J. Dairy Sci. 94:2114–2117
doi:10.3168/jds.2010-3896

Technical note: Efficient parentage assignment and pedigree reconstruction with dense single nucleotide polymorphism data

B. J. Hayes*†[1]
*Biosciences Research Division, Department of Primary Industries Victoria, Melbourne, Victoria, Australia 3083
†Dairy Futures Cooperative Research Centre, Melbourne, Victoria, Australia 3083

ABSTRACT

Large numbers of dairy cattle are now routinely genotyped for dense single nucleotide polymorphism (SNP) arrays for the purpose of predicting genomic estimated breeding values. Such SNP arrays contain very good information for parentage assignment and pedigree reconstruction. The main challenge in using this information for parentage assignment and pedigree reconstruction is development of computationally efficient strategies that enable a candidate animal to be assigned its sire and dam with the large volume of data. Here we describe an efficient algorithm for parentage assignment with SNP data and demonstrate very accurate assignment with 50,000-SNP and 3,000-SNP panels. The computer code implementing the algorithm is given in the Appendix.
Key words: parentage, single nucleotide polymorphism, efficient

Technical Note

Large numbers of dairy cattle are now routinely genotyped for dense SNP arrays to predict genomic EBV (e.g., VanRaden et al., 2009; Harris and Johnson, 2010). This information can also be used for accurate parentage assignment and pedigree reconstruction. Incorrect parentage assignment decreases genetic gain in dairy industries by approximately 15% (Banos et al., 2001). Fisher et al. (2009) demonstrated that a 40-SNP panel could be used to assign parentage if mating records and birth dates were also used. However, with the release of a low-cost, 3,000 (**3K**) bovine SNP panel, it is likely that very large numbers of cattle will be genotyped for this panel rather than for smaller panels. The main challenge in using these larger SNP panels for parentage assignment and pedigree reconstruction is development of computationally efficient strategies that enable a candidate animal to be matched to its sire or its dam with the large volume of data.

With SNP data, an animal can be excluded from being the parent of an individual if, at any locus, the individual and the prospective parent are both homozygous but for different alleles ("opposing homozygotes"). Thus, given the current candidate, an efficient strategy would be to loop through potential sires and dams, compare the genotypes of the current candidate with the potential parent from the first SNP in the data set, and then eliminate the animals from the list of potential parents as soon as an opposing homozygote genotype is encountered. However, this strategy encounters 2 potential problems. The first is that if the individual has a monozygotic twin, the twin will also have no opposing homozygotes compared with the current candidate. This problem can be avoided if the date of birth is known and a condition is set that potential sires and dams must be at least 1 yr older than the current candidate. The second problem is that genotyping errors can result in even the true parents apparently having a small number of opposing homozygotes with their progeny. This problem can be dealt with if the distribution of genotyping errors is known.

The first task then is to determine the distribution of genotype errors in the genotype data. We did this empirically in a data set of 464 Holstein-Friesian cows, with known sires among a further data set of 2,126 Holstein-Friesian bulls. Four hundred and three Jersey bulls were also included to ensure that the method was not confounded by having other breeds in the data set. With 43,115 SNP genotypes (see Verbyla et al., 2009 for a description of quality control on the SNP data), the number of opposing homozygotes was many times larger when a cow was compared with a bull that was not its sire than when compared with its sire, and the 2 distributions did not overlap (Figure 1). This was also true when the SNP on the Illumina (San Diego, CA) Bovine 3K chip (with 3,123 SNP, extracted as a subset from the 43,115-SNP set described above) was used for the same purpose (Figure 1).

One problem frequently encountered with parentage assignment is distinguishing between the sire of an indi-

Received October 5, 2010.
Accepted January 6, 2011.
[1]Corresponding author: ben.hayes@dpi.vic.gov.au

