Research Article • DOI: 10.2478/ijafr-2019-0001 IJAFR • 58 • 2019 • 1-12

Irish Journal of Agricultural and Food Research

Low-density genotype panel for both parentage verification and discovery in a multi-breed sheep population

D.P. Berry[1†], N. McHugh[1], E. Wall[2], K. McDermott[2], A.C. O'Brien[1]

[1]Teagasc, Animal & Grassland Research and Innovation Centre, Moorepark, Fermoy, County Cork, Ireland
[2]Sheep Ireland, Highfield House, Shinagh, Bandon, County Cork, Ireland

Abstract

The generally low usage of artificial insemination and single-sire mating in sheep, compounded by mob lambing (and lambing outdoors), implies that parentage assignment in sheep is challenging. The objective here was to develop a low-density panel of single nucleotide polymorphisms (SNPs) for accurate parentage verification and discovery in sheep. Of particular interest was where SNP selection was limited to only a subset of chromosomes, thereby eliminating the ability to accurately impute genome-wide denser marker panels. Data used consisted of 10,933 candidate SNPs on 9,390 purebred sheep. These data consisted of 1,876 validated genotyped sire–offspring pairs and 2,784 validated genotyped dam–offspring pairs. The SNP panels developed consisted of 87 SNPs to 500 SNPs. Parentage verification and discovery were undertaken using 1) exclusion, based on the sharing of at least one allele between candidate parent–offspring pairs, and 2) a likelihood-based approach. Based on exclusion, allowing for one discordant offspring–parent genotype, a minimum of 350 SNPs was required when the goal was to unambiguously identify the true sire or dam from all possible candidates. Results suggest that, if selecting SNPs across the entire genome, a minimum of 250 carefully selected SNPs are required to ensure that the most likely selected parent (based on the likelihood approach) was, in fact, the true parent. If restricting the SNPs to just a subset of chromosomes, the recommendation is to use at least a 300-SNP panel from at least six chromosomes, with approximately an equal number of SNPs per chromosome.

Keywords

DNA • genomic • genotype • parent • single nucleotide polymorphism

Introduction

Inaccurate parentage recording is known to contribute to biased variance components (Van Vleck, 1970) and biased genetic evaluations (Israel and Weller, 2000; Banos *et al*., 2001), both of which subsequently affect genetic gain (Visscher *et al*., 2002). Unbiased estimates of coancestry among candidate mates require accurate pedigree recording. Therefore, tools to verify or discover parentage could be advantageous in facilitating optimised breeding programmes. It is nonetheless important that any tools to aid parentage assignment should be robust, technically accurate, inexpensive, and require minimal effort by producers and breeders. Moreover, the tools should ideally be such as not to stifle collaboration in (genomic) data exchange.

The exploitation of genomic technologies in animal production systems is intensifying, primarily in response to improved accuracy of prediction of true genetic merit when the genotype of an animal is incorporated into genetic evaluations (Hayes *et al*., 2009; Spelman *et al*., 2013). The cost of the commonly termed single nucleotide polymorphism (SNP) chips, consisting of genome-wide dense genomic markers, is nonetheless limiting large-scale adoption in some lower value species like sheep (Rupp *et al*., 2016). Alternatively, potentially lower cost genotyping technologies such as genotype-by-sequencing (Nielsen *et al*., 2011) can be used for genotyping, but such technologies are currently limited by the number of genomic markers. In the absence of routine genotyping for a large number of genome-wide genomic markers for use in genome-wide-enabled selection, lower cost genomic tools that verify or discover parentage can be useful to advance genetic gain through traditional quantitative genetic approaches. A balance must, however, be achieved between the cost of procuring a genotype and the accuracy of parentage assignment; both are likely affected by the number of genomic markers and how these markers are selected.

Several studies have developed panels of SNPs for parentage verification and discovery in sheep (Clarke *et al*., 2014; Heaton *et al*., 2014). Clarke *et al*. (2014) documented the ability of a specifically chosen 84-SNP panel to assign a ram to 99% of the progeny in their study; they, however, were not able to verify if the sire assigned was truly the correct sire, although in most instances, the probability of the next most likely candidate sire was very low.

