the United States. To assure that a genotype is assigned to the correct animal and that the animal's pedigree is correct, the pedigree parents are checked. As of January 2018, the sire was validated for 97% and the dam for 39% of the 2.2 million genotyped animals that passed edits. The genotype is compared with all other genotypes to detect unknown parent-progeny relationships or identical genotypes. If a parent is not confirmed, the grandsire is checked. If a grandsire is unknown or designated as unlikely, possible grandsires are proposed. If SNP conflicts for a parent-progeny pair are concentrated on a single chromosome, a chromosomal deletion or other abnormality is considered; 102 such cases have been detected. All comparisons consider the SNPs in common between the genotypes from the current 30 chip types. Comparison of each genotype with all others is a major and increasing consumer of computer resources. Because processing time has continued to increase, ways to reduce the time have been investigated. In 2012, a set of 1,000 SNPs that are present on nearly all chips was selected for preliminary screening. To further speed up processing, a set of 100 SNPs recently was selected based on minor allele frequency, call rate, and Mendelian consistency. Tests with the 100-SNP set showed that excluding cases with more than three opposite homozygotes could eliminate 99.7% of genotypes without eliminating any confirmed parent-progeny relationships. A continuing effort is required to maintain extensive checking and pedigree correction within the time available for processing incoming genotypes and applying updates caused by pedigree changes. In addition to grandsire checking when genotypes are loaded, maternal grandsire and maternal great-grandsire are checked and discovered using haplotypes from the imputation process as part of the genomic evaluation. For dams with unknown sires, the discovered maternal grandsire is assigned as her sire. The genotypes provide a rich source of information for validation and discovery of genetic relationships.

Key words: genotype validation, parentage discovery, genomic evaluation, pedigree

Introduction

Genomic selection (Wiggans *et al.*, 2017) based on single-nucleotide polymorphisms (SNPs) has led to the collection of over 2.3 million genotypes by the Council on Dairy Cattle Breeding in the United States. To assure that a genotype is assigned to the correct animal and that the animal's pedigree is correct, the pedigree parents are checked. As of January 2018, the sire was validated for 97% and the dam for 39% of the 2.2 million animals with genotypes that passed edits. Validation is by counting opposite homozygous calls for parent and progeny as reported by Wiggans *et al.* (2011). As a further check on accuracy of pedigree and correctness of genotype assignment, the genotype is compared with all other genotypes to find unknown parent-progeny relationships or identical genotypes.

Recently, animals with both parents confirmed are compared only with animals born from 500 days before their birth to the present to reduce processing time. If a parent is not confirmed, that parent's sire is checked. Preliminary grandsire checking is by counting opposite homozygotes as with parent-progeny relationships but with a higher threshold. This threshold declines as the number of comparisons increases. If a grandsire is designated as unlikely, possible grandsires are proposed. Animals with an unlikely grandsire are excluded from genomic evaluation to avoid

eventual pedigree changes and resulting evaluation volatility.

If SNP conflicts for a parent-progeny pair are concentrated on a single chromosome, a chromosomal deletion or other abnormality may be declared, and the parent confirmed; 102 such cases have been detected. Confirmations consider the SNPs in common between the genotypes of the two animals from the current 30 chip types. Comparison of each genotype with all others is a major and increasing consumer of computer resources. Because of this, ways to reduce processing time have been developed. In 2012, a set of 1,000 SNPs that are present on nearly all chips was selected for preliminary screening. The objective of this research was to investigate ways to further speed up processing time.

Materials and Methods

To determine if a small set of SNPs could be used to quickly identify a very small set of genotypes worthy of further investigation, a set of 100 SNPs was selected based on minor allele frequency, call rate, and Mendelian consistency. Cooper *et al.* (2013) reported that calling accuracy declines with call rate. To facilitate fast comparisons, genotypes were stored as 1-byte integers with values of 0, 1, 2, and 3 (3 indicates no call), which were used as subscripts in a two-dimensional matrix (one dimension for each animal) that returned a 0 for no conflict and a 1 for conflict. These were summed over the 100 SNPs. A similar matrix was used to accumulate the number of cases in which both animals had called genotypes that were homozygous.

To determine the appropriate threshold to identify possible parent-progeny relationships, all genotypes with confirmed parent-progeny relationships were processed to determine the highest number and percentage of conflicts for those relationships. A similar process was applied to detection of possible grandsires to determine an appropriate threshold for them. The effectiveness of the threshold for parent-progeny relationships was determined by processing all pairs of genotypes to determine what portion of all comparisons could be eliminated based on 100 SNPs.

The 1,690,706 Holsteins available resulted in 1.43 trillion comparisons of 100-byte genotypes. The ratio of conflicts to comparisons was set to 0.084; genotype pairs with lower values were retained as possible parent-progeny sets.

Results and Discussion

The 100-SNP set was effective in identifying a small set of possible parent-progeny relationships quickly. For Holstein, slightly over 4 million of the 1.43 trillion comparisons were retained (99.72% excluded). Processing took 81.4 minutes using 20 processors. In initial processing, the threshold was set at three conflicts. This threshold is in line with the recommendations of the International Society for Animal Genetics (2012) for a panel of 100 SNPs. By changing the threshold to 0.084, the number of conflicts divided by the number of pairs with both genotypes homozygous, all true parent-progeny relationship pairs were included and fewer other pairs. A preliminary test with 50 SNPs showed that only about half the animals could be eliminated while retaining all the parent-progeny relationships.