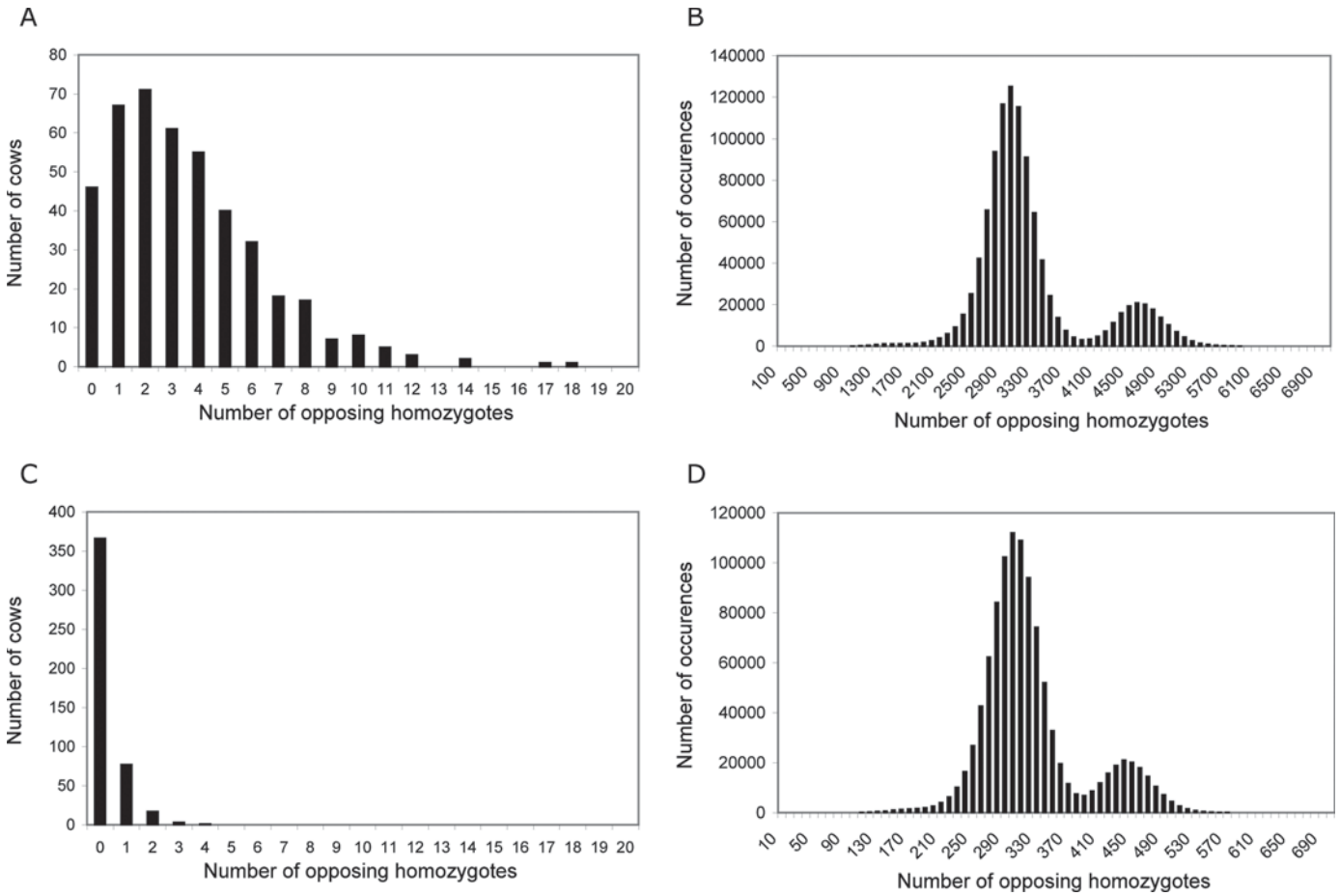

Figure 1. Distribution of number of opposing homozygote genotypes between animals (A) when a cow and its real sire are compared at 50,000 SNP genotypes; (B) when a cow and individuals other than its true sire are compared at 50,000 SNP genotypes; (C) when a cow and its real sire are compared at 3,000 SNP genotypes; (D) when a cow and individuals other than its true sire are compared at 3,000 SNP genotypes. The second normal distribution to the right of the large distribution in (B) and (D) results from comparing the Holstein-Friesian cow genotypes to Jersey bull genotypes.

vidual and the full brother of the sire. In our data set, 44 cows had both a sire and a first uncle (full brother of the sire) in the data set. The lowest number of opposing homozygotes for a cow and its full uncle was 1,033 for the 50,000 (**50K**) array and 88 for the 3K array, which is well above the maximum number of conflicts due to genotyping error encountered when a cow and its real sire were matched. The maximum number of conflicts due to genotyping errors was <20 for 50K and <5 for 3K (Figure 1).

We then tested the 50K-, 3K-, 300-, 150-, and 100-SNP panels for their accuracy in assigning sires to the cows. The 300-, 150-, and 100-SNP panels were chosen to be evenly spaced along the genome. We assessed how many of these cows could be assigned to a sire unambiguously (e.g., no more than 1 match with fewer than 20 opposing homozygotes for the 50K and 3K panels, and no opposing homozygotes for the smaller SNP panels). The accuracy of assignment was determined (the percentage of sire assignments that were correct; Table 1). For the 50K-, 3K-, and 300-SNP panels, all cows were matched to a sire unambiguously, and all cows were matched to their sires correctly. For smaller SNP panels, a proportion of cows could not be matched unambiguously to a sire. Given the widespread use of AI in the dairy industry and the global exchange of semen, one challenge with using SNP for parentage may be that no one country has a database containing SNP genotypes for all the sires used in that country. One potential solution is to have an international database of sire genotypes (e.g., hosted at Interbull) available to all participating countries. If concerns existed that such data could be used to predict genomic EBV, a standard 300-SNP panel of genome-wide SNP could be used. Such a panel is too small for predicting genomic breeding values, even with imputation.