Detecting Relationships among Genotypes in a Rapidly Growing Collection

G. R. Wiggans, G. B. Jansen, L. R. Bacheller, and J. A. Carrillo
Council on Dairy Cattle Breeding, Bowie, MD 20716, USA

Abstract

To correct pedigree errors and discover genotype misassignments, the Council on Dairy Cattle Breeding in the United States compares each new genotype with existing genotypes. With over 6 million genotypes as of May 2022, this is a computationally demanding task. The process was recently revised to maintain a table of genotype pairs that are similar enough to qualify as having a parent-progeny relationship or to be identical. Those genotype pairs are identified by a unique genotype identification and thus are unaffected by changes in genotype assignment to animal. Having those pairs substantially reduces processing time when propagating the effects of pedigree or assignment changes on the usability of genotypes. A set of 3,552 SNPs selected based on call rate and Mendelian consistency is used for the comparisons. Determination of percentage of conflicts stops after 96 and 1,000 SNPs if members of a genotype pair are unlikely to be related. The memory required to store the set of genotypes that is being searched is minimized by using just 2 bits per SNP. The time to access those genotypes is minimized by using memory mapping, which effectively makes the disk where the genotypes are stored an extension of memory. New or updated genotypes are compared with a restricted set of genotypes (one per animal) to reduce processing time. All animals with genotyped progeny are checked. Remaining genotypes are compared in birth date order so that no genotypes from animals born more than 12 years earlier are checked. This limit is reduced to 5 years if both parents of the animal are confirmed. Non-AI bulls with no progeny born in the last 5 years are skipped. Initial determination of unlikely grandsires is done using SNP-at-a-time comparisons and the genotype of the other parent (if available) based on the same 3,552 SNPs. During weekly and monthly evaluations, grandsires are validated using imputed haplotype comparisons. The reliance on the new procedure for discovery of close relatives eliminates the need to access full genotypes of all animals as was previously done. Previously, to minimize database access, all genotypes were loaded in memory from a file. Now, only those full genotypes needed to confirm pedigree relationships are retrieved from the database. The genotypes in the database are compressed, which reduces storage by 75%. These modifications allow comprehensive genotype checking while keeping processing time within acceptable limits.

Key words: genomic evaluation, parentage validation, parentage discovery

Introduction

The Council on Dairy Cattle Breeding (Bowie, MD) maintains the US collection of genotypes used for genomic evaluation of dairy cattle. It included over six million genotypes as of May 2022 (Council on Dairy Cattle Breeding, 2022). Genotypes are compared with those of parents and progeny to confirm pedigree and accuracy of genotyping and with all other genotypes to discover parent-progeny relationships and identical genotypes (Wiggans *et al.*, 2018). As the collection grows with over 1 million added in 2021, the time required for these comparisons has increased. The system in addition to validating new genotypes determines which genotype should be designated as usable if there is a pedigree error or other conflict. The genotype with the greater number of confirmations is generally designated as the usable one. If an animal has

more than 1 genotype, the one with the largest number of usable SNP genotypes is designated as the primary usable one.

Materials and Methods

Factors Affecting Speed of Discovery

The time required to discover parent-progeny and identical relationships is determined by number of SNPs compared, number of animals checked, method used for comparison, and times the comparison is done. The previous system relied on 1,000 SNPs for most comparisons and ended the comparison early if enough conflicts had been encountered so that a parent-progeny relationship was unlikely (Wiggans and Bacheller, 2014). Checking was limited to genotypes designated as primary. Comparisons relied on using the SNPs as subscripts in a matrix to determine conflicts. A set of vectors was required to identify the 1,000 SNPs on the genotypes that were being compared. With pedigree updates and identification changes, comparisons were run every time there was a change.

Characteristics of New Discovery System

With the goal of maintaining reasonable processing times for a collection of genotypes that was increasing in size at an increasing rate, improvements in all the factors that affect processing time have been implemented. A set of 3,552 SNPs was selected based on call rate and Mendelian consistency. That set includes most of the 1,000 SNPs used previously, and the set is sequenced so that the first 96 are the most informative, which enables most genotypes to be excluded after checking only those 96 SNPs This number was based on Wiggans *et al.*, 2018 who found that checking 100 SNPs was sufficient to exclude most genotypes that were not a parent or progeny. A second decision point is after 1,000 SNPs. A genotype is rejected as a close relative if the percentage of conflicts compared with those not missing is >12% at 96 SNPs, 3.1% at 1,000 SNPs, or 0.5% at 3,552 SNPs. To minimize memory required, the SNPs are stored in 2 bits (4 SNPs per byte) with 3 designating a missing value. The not-missing SNP genotypes are represented as the count of the A allele. The time to access those genotypes is minimized by using memory mapping (Kerrisk, 2021), which effectively makes the disk where the genotypes are stored an extension of memory.

