SNP selection for nationwide parentage verification and identification in beef and dairy cattle

M.C. McClure[1], J. McCarthy[1], P. Flynn[2], R. Weld[2], M. Keane[1], K. O'Connell[1], M.P. Mullen[3], S. Waters[3] and J.F. Kearney[1]

[1]Irish Cattle Breeding Federation, Bandon, Co. Cork, Ireland
2[2]Weatherbys Ireland, Johnstown, Co. Kildare, Ireland
[3]Animal & Grassland Research and Innovation Centre, Teagasc, Grange, Dunsany, Co. Meath, Ireland

Abstract

As parental verification in livestock species moves from microsatellite- to single nucleotide polymorphism (SNP)-based methods, the accuracy of pedigree verification will increase if robust methods and high quality SNP are used. In beef and dairy cattle, the international standard for SNP-based verification has been to use the ISAG100 or ISAG200 SNP set for Bos taurus breeds. We show that while these SNP sets do provide a higher level of accuracy than microsatellites, more SNP should be used for parentage verification and prediction, and some SNP should not be used due to genotyping quality. The Irish Cattle Breeding Federation (ICBF) is in the unique position of having access to both beef and dairy genotypes on Irish cattle, and through recent government schemes a large portion of ICBF genotypes come from commercial herds which have both purebred and crossbred animals. By analysing different SNP levels across beef and dairy cattle, we were able to determine that at a minimum >500 SNP are needed to consistently predict only one set of parents. If only the ISAG200 SNP are used for parentage prediction, then >1 sire or dam can be predicted at <1% misconcordance rate levels. Since if >1 parent can be predicted using the ISAG200 SNP set, then in theory it is also possible to validate the wrong parent for an animal. Recent analysis of SNP clustering patterns in Illumina BeadStudio software indicated that some SNP, including 3 from the ISAG200 panel, have clustering issues which only become apparent when thousands of samples from multiple breeds are analysed together. Minor allele frequency (MAF) and call rates (CR) were calculated for the Illumina LD base SNP across >180,000 genotyped animals that represent >20 breeds, from this 2 ISAG200 SNP with MAF <0.01 were identified. ICBF currently uses 800 SNP for parentage validation and prediction, which is comprised of 195 SNP from the ISAG200 panel and 605 SNP based on their minor allele frequency (>0.47) and SNP genotyping quality in >160,000 Irish beef and dairy animals. The use of a larger set of high quality SNP has resulted in a highly accurate pedigree validation and prediction regardless of the animals breed composition or pedigree status.

Introduction

Pedigree verification has been performed with DNA makers for almost 50 years in cattle. It was initially performed with the analysis of blood groups (Stormont, 1967), then microsatellite markers (MS) (Davis and DeNise, 1998), and is currently in a transition phase to single nucleotide polymorphisms (SNP). While the cost and availability of each new technology has hindered its initial use, the benefit of reducing pedigree errors cannot be ignored. A 10% pedigree error rate has been estimated to have a 6-13% effect on the inbreeding coefficient, 11-18% reduction on genetic trends in estimated breeding values (EBV), and a 2-3% loss in the response to selection (Visscher *et al.*, 2002, Banos *et al.*, 2001), and causes a downward basis in heritability estimates (Israel and Weller, 2000). While sire error rates have been estimated at 3-23% in different national Holstein-Friesian

populations missing sire rates (10-40%) can also be substantial, especially when using pedigree data from commercial herds (Sanders *et al.*, 2006, Harder *et al.*, 2005). Missing sire data can have a large effect on the response to genomic selection and the variance of breeding value (Harder *et al.*, 2005). While sire errors have a larger effect than missing sire information on genomic progress, their effect on genetic gain is additive (Sanders *et al.*, 2006).

