The number of single nucleotide polymorphisms and on-farm data required for whole-herd parentage testing in dairy cattle herds

P. J. Fisher,*†¹ B. Malthus,* M. C. Walker,* G. Corbett,‡ and R. J. Spelman*

*Livestock Improvement Corporation Ltd., Hamilton 3240, New Zealand

†AgResearch, Invermay, Mosgiel 9053, New Zealand

‡ZyGEM Corporation Ltd, Hamilton 2001, New Zealand

ABSTRACT

New platforms utilizing single nucleotide polymorphisms (SNP) offer operational advantages over the conventional microsatellite-based ones, making them a promising alternative for parentage exclusion. Through simulation and empirical data, a 40-SNP panel (where the minor allele frequency was 0.35 on average) was shown to be a comparable or better diagnostic tool than the current 14-microsatellite panel that is used to parentage test New Zealand dairy animals. The 40 SNP alone did not have sufficient power of exclusion to match more than 75% of the progeny to the correct sire and dam. Utilizing mating records and grouping progeny and dams by birth and calving dates, respectively, decreased the number of sire-dam combinations that each progeny was tested against and dramatically increased the utility of the SNP. These results highlight the importance of combining genotypes with on-farm data to maximize the ability to assign parentage in the New Zealand dairy herd.

Key words: bovine, parentage testing, pedigree record, single nucleotide polymorphism

INTRODUCTION

Parentage testing in dairy cattle is carried out for several reasons, such as to increase genetic gain (through increased accuracy of animal evaluation and reduction of inbreeding) and to maintain accurate pedigree information (Davis and DeNise, 1998; Banos et al., 2000; Israel and Weller, 2000; Spelman, 2002). Traditionally, farmers and breeders have relied on matings and birth records to assign parents. However, the gathering of precise pedigree data can both be inconclusive and disruptive to farm management practices. For example, assigning dams to calves requires that parturition or suckling, or both, be observed (Dodds et al., 1996, 2005).

Parentage error rates of 4 to 6% (mother-daughter) and 12 to 15% (sire-daughter) have been indicated in New Zealand, where the national dairy herd is almost exclusively pastorally farmed (Spelman, 2002).

In the last decade, co-dominant DNA markers such as the multiallelic microsatellites have been used increasingly for parentage assignment (Geldermann et al., 1986; Kashi et al., 1990; Usha et al., 1995; Glowatzki-Mullis et al., 1995; Heyen et al., 1997). More recently, use of SNP has become more prominent (Heaton et al., 2002; Glaubitz et al., 2003; Werner et al., 2004). Single nucleotide polymorphisms offer practical improvements over microsatellites including easier automation and scoring and lower mutation rates (Amorim and Pereira, 2005; Anderson and Garza, 2005).

This study aimed to investigate by stochastic modeling the number of SNP that are required to give power of exclusion comparable to the microsatellites used currently. The simulation results were then empirically tested with 1 dairy herd that had been genotyped with both microsatellites and SNP. A secondary objective was to investigate the value of combining on-farm data with genotypes to aid in parentage.

MATERIALS AND METHODS

Parentage Matching Simulations

Stochastic simulations were undertaken to test the exclusion power of both microsatellites and SNP. For a given number of markers and allele frequencies, genotype profiles for a sire and dam were simulated. Using the laws of Mendelian inheritance, a single daughter was generated for the simulated sire and dam. In addition, another 10,000 putative sires' genotype profiles were simulated using the same marker and allele frequencies as the true sire. The daughter was then tested for parentage against these sires, and the percentage of occurrences of incorrect acceptance of parentage for a sire was calculated. Fifty replicates of this process were undertaken for each scenario. The power of exclusion was calculated as 100 – percentage of incorrect acceptance of sire. The scenarios investigated were: 1) the putative

Received February 7, 2008.

Accepted September 15, 2008.