To minimize the number of genotypes compared (as in the current system), only one genotype per animal is compared, but unusable genotypes (including those not assigned to an animal) also are compared. For animals with both parents confirmed, comparison is limited to those born not more than 5 years earlier than the subject animal. For other animals, comparisons extend back to those born 12 years earlier. Genotypes where the birth date is unknown use the date the genotype was loaded in place of the birth date. To skip comparisons with bulls that have never been used, non-AI bulls with no progeny born in the last 5 years are excluded. All animals with genotyped progeny are checked.

Comparisons are done using a two-dimensional matrix with the genotypes as subscripts to determine conflicts and if both genotypes have a call for the SNP. A matrix also was used in the previous system (Wiggans *et al.*, 2018).

The major change from the previous system is that all genotype pairs that are similar enough to have a parent-progeny or identical relationship are stored in a close-relative table and identified by genotype (rather than animal); thus, genotype pairs are unaffected by genotype reassignment. This eliminates the need to repeat genotype comparisons with all genotypes. Determination of usability involves analyzing the effect of a change in usability on other genotyped animals and often requires multiple rounds until no more animals have changes.

The file of 3,552 SNP genotypes used for discovery contains a row for every genotype in the genotype table, including those that have been withdrawn. This simplifies maintaining the association between the database table and the file used for discovery as the sequence

number in the database can be used to address the corresponding 3,552 SNP genotype. To identify which genotypes are to be used for discovery, a pointer file is created and contains the displacements (sequence number – 1) of the genotypes to be checked. Those that have been withdrawn, are not the primary genotypes for an animal, or are 3K genotypes for animals without genotyped progeny are not included. The sequence numbers are ordered by birth date for animals without genotyped progeny. Those with progeny are at the end, which causes their genotypes to be checked in all comparisons as iteration starts at the end. This enables the comparisons to be limited to a specific birth date by setting the end of the loop. The end point is determined by a vector of the earliest position in the pointer file of each birth date since January 1, 2009. Provision is made for withdrawn genotypes to be restored, the genotype designated as primary to change, old bulls first getting progeny to be added, and a genotype to be updated. The reason that iteration starts at the end is to allow new genotypes to be added to the end of both the discovery and pointer files. The pointer file is updated weekly, which enables the exclusion of recently withdrawn genotypes and reestablishment of birth date order.

Determination of Grandsire Likelihood

The same set of 3,552 SNPs is used to determine if the grandsire is unlikely. For this purpose, the genotype of the other parent is used if available so that a heterozygous SNP can count as a conflict if the parent and the sire of the other parent are the same homozygote. The threshold is 8% without and 13% with the other parent. During weekly and monthly evaluations, grandsires are validated using imputed haplotype comparisons, which may cause an unlikely designation to be removed.

Results & Discussion

The previous system used an initialization file that included all genotypes to minimize database access during processing (Wiggans *et al.*, 2018). With discovery now done using a discovery file, accessing all genotypes is no longer necessary; therefore, they have been removed from the initialization file, which substantially reduces memory requirements. The database is accessed for full genotypes of the animal and its parents and to validate discovered ancestors. The smaller initialization file also reduces start-up time. The compression of 4 SNPs per byte is also applied to the database table so that genotypes now require only 25% of the disk space used previously. All submissions are subjected to initial checking to enable detection of issues such as problems as systematic misassignment of genotypes, or low call rate or excessive SNP parent-progeny conflicts requiring reclustering.

Conclusions

The discovery phase of this update of the program that checks and loads genotype was implemented in June 2022 and resulted in substantial time savings, particularly for updates when no new genotypes were added so that comparisons with all genotypes were not needed. Limits on the birth years of genotypes checked reduced the growth in processing time somewhat; however, if the number of genotypes received per year continues to increase, processing time will increase. The reduction in time for updates is particularly important because processing times of >5 hours was not uncommon with the previous system.

Acknowledgments

The authors thank S.M. Hubbard for technical review. The Council on Dairy Cattle Breeding is an equal opportunity provider and employer.

References

Council on Dairy Cattle Breeding. 2022. Genotype counts by chip type, breed code, and sex code in database as of 2022-05-30. https://queries.uscdcb.com/Genotype/cur_freq.html. Accessed June 8, 2022.

Kerrisk, M. 2021. mmap(2) — Linux manual page. Linux Programmer's Manual. https://man7.org/linux/man-

pages/man2/mmap.2.html. Accessed June 9, 2022.

Wiggans, G., and Bacheller, L. 2014. Improved discovery of maternal grandsires. Changes to evaluation system (April 2014). https://queries.uscdcb.com/reference/changes/eval1404.htm. Accessed June 8, 2022.

Wiggans, G.R., VanRaden, P.M., and Bacheller, L.R. 2018. Methods for discovering and validating relationships among genotyped animals. *Interbull Bulletin 53,* 27–30.