†Corresponding author: Donagh P. Berry
E-mail: donagh.berry@teagasc.ie

The accuracy of parentage assignment is predicated on having access to the genotypes of the parent(s). Concerns are sometimes expressed on the sharing of genotypes, even at a low density for parentage assignment. This is because of the perceived ability to impute, albeit with relatively low accuracy, from low-density genotype to the high-density genotype (Judge *et al*., 2016), necessary to undertake genomic evaluations. Limiting the SNPs selected for parentage testing to only a proportion of the genome could put such concerns at ease since only a proportion of the genome could then be imputed to high density with reasonable accuracy. Such a strategy, however, may compromise the accuracy of parentage assignment and may actually require denser SNP panels. Nevertheless, the answer to this question is unknown. The objective of the present study was to quantify the accuracy of parentage verification and discovery *a posteriori* using purposely selected SNP panels of varying sizes. A particular focus of the present study was whether the SNPs selected could be limited to just a subset of chromosomes and what impact this would have on genome-wide imputation to higher density genotypes. Both exclusion and likelihood methods for parentage verification and assignment were evaluated, the latter being important where strong relationships exist among candidate parents, which often exist in sheep.

Materials and methods

Genotype data

Biallelic SNP genotype data were available from 12,844 individuals genotyped on either the Illumina OvineSNP50 BeadChip (n = 3,500 animals) or a custom Illumina Infinium panel (n = 9,344). The Illumina OvineSNP50 BeadChip contains 51,135 biallelic SNPs (excluding intensity-only SNPs). The custom Illumina Infinium panel consists of 14,940 SNPs. All animals had a call rate >95% on their respective panel. All available genotypes from both platforms were initially used to validate all parentage prior to the subsequent analyses. The genomic position of all SNPs was based on the OAR3.1 genome build. Subsequently, only the 10,933 autosomal SNPs, of known genomic location, common to both platforms were retained for further analysis.

Genotype quality control

The 88 autosomal International Sheep Genomics Consortium (ISGC) SNPs for parentage assignment were identified. Only 87 of these SNPs (i.e., DU417675_79 SNP not present) were part of the 10,933 SNPs common to both genotype platforms. All ISGC SNPs were forced on the parentage panels to be developed in the present study and were therefore not subjected to the subsequent SNP quality control measures. The exception was the panels limited to a selection of chromosomes; in that, the ISGC SNPs not on the selected chromosomes were not considered.

Only the 10,933 autosomal SNPs were considered for inclusion on the parentage genotype panels. A total of 141 SNPs with any opposing homozygous genotypes between the previously validated 1,876 sire–offspring pairs available in the present study were not considered further. No edit was applied based on genotypes of dam–offspring pairs as these were to be used in the subsequent validation analysis. A total of 720 SNPs with a call rate <99.5% was also not considered further. An additional one SNP was discarded, which had poor genotype cluster resolution dictated by a GenTrain score <0.55. The GenTrain score is a clustering algorithm propriety to Illumina Inc. and is useful as a measure of the distinction between genotype calls for a given SNP. Berry *et al*. (2016) evaluated the reproducibility of genotypes from the same SNP locus generated on technical duplicate samples for the same Illumina platform or different Illumina platforms. A total of 1,400 of the remaining SNPs in the present study, reported by Berry *et al*. (2016) to not have 100% concordance on the same Illumina genotype platform, were not considered further. Neither were an additional 1,278 SNPs that had <100% concordance on different Illumina genotype platforms were considered (Berry *et al*., 2016).

Only flock book-recorded animals from the main breeds of Belclare (n = 1,074), Charollais (n = 2,778), Suffolk (n = 1,799), Texel (n = 3,056), and Vendeen (n = 683) were retained. The minor allele frequency (MAF) of each SNP within each breed was calculated. SNPs with an MAF <0.20 in either of the five breeds or an MAF <0.25 across the entire population were discarded, as were SNPs that deviated ($P < 0.001$) from the Hardy–Weinberg equilibrium within breed. Following all edits, 2,379 candidate SNPs from 9,390 animals remained.

Development of parentage panel

In all, 39 genotype panels were generated (which included two panels already in existence but tested here for accuracy of parentage verification and assignment). A series of 10 genotype panels with 87 (ISGC SNPs only), 100, 150, 200, ..., 450, and 500 SNPs were generated; all chromosomes were represented in the panels with 100 to 500 SNPs. The mean polymorphic information content per breed for each panel developed was also estimated. For comparative purposes, an SNP panel that included just the 84 SNPs suggested by Clarke *et al*. (2014) for parentage assignment in New Zealand sheep was evaluated; 83 of the SNPs were in common with the ISGC panel. Similarly, a panel comprising the 163 SNPs identified by Heaton *et al*. (2014) to be useful for parentage testing across diverse breeds of sheep (50 in common with the ISGC panel) was evaluated; genotypes on only 159 of the proposed 163 SNPs were available for use.