¹Corresponding author: paul.fisher@agresearch.co.nz

Table 1. Coefficients of co-ancestry (%) for the 11 assumed commercial sires of the test herd from known pedigree records

Sire no.	1	2	3	4	5	6	7	8	9	10	11
1	51.0	11.0	10.0	6.9	3.7	8.3	9.7	2.9	9.6	2.2	10.2
2		53.4	6.4	15.5	1.7	6.6	6.4	2.4	6.5	2.8	8.3
3			50.4	6.8	4.1	4.9	6.0	3.8	5.8	2.8	7.1
4				51.7	2.1	2.6	3.5	2.9	3.6	1.6	4.9
5					52.1	4.4	16.5	7.0	15.7	6.0	8.2
6						51.0	7.8	2.4	6.1	3.4	5.9
7							52.1	6.3	18.9	5.4	8.6
8								52.0	6.1	14.4	5.0
9									51.9	5.4	8.5
10										50.5	3.6
11											53.1

sires were unrelated, 2) the putative sire was half- and full-sib brother of the true sire, and 3) the dam's genotype information was or was not included. Simulation was undertaken for only sire genotypes, which reflected the most common scenario within the industry where dams are usually not typed. These conditions were simulated with either a 14-microsatellite panel (with 3 or 4 alleles at equal frequency) or SNP panels ranging from 24 to 72 in number and marker minor allele frequencies (**MAF**) ranging from 0.2 to 0.5. The microsatellite, or simple sequence repeat markers (**SSR**), were chosen to have 3 to 4 equal frequency alleles, because this best represented the current commercial marker set, with heterozygosity and polymorphism information content (**PIC**) scores (Botstein et al., 1980) between 60 and 67% (based on 3 yr of commercial parentage test results for Jersey and Holstein-Friesian breeds in New Zealand). Three equal frequency alleles produce a PIC score of 0.67; therefore, we were conservative in favor of the 3- to 4-allele SSR marker definitions when making SSR-SNP comparisons.

Animals and DNA Samples

The commercial dairy herd consisted of 213 Jersey animals, where 159 dams had been artificially mated to 11 commercial sires. Known pedigree information was used to calculate co-ancestry coefficients of these sires (Table 1). A total of 43 daughters born between July 24 and September 1, 2003, were retained from these matings. This 40-d spread is typical in New Zealand where mating is designed to produce a 6-wk calving period.

Genotyping

The Magnesil Blood Genomic Max Yield System 1 × 96 kit (Promega Corporation, Madison, WI) was used to extract DNA, according to manufacturer's instruc-

tions for a 200-μL prep, but was scaled down to use 20 μL of blood.

Microsatellite genotyping was carried out as follows: 14 microsatellite markers were amplified in a multiplexed PCR reaction using PCR reagents from Qiagen (Qiagen, Hilden, Germany). The PCR was carried out under standard conditions at the GeneMark genotyping laboratory (Hamilton, New Zealand). Electrophoresis was carried out on an ABI3100 sequencer, and analysis was undertaken using GeneMapper software (Applied Biosystems, Foster City, CA). The 14 microsatellites consist of recommended International Society of Animal Genetics parentage testing markers. They are BM1864, BM2113, ETH10, ETH225, INRA23, SPS115, TGLA122, TGLA126, TGLA227, AGLA293, ETH3, MGTG4B, TGLA53, and TGLA57.

Genotyping of 99 SNP was carried out by Sequenom Inc. (San Diego, CA) using the MassEXTEND iPLEX assay. The markers were either selected from a panel of beef genotyping markers (Michael Heaton, US Meat Animal Research Centre, Clay Center, NE) or they were on the bovine 10K GeneChip (Affymetrix, Santa Clara, CA). The Affymetrix SNP were selected if they were able to be genetically mapped and had MAF above 0.40 for one breed and above 0.30 for the second breed from maternally inherited F1 dam alleles of a 800-dam Holstein-Friesian × Jersey crossbred trial (Spelman et al., 2001).

Twenty-seven markers that failed to pass technical criteria, departed from Hardy-Weinberg equilibrium, or mapped to the same position were not assessed further, leaving 72 SNP for further analysis (14 of these were from the US Meat Animal Research Centre set and 58 were derived from the Affymetrix GeneChip).