As with all technology there is a need to balance the cost with its performance. For parentage validation the question has often been how many markers are needed to obtain a high probability that the parents, usually just the sire, are correct. For MS markers the international standard is the International Society of Animal Genetic (ISAG) panel of 12 markers (http://www.isag.us/Docs/CattleMMPTest_CT.pdf) although additional or different MS marker sets are used at times for higher parentage accuracy and in research settings (Van Eenennaam *et al.*, 2007, Fernández *et al.*, 2013, Sanders *et al.*, 2006). Given their limited variability (i.e. biallelic nature) and therefore inherent lower resolving power, more SNP are needed to provide the same parentage discriminating power of a MS panel. The current ISAG recommended panel of 100 SNP (ISAG100) has a parental exclusion probability (PE) of >0.999 and the ISAG200 panel of 200 SNP has a PE >0.9999999 (*www.isag.us/docs/Workshop report CMMPT 2014.pdf*). The ISAG100 panel is used by many groups world-wide for initial parentage validation and until recently by the Irish Cattle Breeding Federation (ICBF).

The ICBF had routinely been using 120 SNP for initial parentage validation which consisted of the ISAG100 and a subset of the additional SNP from the ISAG200 panel. For parentage prediction the ICBF had been using the set of 2,000 SNP which are common across all commercial Illumina SNP panels (3K, LD, 50K, and HD) (Illumina Inc., 2010, Illumina Inc., 2011b, Illumina Inc., 2011a, Matukumalli *et al.*, 2009). For parentage validation only 1 misconcordance was allowed (which is where for a SNP the sire is AA and the offspring is BB) to account for potential genotyping errors. For parentage prediction up to 10 misconcordances were allowed. This equates to a 1% and 0.5% misconcordance rate for the validation and prediction processes. Occasionally an animal would be presented for parentage validation and its listed sire would fail due to 2-3 misconcordances from the 120 SNP panel but then would be predicted as the sire from the 2,000 SNP panel. Upon investigation ICBF staff would find evidence that the sire had a genotype error for those 2-3 SNP where he was called homozygous but he should have been heterozygous based upon his other SNP validated progeny. The genotype errors did not prevent the sire being predicted due to the large number of SNP used. While 2,000 SNP used for parentage prediction was very accurate the computational time needed was a concern, especially as the number of animals being genotyped began to sharply rise in 2014. Given the issues generated with low SNP numbers and large SNP numbers being used for parentage validation and prediction a more optimal number of SNP for both validation and prediction was needed.

At the ICAR meeting in Berlin, Germany (May, 2014), the question was raised on how many SNP are used by various groups for parentage prediction. As no consensus was given, the ICAR- Parentage working group asked for work to be done in determining a minimum SNP panel needed for parentage prediction. ICBF agreed to take up this task as their database contains genotypes from multiple SNP panels, beef and dairy breeds, both commercial and purebred animals, and it dovetailed nicely with already initiated projects.

Material, methods, and results

Minimum SNP needed to generate 1 predicted sire

The ICBF database contains animals genotyped on the Illumina 3K, LD, 50K, and HD SNP chips along with those genotyped on the custom International Beef and Dairy (IDB) chip (Mullen *et al.*, 2013), with 56,147 genotyped animals being present in June, 2014. The animals genotyped represent a mixture of Irish beef, dairy, purebred, crossbred, pedigree, and commercial male and females from >20 Bos taurus breeds. As the Illumina 3K has not been commercially available since September, 2011 (personal communication, André Eggen, 23 Feb. 2015), and given its higher variability in genotyping accuracy (Wiggans *et al.*, 2012) 3K genotypes were not used for this study. Across multiple commercial and custom bovine SNP chips the 6,909 SNP from the LD beadchip are a common core of SNP. Minor allele frequencies (MAF) on the core LD SNP were calculated across all genotyped animals in the ICBF database and SNP were ranked on their MAF. Panels of SNP (200, 300, 350, 400, 450, 500, 600, 700, 800, 900, 1000, 1100, and 1250) were designed that contained the ISAG200 SNP and increasing numbers of top MAF ranked SNP.