The denser genotype panels (i.e., ≥100 SNPs) were generated by simply adding marginally informative SNPs (described later) to the immediately lesser dense panel. Hence, the SNPs included on a given genotype panel included all SNPs included in all lesser dense panels. The number of SNPs chosen per chromosome for each panel density was proportional to the length of the chromosome. The exceptions were the ISGC panel and the 100-SNP panel where the minimum number of SNPs per chromosome could not be less than the maximum number of SNP per chromosome on the ISGC panel. To quantify the impact of selecting SNPs from only a subset of chromosomes, selection of the most informative SNPs (described later) was re-undertaken but with the SNPs chosen from only the first 3, 6, 9, or 12 chromosomes. An equal number of SNPs per chromosome was chosen in this scenario except where the *modulo* operation of the number of required SNPs divided by the number of chromosomes was greater than zero; a single extra SNP was chosen from the first n chromosomes where n represented the modulus.

The within-breed pairwise linkage disequilibrium (LD) between all candidate SNPs in a given chromosome was calculated for each breed independently. The strongest of the within-breed LD values estimated for each of the pairwise SNP comparisons was retained as the measure of LD for later use in the SNP selection. The initial SNPs selected per chromosome for the developed panels were the 87 SNPs on the ISGC panel. In the 100-SNP panel selected from the entire genome, because SNPs on chromosomes 16 and 26 were not included in the ISGC panel, the SNP on both chromosomes with the greater MAF was initially selected. SNPs were sequentially selected thereafter based on the method documented by both Wellmann *et al*. (2013) and Judge *et al*. (2016) for selecting informative SNPs in the development of low-density genotype panels for imputation to higher density. The algorithm selects, on a chromosome-by-chromosome basis, marginally informative SNPs based on a combination of genomic location relative to the SNPs already selected, the extent of LD with the already selected SNPs, MAF, and call rate; such SNPs were deemed to be highly informative (i.e., combination of MAF concurrent with LD with already selected SNP) and of high quality (i.e., high call rate). An additional criterion included in the present study was that SNPs within 1 Mb of selected SNPs were immediately excluded from further consideration.

Parentage verification and discovery

Parentage verification and discovery were undertaken for sires and dams separately without considering the other parent. Because all genotyped parent–offspring pairs were already verified using at least 10,933 autosomal SNPs, a comparison could also be made between the assigned parent for each parentage SNP panel and the verified parent from the 10,933 SNPs. The recorded date of birth of all animals was also used when matching candidate parents to offspring, and candidate parents could not be less than 10 months older than their possible progeny.

Exclusion of candidate parents was dictated by the number of opposing homozygotes between the offspring and candidate parent. The number of informative SNPs per panel actually included in this exclusion analysis (i.e., SNPs where the genotype was called for both the offspring and parent and neither called genotype was heterozygous) was also calculated. A likelihood-based approach, as described in detail by Dodds *et al*. (2005), was also used for parentage assignment. The allele frequencies used in the likelihood-based calculations were the across-breed allele frequencies from the entire dataset of 9,390 animals. Because the candidate SNPs selected in the present study were already screened to have good concordance between panels, a genotype error rate of 0.01% was assumed in the likelihood-based approach; the exception was the evaluation of the 84-SNP panel proposed by Clarke *et al*. (2014) where an error rate of 0.5% was assumed in-line with what was used in the aforementioned study. The logarithm of the odds (LOD) score (i.e., natural logarithm of the likelihood of the candidate parent being the true parent divided by the likelihood of the candidate parent not being the true parent) was calculated (Dodds *et al*., 2005), and within parental gender, the parent with the greatest LOD score was assigned as the most likely. The delta statistic (Δ) representing the difference between the highest and second highest LOD scores was also calculated as a measure of confidence in the parentage assignment relative to the next most likely candidate. If the second most likely parent had a negative LOD score, then the delta statistic was the LOD score of the most likely parent. The probability of the most likely chosen parent being the true parent was also calculated as outlined by Dodds *et al*. (2005).