Ten panels with varying SNP number were generated postgenotypically from among the 72 genotyped SNP. The 72 SNP were ranked on their MAF, and the 10 panels were generated from the 24, 28, 32, 36, 40, 44, 48, 52, 56, or 60 markers with the greatest MAF.

Parentage Matching with Real Genotypic Data

In the test herd, the 43 daughters were matched to the possible sires and dams using software designed in-house. With this program, each daughter is simultaneously tested against all possible sire and dam combinations, and all of the possible parentage links are stored for each sire and dam. Once this has been completed for each daughter, unique links are identified at the progeny and dam level. A unique link is where a progeny has only 1 dam possible or a dam has only 1 progeny possible. In the instance where one of the multiple progeny of a dam has only that dam as its mother, all of the other possible progeny links for that dam are removed. Where 2 progeny share sire and dam and have the same birth dates (within a day of each other), the offspring are assumed to be twins. By removing the other progeny links for this dam, some of these progeny may now have a unique dam solution (i.e., previously they had 2 possible dams and now with the dam A link being removed, dam B is the unique solution). The software iterates through all of the solutions until there are no more links that can be removed. The unique link identification is only undertaken for dams and progeny; dairy cattle sires have a large number of progeny in each herd.

To decrease the initial number of sires and dams that each progeny is tested against, the software can utilize mating data, calving date for the dam and birthdate for the progeny. The user has the ability to initially restrict the progeny to the known sire-dam combinations as derived from the mating data and also restrict the progeny to be only tested against dams that calved in the same period as when the calf was born. Where genotypic information is consistent with on-farm data, the records are assumed to be correct; the priority for this set of procedures is to provide a more powerful level of exclusion than would be observed by genotypes or on-farm data alone. False mating records are expected to be detected by the genotyping alone; there is an option to accept chosen numbers of mismatches as either misparentage or as genotyping errors. For the remaining unallocated calves and dams, the criteria around mating information and calving and birth date relationships can be relaxed (or removed entirely).

RESULTS

Simulations

Simulation of the microsatellite panel gave an exclusion probability of 0.95 when the putative sire was a half-sib to the true sire and no dam genotypes were used. In other words, the putative sire was incorrectly

Table 2. Probability of parentage exclusion, using simulated data, for different numbers of SNP and minor allele frequencies (MAF) and when the putative false sire is a half-sib to the true sire

Panel	MAF			
	0.2	0.3	0.4	0.5
24 SNP	0.55	0.78	0.88	0.91
36 SNP	0.69	0.90	0.96	0.97
48 SNP	0.80	0.96	0.99	0.99
60 SNP	0.86	0.98	1.00	1.00
72 SNP	0.98	0.99	1.00	1.00

matched as the sire 5% of the time. For the same simulation scenario, 48 SNP with an average MAF of 0.3 or 36 SNP with an average MAF of 0.4 are required to have a similar exclusion probability as the microsatellites (Table 2).

The degree of relatedness of the putative sire to the real sire (represented as unrelated, half-sibs, or full-sibs) and the use of dam genotypes affected the exclusion probability. Although representing the sires in only 3 levels of relatedness would not likely provide very precise proxies for many real herd situations, the data was consistent in one factor; in every tested scenario, 36 SNP with an average MAF of 0.4 had an exclusion probability equivalent or better than with the microsatellite markers (Table 3).

Real Data

Without the utilization of genotypes, the parent-matching software could identify only 2 of the 43 daughters with unique dam assignments utilizing calving and birth dates.

Using the genotypes from the 14 microsatellites, 72% of the daughters (i.e., 31 out of 43) were matched to both parents unambiguously when neither mating nor calving-birth dates were used (Table 4). This increased to 79% when mating data were included and to 95%

Table 3. Probability of exclusion for 36 SNP, simulating true and false sires that are unrelated, half-sibs or full-sibs with an average minor allele frequency of 0.4, and 14 microsatellites averaging 3.5 alleles at uniform frequency

Scenario			
Putative sire	Dam genotypes	SNP	Microsatellites
Unrelated	No	1.00	0.99
Unrelated	Yes	1.00	1.00
Half-sib	No	0.96	0.95
Half-sib	Yes	1.00	0.99
Full-sib	No	0.88	0.84
Full-sib	Yes	0.99	0.98