Identification of possible grandsires is currently done using a 1,000-SNP set (Wiggans and Bacheller, 2014). Substantial speed-up in this process was achieved recently by simplifying conflict detection by not checking for a conflict when the animal is heterozygous. This checking is possible when the other parent (the one that is not the progeny of the grandsire being searched for) is genotyped. The same 100 SNPs as proposed for parent-progeny relationships but with a threshold of seven could be used to detect possible grandsires. Only males need to be processed, and a further reduction in processing can be achieved by processing genotypes from only those bulls that have progeny. Currently, an age check eliminates bulls that are too young to have grandprogeny. Because a potential grandsire can have a small number of conflicts as the result of other relationships, such bulls are excluded from the list of possible grandsires. Sometimes this causes the true grandsire to be missed because he also is related through another path.

A more reliable method to discover grandsires is to compare haplotypes rather than SNPs (VanRaden *et al.*, 2013). This method does require that the genotype be phased and imputed. In searching for a maternal grandsire, the true maternal grandsire is expected to have 45% of his haplotypes in common with the animal's maternal haplotypes. This method is currently used to check and discover maternal grandsires and maternal great-grandsires using haplotypes from the imputation process as part of the genomic evaluation. For dams with unknown sires, the discovered maternal grandsire is assigned as her sire.

In a recent test, haplotype detection of maternal grandsires and maternal great-grandsires took 8 hours with 20 processors to process all 1.8 million Holsteins (except the ~30% with genotyped dams that did not need processing). Checking haplotypes is efficient because all 100 markers in a haplotype are represented by a single haplotype identification so a single conditional (if) statement effectively compares all of them. Accuracy is gained by using the SNP genotypes of the animal and other parent to evaluate the potential grandsire for all 60,671 markers; this avoids accuracy loss from using only the observed SNP subset, reducing the number of SNPs, or comparing only the animal and possible grandsire.

The haplotype method could be applied weekly as new genotypes are received. This would allow the genotype loading process to be faster by removing the search for possible grandsires and would provide a more accurate list of possible grandsires.

The haplotype method can provide a likely grandparent even if the parent is not identified. If the parent is unknown, an identification number would need to be constructed to enable storing the grandsire where pedigree is stored as IDs of an animal and its parents. By doing this, the numerator relationship matrix would correspond more closely to the genomic relationship matrix.

Potential Further Speed-Up in Processing

In addition to the reduction in comparisons already imposed for genotypes with both parents confirmed, animals with a different confirmed parent cannot have an identical genotype. Therefore, they could be skipped when searching for identical genotypes.

Further speed-up may be possible by limiting the total number of SNPs checked. As new chips have been introduced, most have included a large portion of the 6,909 SNPs on the Illumina BovineLD BeadChip (Boichard *et al.*, 2012). Using the same criteria as for selecting the 100 SNPs above, a set of 4,366 SNPs were selected. These SNPs would be the only ones used to confirm identical genotypes, discover parent-progeny relationships, and determine if the grandsire was unlikely. The primary advantage of this smaller set is to reduce start-up time and memory requirements. Also, the same set of SNPs would be stored for all chips, which would eliminate the complexity of matching SNPs from different chips. Currently, all genotypes are loaded in memory from a sequential file at the start of processing to minimize database access. If only a portion of the SNPs was stored, the start-up time should be substantially reduced.

As part of the genotype quality-control process, accessing full genotypes of parents would still be necessary to assess Mendelian consistency of all SNPs in common between the animal and its parents. The 4,366-SNP set may not be the ultimate solution because it would continue to grow and contribute to the start-up time, whereas the time to directly access the database for the relatively few genotypes needed would be largely independent of the size of the database. As the number of genotypes continues to increase, efficiencies in storing and comparing them can be achieved by employing bit-level processing (Chang *et al.*, 2015).

Conclusions

A continuing effort is required to maintain extensive checking and pedigree correction within the time available for processing incoming genotypes and applying updates caused by pedigree changes. The genotypes allow for validation and discovery of parents and, with less certainty, grandsires and even great-grandsires. The genotypes provide a rich source of information for validation and discovery of genetic relationships.

Acknowledgements

Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by USDA; USDA is an equal opportunity provider and employer.

References

Boichard, D., Chung, H., Dassonneville, R., David, X., Eggen, A., Fritz, S., Gietzen, K.J., Hayes, B.J., Lawley, C.T., Sonstegard, T.S., Van Tassell, C.P., VanRaden, P.M., Viaud-Martinez, K.A. & Wiggans, G.R. 2012. Design of a bovine low-density SNP array ooptimized for imputation. *PLoS ONE 70*, e34130.

Chang, C.C., Chow, C.C., Tellier, L.C.A.M., Vattikuti, S., Purcell, S.M. & Lee, J.J. 2015. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience 4*, 7.

Cooper, T.A., Wiggans, G.R. & VanRaden, P.M. 2013. *Short communication:* Relationship of call rate and accuracy of single nucleotide polymorphism genotypes in dairy cattle. *Journal of Dairy Science 96*, 3336–3339.

International Society for Animal Genetics. 2012. Guidelines for cattle parentage verification based on SNP markers. http://www.isag.us/Docs/Guideline-for-cattle-SNP-use-for-parentage-2012.pdf.

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. *Journal of Dairy Science 96*, 1874–1879.

Wiggans, G. & Bacheller, L. 2014. Improved discovery of maternal grandsires. Changes to evaluation system (April 2014). https://queries.uscdcb.com/reference/changes/eval1404.htm. Accessed Jan. 11, 2018.

Wiggans, G.R., Cole, J.B., Hubbard, S.M. & Sonstegard, T.S. 2017. Genomic selection in dairy cattle: The USDA experience. *Annual Review of Animal Biosciences 5*, 309–327.

Wiggans, G.R., VanRaden, P.M. & Cooper, T.A. 2011. The genomic evaluation system in the United States: Past, present, future. *Journal of Dairy Science 94*, 3202–3211.