Of the 221 genotyped sires that had validated genotyped progeny included in the analysis, six of them also had genotyped full-sibs, 15 had a genotyped half-sib, and five sires had at least two half-sibs genotyped. Of the 1,876 genotyped dams with at least one genotyped progeny in the dataset, 104 had a genotyped full-sib and 777 had a genotyped half-sib. Hence, strong relationships existed in the dataset of candidate parents.

Imputation

To investigate the impact of imputation to the higher density genotype platform (i.e., OvineSNP50 BeadChip) from only SNPs selected for use in the parentage panel, the population of animals genotyped on the higher density panel (i.e., 49,386 autosomal SNPs) were stratified into a reference and validation population. The entire population consisted of 648 Belclare, 650 Charollais, 712 Suffolk, 418 Texel, and 619 Vendeen animals. The youngest 75 animals in each

breed were chosen as the validation animals with all other animals selected to be the reference animals. Within-breed imputation was undertaken, across the entire genome simultaneously, using FImpute V2.2 (Sargolzaei *et al*., 2014). Pedigree was included in the imputation process, although parental genotypes of validation animals were not available. All 49,386 autosomal SNP genotypes from the OvineSNP50 BeadChip were available in the reference population animals. Genotypes in the validation animals were masked to represent the different parentage panels being developed in the present study. Accuracy of imputation was based on the correlation of the entire 49,386 SNP genotypes of the validation animals from the imputation process versus their real genotypes.

Results

Selected SNP characteristics

The number of SNPs chosen per chromosome for each of the panel densities when considering all chromosomes is given in Appendix 1. The mean SNP MAF per breed for each panel density is shown in Figure 1. The mean MAF in the Charollais breed was always the greatest, while the mean MAF in the Suffolk breed was always the lowest (Figure 1). The mean MAF per panel tended to increase as panel density increased although this was not always the case (Figure 1). The mean distance between selected SNPs within chromosome increased from the ISGC panel to the panel with 200 SNPs but remained similar thereafter.

Summary statistics of the mean within-chromosome LD statistics per breed is shown in Figure 2. The LD between SNPs, within chromosome, was, on average, weak. Similar to the mean MAF, the mean polymorphic information content per panel was greatest for the Charollais and lowest for the Suffolk. The overall panel mean polymorphic information content increased at a diminishing rate with increasing panel density and was greater for the Charollais and least for the Suffolk.

Accuracy of parentage verification using exclusion based on SNPs across the entire genome

The number of discordant genotypes per parentage panel between each animal and its validated sire and dam (using the 10,933 SNPs) is given in Table 1. With the exception of the ISGC SNPs, any SNP that displayed any level of discordance between an offspring and its validated sire was not considered in the present study; this edit, however, was not included for dam–offspring comparisons. A particular point to note is that when parentage testing is undertaken based on exclusion of individuals with opposing homozygous genotypes, the number of SNPs used in such an approach is not necessarily the total number of SNPs on the panel. This is because SNPs with a heterozygous genotype in either of the two animals are not considered, nor are SNPs where the genotype was not called in either animal. In fact, based on the 87-SNP ISGC panel, the mean number of SNPs used to exclude male