To test how many sires are predicted at different SNP densities at the standard accepted 0.5-1% misconcordance rate 7,092 animals, which had previously had their listed sire SNP validated (420 also had the dam SNP validated), were run through sire prediction with 56,147 animals being in the reference population. The prediction program will return all sires and dams that had less than the set number of SNP misconcordances (<1.0% for <500 SNP, <0.5% for >600 SNP), and were >18 months older than the animal. The breed, pedigree status, herd, farm location, and AI status of the animal and predicted parents were not considered, only age difference to the animal and misconcordance counts (Table 1).

Table 1. Major breed percent in reference populations.

Breed	Jun-14	Mar-15
HOL	68.65	30.08
LIM	7.94	22.34
CHA	9.09	19.2
AAN	4.42	7.22
SIM	2.35	6.77
HER	3.23	4.73
BBL	1.01	2.75
MSH	0.06	1.57
SAL	0.04	0.92
JER	0.17	0.67
PAR	0.17	0.64
LMS	1.91	0.59
BAQ	0.04	0.59
AUB	1	0.53
PIE	0.55	0.17
MON	0.14	0.05
IRM	0	0.03
NWR	0.09	0.03
RED	0.01	0.03

The result from this analysis showed that using the ISAG200 SNP for sire prediction would result in >1 sire being predicted at <1.0% misconcordance levels for 4.2% of the animals (Figure 1). Only when >500 SNP were used would only 1 sire be predicted at <1.0% misconcordance level. To provide an extra 'buffer' the 800 SNP set was chosen for parentage validation and prediction, with the intention of revaluating this later on. This was carried out because a large portion of the animals were dairy or dairy crossbreds (68%) in the 52,909 data set and ICBF was preparing to genotype >120,000 commercial and purebred beef animals from ~35,000 herds via the 2014 Beef Genomics Scheme (*http://www.icbf.com/?p=1725*). Going forward at ICBF the same 800 SNP set will be used for sire

prediction and validation as in theory if one can predict >1 parent then one could also accidently validate the wrong sire if <500 SNP are used for initial validations (Figure 1 and Table 2).

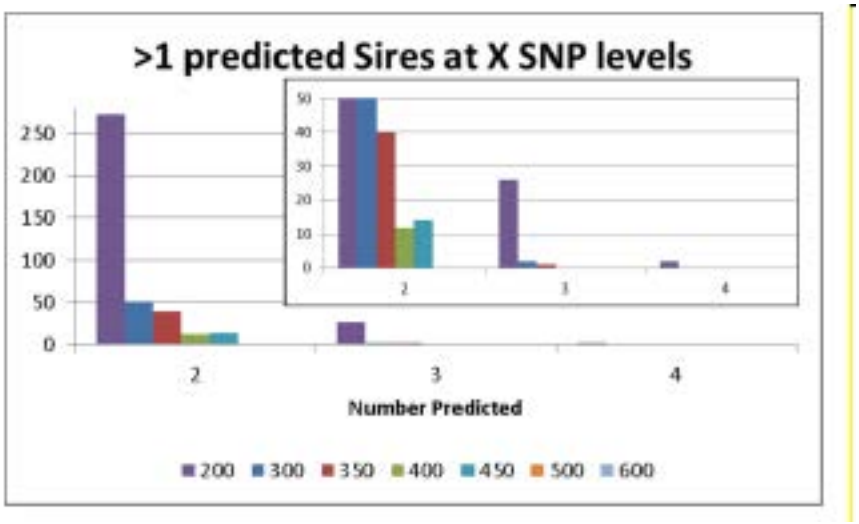

Figure 1. Count of predicted sires with <0.5-1% misconcordances at varying SNP density levels.