Table 4. Percentage of daughters uniquely matched with sire and dam, using real data with and without mating data or birthdates, or both, for parentage matching for a variety of SNP panels and a 14-microsatellite panel

Panel	MAF ¹	PIC ²	Birth-calving data		No birth-calving data	
			Matings	No matings	Matings	No matings
Microsatellite	NA ³	0.70	95	93	79	72
24 SNP	0.40	0.36	98	88	11	6
28 SNP	0.39	0.36	98	93	33	14
32 SNP	0.37	0.36	98	100	43	30
36 SNP	0.36	0.35	100	100	68	53
40 SNP	0.35	0.35	100	100	82	72
44 SNP	0.33	0.34	100	100	86	78
48 SNP	0.32	0.34	100	100	86	80
52 SNP	0.31	0.34	100	100	91	81
56 SNP	0.30	0.33	100	100	91	81
60 SNP	0.29	0.33	100	100	91	81

¹MAF = minor allele frequency.²PIC = polymorphic information content.³NA = not applicable.

when both mating and birth-calving records were used.

The average MAF ranged from 0.40 for the 24-SNP panel to 0.29 for the 60-SNP panel (column 2, Table 4). This is because of the way that the panels were formed with the addition of markers with progressively lower MAF. The PIC values of the markers are provided to compare levels of informativeness between the SNP and the SSR panel (column 3, Table 4). The SNP panels that contain 40 or more markers have greater matching abilities than the microsatellite panel irrespective of whether pedigree information was included or not (Table 4). Given that these SNP had an average MAF of 0.35, this scenario is comparable to the simulated predictions in Table 2. In the absence of mating and calving-birth date data, 72% of daughters (31 out of 43 animals) were parentage-matched with the 40-SNP panel. When mating data and calving-birth date alignments were used, this increased to 100%. The latter of these 2 types of pedigree information provided the greatest increase in matching ability. This result is consistent with the idea that calving-birth date alignments provide greater exclusion power by decreasing the number of possible false matches that could be considered. For example, assume each cow has on average 1.3 inseminations over the mating period. When mating records are included, each of the 43 daughters is tested against 159×1.3 dam-sire combinations compared with the 159×11 dam-sire combinations without mating records. However, when using the calving or birth date alignments, there were on average just 4 cows calving on the birth date of the daughter. Therefore only 4×11 dam-sire combinations were tested on average and only 4×1.3 when the mating records were added to the birth and calving date information.

The calving and birth date alignments were set initially to be on the same day. Relaxing the constraint to any birth or calving date match within ± 4 d only minimally decreased the parentage-assigning ability of the microsatellite panel and the 36- to 44-SNP panels (Table 5). For the more powerful of these panels (40- and 44-SNP), the matching ability remained sound even when the constraints were relaxed to ± 7 d; therefore, precision is considered preferable but not critical for calving and birth date records.

To estimate the number of SNP required to uniquely identify all daughters in the absence of on-farm information, further simulation was undertaken. Using the structure of the genotyped herd, SNP with a range of MAFs (0.2 to 0.5) were simulated for sires, dams, and daughters. The simulation involved several generations of breeding before reaching the final herd state. This was undertaken to mimic the inbreeding and co-relatedness of the animals in the real herd. Simulated data for 40 SNP with MAF of 0.25 gave power of exclusion results that were most similar to 40 real data SNP with MAF of 0.35. In the simulation, 19 out of 20 replicates had 100% resolution when 65 SNP were simulated. Therefore, we used these extrapolations to predict that 65 to 70 SNP (with MAF = 0.35) will be required for accurate parentage testing in this dairy herd when herd records are not used.