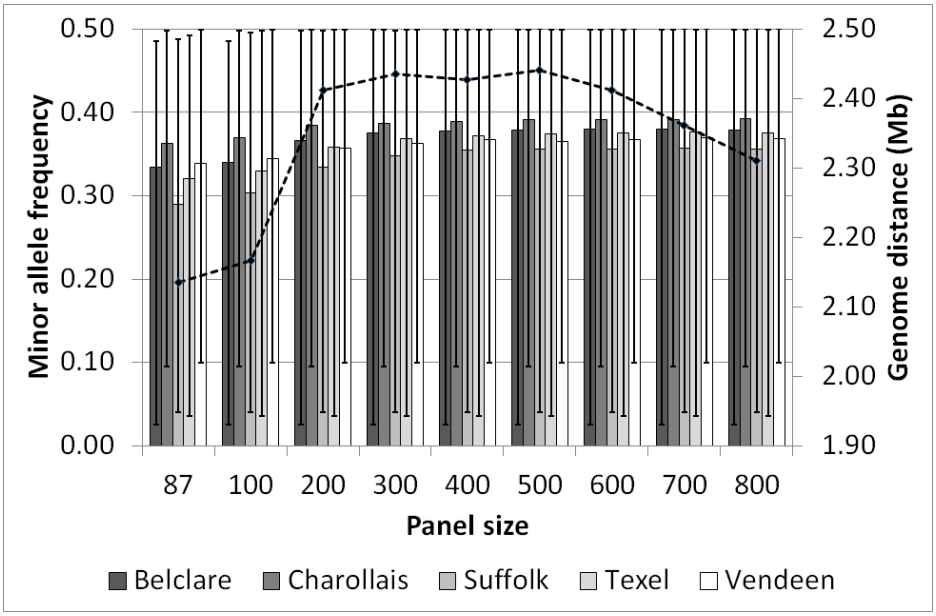

Figure 1. Mean MAF per breed (minimum and maximum MAF represented by standard error bars) for each breed represented by the histograms (from left to right: Belclare, Charollais, Suffolk, Texel, Vendeen), as well as the mean number of megabases between SNPs within chromosome represented in the broken line, for each SNP panel. MAF = minor allele frequency, SNP = single nucleotide polymorphism.

animals as potential sires was just 32 with some comparisons being based on only 16 SNPs (Table 1). When using an SNP panel of 150 SNPs, on average, only 51 SNPs were used in the exclusion of sire–offspring (Table 1).

Irrespective of the parentage panel size, of the 1,876 validated offspring–sire comparisons made, only five such comparisons had a discordant genotype (was one discordant SNP per comparison), and these discordances all appeared in just three SNPs, namely DU425259_620 (three discordances), DU518561_359 (one discordance), and CZ920359_258 (one discordance); all three SNPs were ISGC SNPs since no restriction was imposed on the presence of discordance between true sire–offspring pairs for these SNPs. On closer examination of the Log-R ratios of the genotypes, the sires appear to have been carrying only one copy of the DU425259_620 and DU518561_359 SNPs, thus causing the apparent conflict; both sire and offspring appear to have been carrying two alleles for the CZ920359_258 SNP. When comparing the 2,784 true dam–offspring genotypes for the ISGC SNP panel, there were five discordant genotypes (two

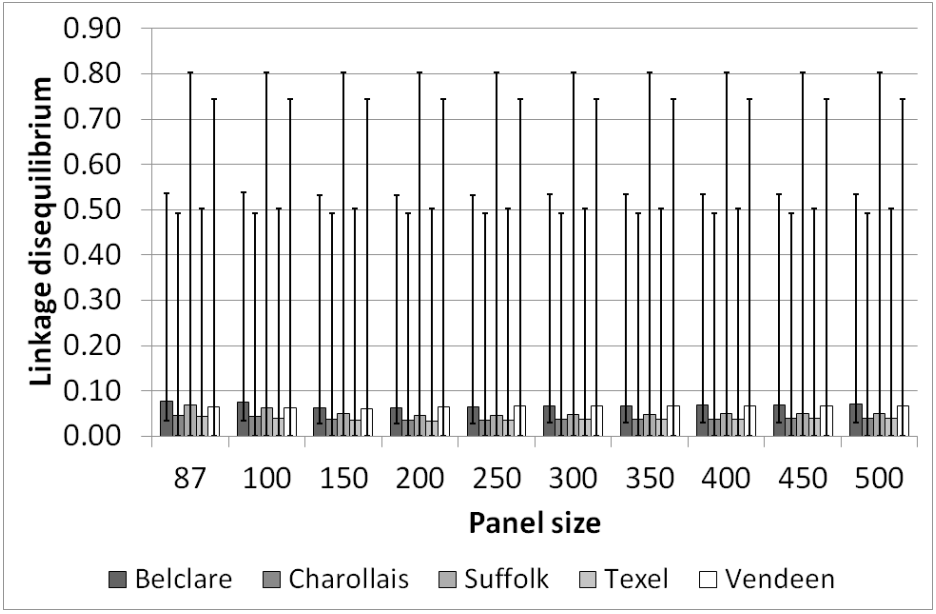

Figure 2. Mean within-chromosome linkage disequilibrium (minimum and maximum linkage disequilibrium represented by standard error bars) among selected SNPs for each panel density by breed represented from left to right: Belclare, Charollais, Suffolk, Texel, and Vendeen. SNP = single nucleotide polymorphism.