*Table 2. SNP excluded from Parentage analysis due to call rate (CR), Minor Allele Frequency (MAF), or Probe Clustering issues (**BOLD**).*

SNPID	Rs#	ISAG100	ISAG200	CR	MAF	Cluster
ARS-BFGL-NGS-103099	rs110896475	0	0	0.952	0.496	**1**
ARS-BFGL-NGS-112652	rs109504357	0	0	0.994	0.499	**1**
ARS-BFGL-NGS-11469	rs111032185	0	0	0.955	0.496	**1**
ARS-BFGL-NGS-118188	rs42156449	0	0	0.993	0.485	**1**
ARS-BFGL-NGS-3547	rs110206613	0	0	1	0.493	**1**
ARS-BFGL-NGS-43361	rs109934313	0	0	0.998	0.494	**1**
ARS-BFGL-NGS-53975	rs109456438	0	0	0.988	0.484	**1**
ARS-BFGL-NGS-62906	rs110146023	0	0	0.998	0.484	**1**
ARS-BFGL-NGS-66558	rs109817790	0	0	0.991	0.495	**1**
ARS-BFGL-NGS-76191	rs110846944	0	1	0.998	0.385	**1**
ARS-USMARC-PARENT-DQ786764-NO-RS	rs109943112	1	1	0.952	**<0.001**	-
ARS-USMARC-PARENT-DQ837645-RS29015870	rs29015870	1	1	**0.276**	0.389	**1**
ARS-USMARC-PARENT-EF034087-NO-RS	rs110665639	1	1	0.984	**0.005**	-
BTA-100621-NO-RS	rs41611675	0	1	0.928	0.452	**1**
BTB-00147175	rs43356919	0	0	0.99	0.491	**1**
BTB-01834338	rs42942345	0	0	0.999	0.498	**1**
HAPMAP47324-BTA-55159	rs41586638	0	0	0.984	0.496	**1**
UA-IFASA-9571	rs41659357	0	0	0.985	0.485	**1**

SNP quality control and MAF revaluation

By March 2015, >180,000 genotyped animals were in the ICBF database, and MAF for the core LD SNP were rechecked and an updated 800 SNP list was produced. This updated list allowed for the identification of SNP that were highly informative (MAF >0.45) across multiple breeds (Table 2). Theee ISAG100 SNP were noted to have either very low MAF (<0.001) or very low call rates (CR<0.3). Analysis of the Illumina BeadStudio files for the low CR SNP (ARS-USMARC-PARENT-DQ837645-RS29015870) revealed that it had clustering issues which were only apparent at high genotyping throughput rates, such as 4,000 samples a week. Analysis of all the ISAG200 and the next top 1000 MAF SNP revealed 15 SNP that have clustering issues. The clustering issues were not breed dependent.

A new set of 800 SNP for parentage validation and prediction was built using the ISAG200 SNP set as a base, minus the 5 listed in Table 1, and the next top 605 SNP based on MAF that did not have clustering issues (h*www.icbf.com/wp/wp-content/uploads/2013/07/ICBF_Parentage_SNP_Selection.csv*). All animals in the ICBF database had their parentage validation reanalysed with the new 800 SNP set to provide a consistent parentage analysis for all animals.

Misconcordance count distance between validated and failed sires

While most sire validated animals (>99%) in the ICBF database have only 0 or 1 misconcordance when using the 800 SNP set, <1% have 3 and 4 misconcordances. For failed animals >99% have >20 misconcordances, but a handful (N=10) have 13-20 misconcordances. LD core SNP analysis validated that all animals with 13-20 misconcordances from 800 SNP were true fails with >2% misconcordance rates.

To assess if <800 SNP could be used for parentage validation and prediction we analysed 8,626 animals that had >1 misconcordance count on their initial listed sire from the new 800 SNP set. SNP sets of 100, 200, 300, 400, 500, 600, and 700 were developed by using the ISAG 200 core set and then the top MAF SNP, the 100 SNP set was develop using the ISAG100 SNP set (*www.icbf.com/wp/wp-content/uploads/2013/07/ICBF_Parentage_SNP_Selection.csv*). Any SNP listed in Table 1 was not used.