The test herd was considered to be smaller than the average New Zealand dairy herd, which consists of approximately 300 milking cows. Some herds are much bigger than this, containing a thousand or more cows. Simulations were carried out to see if unambiguous parentage assignment would be more difficult in larger herds. We found that there was a reduction in parentage matching ability, although this was minimal

Table 5. Percentage of daughters uniquely matched with sire and dam using real data for different criteria including calving and birth date matching with or without mating data for a variety of SNP panels and a 14-microsatellite panel

Panel	Number of days relaxed ¹							
	0	1	2	3	4	5	6	7
No mating data								
Microsatellite	93	93	93	93	93	91	88	84
36 SNP	100	100	93	93	88	84	79	74
40 SNP	100	100	95	95	95	93	93	93
44 SNP	100	100	98	98	98	98	98	98
Matings included								
Microsatellite	95	95	95	95	95	93	91	91
36 SNP	100	100	100	100	98	95	93	93
40 SNP	100	100	100	100	98	98	98	98
44 SNP	100	100	100	100	100	100	98	98

¹Relaxing refers to increasing, 1 d at a time, the allowed misalignment of offspring birth date with prospective dams' calving dates.

for more informative marker panels (data not shown). For example, 40 simulated SNP with $MAF = 0.25$ gave an exclusion rate of 97% for a 159-dam herd. This decreased to 94% for a 1,000-dam herd. However, when more informative panels were simulated (e.g., 40 SNP where $MAF = 0.4$), the exclusion rate was 100% for the 159-dam herd and 99.4% in the 1,000-dam herd.

DISCUSSION

We were able to uniquely assign parentage to only 33 of the 43 test herd daughters when using genotyping data (up to 72 SNP) alone. This is primarily from insufficient SNP with high MAF. For instance, only 40 SNP had MAF of 0.20 or above, and the SNP that were introduced to create the larger panels were progressively less informative. Glaubitz et al. (2003) used simulations to predict that 100 SNP with $MAF = 0.20$ would be required to be comparable to a panel of 16 to 20 independently segregating microsatellites. That is, 5 to 6 SNP were as informative as 1 SSR in their data. However, we have shown that we can obtain SNP panels whose average MAF is much greater than that (i.e., 40 SNP that average $MAF = 0.35$). The assumption that simulated data for our SSR panel had PIC scores of 0.67 on average was validated in the test herd, and by both simulation and use of real data, it was shown that with $MAF = 0.35$, forty SNP were more powerful for parentage mapping than the current panel of 14 SSR.

The test farm was assumed to be representative of the genetic structure found in other herds in New Zealand and possibly other countries. The PIC and MAF scores were not dissimilar to other New Zealand sample groups (including commercial SSR genotypes and crossbred trial dam frequencies). Also, the New Zealand herd is not considered to be any more inbred

than other herds, and perhaps less so. For example, although the effective population sizes of both Jerseys and Holstein-Friesians among the country's 4 million cow dairy herd are small (that is, 135 for Jerseys with similar values for Holstein-Friesians, De Roos et al., 2008), this is larger than Holstein-Friesian herds in other dairy-producing countries such as Australia, the Netherlands, and the United States (De Roos et al., 2008).

Although we predicted that 65 SNP with MAF of 0.35 are sufficient to provide accurate parentage in the absence of on-farm data in New Zealand dairy herds, the inclusion of on-farm data in the test herd circumvented the need to type more than 40 SNP. We showed that of the 2 types of on-farm data supplied, calving and birthing alignments increased the ability to uniquely match daughters to parents the most. This is likely because date alignments allow exclusion of a much larger proportion of the potential false parents in the herd. In the majority of cases in New Zealand, mating, calving, and birthdate data are routinely recorded and stored on a national dairy herd database. Therefore, the combination of on-farm and genotypic data for improved parentage exclusion is a pragmatic, realistic, and effective option.

CONCLUSIONS

Both the simulations and the test herd indicated that 40 SNP (where MAF average 0.35) would be at least as effective for parentage matching as the current 14-microsatellite panel. In addition, it will be possible to improve on this through identification of more markers with MAF above 40% in the populations of interest and also ensure that the chosen SNP are not located within 50 cM of each other. This will be particularly useful

for larger herds. The combination of genotypes from a 40-SNP panel with on-farm data will be sufficient to carry out whole-herd parentage testing in New Zealand dairy herds with a high degree of accuracy.