Table 1. Mean (minimum, maximum) additional number of single nucleotide polymorphism (SNPs) tested (i.e., neither the parent nor offspring had a heterozygote genotype or missing genotype) and additional number of discordant SNPs for the 1,876 validated true sire–offspring pairs (sires) and 2,784 validated true dam–offspring pairs for each increase in panel density size

SNPs	Sires		Dams	
	SNPs tested (min, max)	Discordant SNPs	SNPs tested (min, max)	Discordant SNPs
≤87	32 (16, 48)	1	32 (15, 51)	1
88–100	4 (1, 9)	0	4 (1, 10)	0
101–150	15 (4, 26)	0	15 (4, 27)	0
151–200	16 (7, 30)	0	16 (6, 28)	0
201–250	15 (5, 26)	0	15 (5, 27)	0
251–300	15 (5, 28)	0	15 (6, 25)	0
301–350	15 (6, 27)	0	15 (4, 27)	0
351–400	16 (6, 27)	0	15 (5, 28)	0
401–450	15 (4, 27)	0	16 (6, 28)	1
451–500	16 (5, 28)	0	16 (5, 27)	1

for CZ920359_258, one for DU364754_308, and two for DU518561_359). Similar to what was observed with the sires, with the exception of the CZ920359_258 SNP, the dams appear to have been missing one allele of the two remaining SNPs, thus contributing to the apparent discordance. In the higher density parentage panels, opposing genotypes existed between validated dam–offspring pairs for only four SNPs, namely, OAR18_54116518, OAR1_162474027, OAR20_14493721, and OAR7_74434542, but these SNPs only existed on panel densities >400 SNPs.

When the number of SNPs evaluated was 87 (i.e., the 88-SNP ISGC panel with one SNP missing) and one discordant SNP between parent–offspring was allowed, there were 581 potential sires identified additional to the actual true sire; this reduced to 190, 6, and 0 when the number of SNPs was 100, 150, and 200, respectively. The number of possible dams identified additional to the true dam (i.e., when one discordant genotype between dam and offspring pairs was allowed) for panels with 87, 100, 150, 200, 250, 300, and 350 SNPs was 13,386, 4,140, 137, 13, 3, 2, and 0, respectively. Therefore, when the panel consisted of 350 SNPs with one discordant SNP allowed, all the correct sires and dam were uniquely identified as the true parents.

When considering the 84 SNPs proposed by Clarke *et al*. (2014), a maximum of one discordant SNP per validated offspring–parent pair was detected, and this occurred in only 10 instances (five out of the 2,784 dam–offspring pairs had one discordant SNP and a further five out of the 1,876 sire–offspring pairs had one discordant SNP). The discordance between genotypes of the validated parent–offspring pairs was not, however, just limited to one SNP, but in fact, one discordance occurred for the DU364754_308 SNP on OAR9, while three discordant genotypes occurred for each of the three SNPs, DU518561_359 (OAR1), DU425259_620 (OAR3), and CZ920359_258 (OAR24).

Of the 159 SNPs proposed by Heaton *et al*. (2014), opposing homozygous genotypes between validated true sire–offspring pairs occurred for only two validated sire–offspring pairs, and they each had only one discordant SNP (CZ920359_258 SNP on OAR24 or DU518561_359 SNP on OAR 1); opposing genotypes between validated dam–offspring pairs existed for only four pairs, and each had only one discordant SNP (either CZ920359_258 SNP on OAR24, which had two discordant events, or DU518561_359 SNP on OAR 1, which also accounted for two discordant events).

Accuracy of parentage assignment using the likelihood-based approach with SNPs across the entire genome

The number of times the true (i.e., validated) sire and dam of an individual was chosen to be the most likely respective parent, as well as the associated minimum calculated probability, is given in Table 2 for the different panels. Also included in Table 2 is the difference in likelihood between the incorrectly predicted parent and the true parent. Based on the 87-SNP ISGC panel, 1,868 of the known 1,876 true sires of genotyped progeny were actually ranked as the most likely sire equating to a specificity of 0.004; the eight true sires, which included representatives from all five breeds, that were not ranked as the most likely were, in fact, ranked the second most likely with a probability of being the sire of the individual lying between 0.09 and 0.47. Of the 1,868 true sires ranked as most likely, the probability of correct assignment ranged from 0.67 to 1.00; the mean (minimum, maximum) delta statistic for these animals was 19.4 (0.7,