While there is clear separation between validated and failed sires at 800 SNP the gap between these two outcomes rapidly narrows or disappears for smaller SNP sets (Figure 2). At the edges of the misconcordance count curves the gap between the number of counts

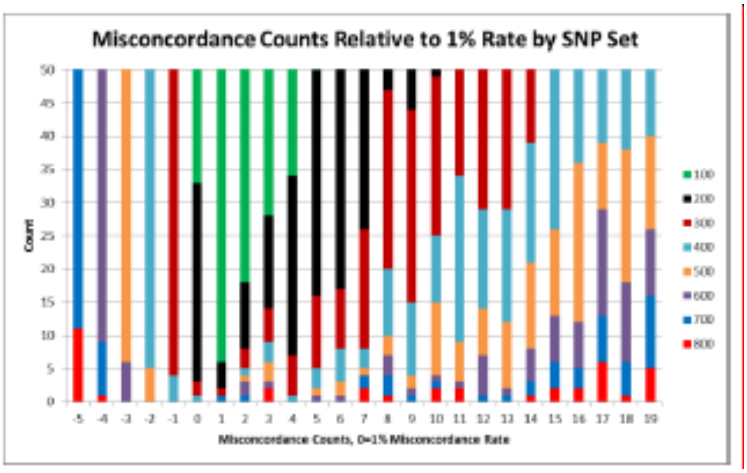

Figure 2. Misconcordance counts above and below the 1% misconcordance rate by SNP set on 8,626 animals with >1 misconcordance when their sire validation was analysed for 800 SNP. Image has been limited to focus on the counts around the 1% rate.

for failed and validated animals widens as more SNP are used. For 800 SNP there is a gap of 6 counts, for700 and 600 SNP its 4, 500 SNP is 3, and <400 it is <1. When you look at the number of SNP counts above the 1% misconcordance rate its only at 800 SNP that one reaches >2 SNP. For animals that passed sire validation at 800 SNP, 14 of them would have failed (>1% misconcordances) if only 100 SNP was used. More importantly for animals that failed sire validation at 800 SNP, 17 would have been validated at 100 SNP, 2 at 200 SNP, and 1 with the 400 SNP set.

Discussion and conclusion

While the ISAG100 and ISAG200 SNP panels do provide a good base for parentage validation via SNP they are not without their limitations. As shown here some of the ISAG SNP have genotyping problems due to their clustering patterns and some have very low MAF. More than 500 SNP are needed to predict only 1 sire when using all animals in your reference group as done at ICBF. If you restrict the prediction to only herd level fewer SNP could be used, but one will be hindered as this does not take into account potential errors from fence jumping breeding stock or mis-recorded semen straws. In Ireland the 800 SNP have proven very effective for initial validation and prediction. To date the only time >1 sire or dam has been predicted was due to a set of identical twin cows. While identical twins will have unique DNA methylation patterns (Kaminsky *et al.*, 2009) to date a method to use this for parentage validation in livestock has not been developed.

This study suggests that >500 SNP be used for parentage validation and that the SNP set described here works well for validation and prediction. ICBF has found that using the 800 SNP set for validation and prediction works well across multiple Bos taurus breeds and removes the possibility of accidently validating or failing a pedigree incorrectly. The number of SNP used by each laboratory, breed society, or national valuation centre will depend on cost and level of acceptable risk for a parentage error. As the cost of SNP genotyping decreases the value of having a near perfect pedigree will soon outweigh the cost of genotyping an animal with additional SNP.

List of references

Banos, G., Wiggans, G. R. & Powell, R. L. 2001. Impact of paternity errors in cow identification on genetic evaluations and international comparisons. Journal of Dairy Science, 84, 2523-9.