REFERENCES

- Amorim, A., and L. Pereira. 2005. Pros and cons in the use of SNPs in forensic kinship investigation: A comparative analysis with STRs. *Forensic Sci. Int.* 150:17–21.
- Anderson, E. C., and J. C. Garza. 2005. The power of single nucleotide polymorphisms for large-scale parentage analysis. *Genetics* 172:2567–2582.
- Banos, G., G. R. Wiggans, and R. L. Powell. 2000. Impact of paternity errors in cow identification on genetic evaluations and international comparisons. *J. Dairy Sci.* 84:2523–2529.
- Botstein, D., R. L. White, M. Skolnick, and R. W. Davis. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* 32:314–331.
- Davis, G. P., and S. K. DeNise. 1998. The impact of genetic markers on selection. *J. Anim. Sci.* 76:2331–2339.
- De Roos, A. P. W., B. J. Hayes, R. J. Spelman, and M. E. Goddard. 2008. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. *Genetics* 179:1503–1512.
- Dodds, K. G., M. L. Tate, J. C. McEwan, and A. M. Crawford. 1996. Exclusion probabilities for pedigree testing farm animals. *Theor. Appl. Genet.* 92:966–975.
- Dodds, K. G., M. L. Tate, and J. A. Sise. 2005. Genetic evaluation using parentage information from genetic markers. *J. Anim. Sci.* 83:2271–2279.
- Geldermann, H., U. Pieper, and W. E. Weber. 1986. Effect of misidentification on the estimation of breeding value and heritability in cattle. *J. Anim. Sci.* 63:1759–1768.
- Glaubitz, J. C., O. E. Rhodes, and J. A. Dewoody. 2003. Prospects for inferring pairwise relationships with single nucleotide polymorphisms. *Mol. Ecol.* 12:1039–1047.
- Glowatzki-Mullis, M. L., C. Gaillard, G. Wigger, and R. Fries. 1995. Microsatellite-based parentage control in cattle. *Anim. Genet.* 26:7–12.
- Heaton, M. P., G. P. Harhay, G. L. Bennett, R. T. Stone, W. M. Grosse, E. Casas, J. W. Keele, T. P. Smith, C. G. Chitko-McKown, and W. W. Laegreid. 2002. Selection and use of SNP markers for animal identification and paternity analysis in U.S. beef cattle. *Mamm. Genome* 13:272–281.
- Heyen, D. W., J. E. Beever, Y. Da, R. E. Evert, C. Green, S. R. Bates, J. S. Ziegle, and H. A. Lewin. 1997. Exclusion probabilities of 22 bovine microsatellite markers in fluorescent multiplexes for semiautomated parentage testing. *Anim. Genet.* 28:21–27.
- Israel, C., and J. I. Weller. 2000. Effect of misidentification on genetic gain and estimation of breeding value in dairy cattle populations. *J. Dairy Sci.* 83:181–187.
- Kashi, Y., E. Lipkin, A. Darvasi, A. Nave, Y. Gruenbaum, J. S. Beckmann, and M. Soller. 1990. Parentage identification in the bovine using “deoxyribonucleic acid fingerprints”. *J. Dairy Sci.* 73:3306–3311.
- Spelman, R. J. 2002. Utilisation of molecular information in dairy cattle breeding. *Proc. 7th World Congr. Genet. Appl. Livest. Prod., Montpellier, France. Communication No. 22-02.*
- Spelman, R. J., F. M. Miller, J. D. Hooper, M. Thielen, and D. J. Garrick. 2001. Experimental design for QTL trial involving New Zealand Friesian and Jersey breeds. *Proc. Assoc. Adv. Anim. Breed. Genet.* 14:393–396.
- Usha, A. P., S. P. Simpson, and J. L. Williams. 1995. Probability of random sire exclusion using microsatellite markers for parentage verification. *Anim. Genet.* 26:155–161.
- Werner, F. A., G. Durstewitz, F. A. Habermann, G. Thaller, W. Kramer, S. Kollers, J. Buitkamp, M. Georges, G. Brem, J. Mosner, and R. Fries. 2004. Detection and characterization of SNPs useful for identity control and parentage testing in major European dairy breeds. *Anim. Genet.* 35:44–49.