Davis, G. P. & Denise, S. K. 1998. The impact of genetic markers on selection. J Anim Sci, 76, 2331-9.

Fernández, M. E., Goszczynski, D. E., Lirón, J. P., Villegas-Castagnasso, E. E., Carino, M. H., Ripoli, M. V., Rogberg-Muñoz, A., Posik, D. M., Peral-García, P. & Giovambattista, G. 2013. Comparison of the effectiveness of microsatellites and SNP panels for genetic identification, traceability and assessment of parentage in an inbred Angus herd. Genetics and Molecular Biology, 36, 185-191.

Harder, B., Bennewitz, J., Reinsch, N., Mayer, M. & Kalm, E. 2005. Effect of missing sire information on genetic evaluation. Arch. Tierz, 48, 219-232.

Illumina INC. 2010. BovineHD Genotyping Beadchip [Online]. Available: *www.illumina.com/Documents/products/datasheets/datasheet_bovineHD.pdf* [Accessed 6/12/2011].

Illumina INC. 2011a. BovineLD Genotyping BeadChip [Online]. Available: *http://res.illumina.com/documents/products/datasheets/datasheet_bovineld.pdf* [Accessed August 31, 2012].

Illumina INC. 2011b. GoldenGate Bovine3K Genotyping BeadChip [Online]. Available: *http://www.illumina.com/Documents/products/datasheets/datasheet_bovine3K.pdf* [Accessed 5/16/2012].

Israel, C. & Weller, J.I. 2000. Effect of misidentification on genetic gain and estimation of breeding value in dairy cattle populations. Journal of Dairy Science, 83, 181-7.

Kaminsky, Z. A., Tang, T., Wang, S. C., Ptak, C., Oh, G. H., Wong, A. H., Feldcamp, L. A., Virtanen, C., Halfvarson, J., Tysk, C., Mcrae, A. F., Visscher, P. M., Montgomery, G. W., Gottesman, Ii, Martin, N. G. & Petronis, A. 2009. DNA methylation profiles in monozygotic and dizygotic twins. Nature genetics, 41, 240-5.

Matukumalli, L. K., Lawley, C. T., Schnabel, R. D., Taylor, J. F., Allan, M. F., Heaton, M. P., O'connell, J., Moore, S. S., Smith, T. P., Sonstegard, T. S. & Van Tassell, C. P. 2009. Development and characterization of a high density SNP genotyping assay for cattle. PLoS One, 4, e5350.

Mullen, M. P., Mcclure, M. C., Kearney, J. F., Waters, S. M., Weld, R., Flynn, P., Creevey, C. J., Cromie, A. R. & Berry, D. P. Year. Development of a custom SNP chip for dairy and beef cattle breeding, parentage, and research. In: INTERBULL, Augustt 23-25, 2013 2013 Nantes, France. Bullentin No. 47.

Sanders, K., Bennewitz, J. & Kalm, E. 2006. Wrong and missing sire information affects genetic gain in the Angeln dairy cattle population. Journal of dairy science, 89, 315-21.

Stormont, C. 1967. Contribution of blood typing to dairy science progress. Journal of dairy science, 50, 253-260.

Van Eenennaam, A. L., Weaber, R. L., Drake, D. J., Penedo, M. C., Quaas, R. L., Garrick, D. J. & Pollak, E. J. 2007. DNA-based paternity analysis and genetic evaluation in a large, commercial cattle ranch setting. J Anim Sci, 85, 3159-69.

Visscher, P., Woolliams, J., Smith, D. & Williams, J. 2002. Estimation of pedigree errors in the UK dairy population using microsatellite markers and the impact on selection. Journal of dairy science, 85, 2368-2375.

Wiggans, G. R., Cooper, T. A., Vanraden, P. M., Olson, K. M. & Tooker, M. E. 2012. Use of the Illumina Bovine3K BeadChip in dairy genomic evaluation. Journal of Dairy Science, 95, 1552-8