Table 2. For each single nucleotide polymorphism (SNP) panel, the number of 1,876 true sire–offspring pairs or 2,784 true dam-offspring pairs that were deemed to be the most likely pair (most likely), the minimum of the probability of any of the pairs being true (probability), and the difference in log-likelihood (delta) between the most likely selected pairing and the true pairing

Panel	Sires			Dams		
	Most likely	Probability	Delta	Most likely	Probability	Delta
≤87	1,868	0.669	0.722 (0.116, 2.301)	2,670	0.396	1.865 (0.216, 8.964)
≤100	1,873	0.648	2.088 (0.305, 2.372)	2,725	0.507	2.341 (0.105, 12.456)
≤150	1,876	0.999		2,781	0.599	3.591 (2.409, 6.642)
≤200	1,876	1.000		2,782	0.992	4.425 (2.702, 6.148)
≤250	1,876	1.000		2,784	0.990	
≤300	1,876	1.000		2,784	0.780	
≤350	1,876	1.000		2,784	1.000	
≤400	1,876	1.000		2,784	1.000	
≤450	1,876	1.000		2,784	1.000	
≤500	1,876	1.000		2,784	1.000	

38.3), although the delta statistic was always >2 when the probability was >0.95. For dams, when based on the 87-SNP panel, 95.9% (i.e., 2,670 out of known true 2,784 dam–offspring pairs) of the true dams were ranked the most likely dam with an associated probability of being the most likely dam varying from 0.40 to 1.00 with a corresponding mean (minimum, maximum) delta statistic of 8.9 (0.02, 39.3). Of the 114 true dams that were not ranked the most likely, their respective probability of being the dam of the individual in question varied from 1×10^{-4} to 0.49.

For the eight sires that were incorrectly ranked to be the most likely sire, their probability of being the true sire varied from 0.52 to 0.90; the corresponding values for the incorrectly assigned dams varied from 0.35 to 0.99. All results were similar, although slightly better than those obtained when only the 84 SNPs proposed by Clarke *et al*. (2014) were used. For example, the number of true sires not ranked as the most likely increased from eight to 14, while the corresponding values for dams increased from 114 to 213. Once the SNP panel included 150 SNPs, the most likely sire determined based on the associated likelihood was always the true sire (Table 2) and the delta statistic for these selected sires was always >7.45; such an outcome did not occur in dams until the panel included 300 SNPs (Table 2) where the delta statistic for these selected dams varied from 0.08 to 231. With a 300-SNP panel, the delta statistic for all true sires was >30.

Based on the likelihood approach of the 159 SNPs proposed by Heaton *et al*. (2014), the true sire was assigned the most likely parent 99.84% (i.e., 1,873 of the 1 876 sire–offspring pairs) of the time, with the probability value of the true sire varying from 0.10 to 0.31 when it was not chosen. Moreover, when the sire chosen was the true sire, the probability that the sire assigned was the most likely sire was as low as 0.58. For dams, 96.6% (2,690 from 2,784 dam–offspring pairs) of the dams assigned as the most likely dam was the true dam with the probability of the true dam not being the most likely dam in these 94 cases varying from 0.001 to 0.492; even at that, when the dam assigned was in fact the true dam, the probability of this assignment being correct was as low as 0.49. Therefore, although the 159 SNPs alone assigned the vast majority of the parents correctly, a small proportion was incorrectly assigned and some of those that were correctly assigned were not assigned with a very high confidence (i.e., probability). For the three sires that were incorrectly ranked to most likely sire, their probability of being the correct sire varied from 0.69 to 0.90; the corresponding values for the incorrectly assigned dams varied from 0.35 to 0.99. Of the three sires that were incorrectly assigned, the first one was a half-sib to the animal (born 11 months earlier), the second was the maternal grandsire to the animal, and the third was the maternal grand-dam's sire.

Parentage verification and discovery using SNPs from a subset of chromosomes

Based on the results of the different SNP panel densities across the entire genome, only a 300-SNP panel was developed. When the 300-SNP panel originated from just the first three, six, nine, or 12 chromosomes and just one discordant genotype was allowed between sire and offspring, only one unique sire was allocated to each individual based on exclusion, and this allocated sire was the true sire. When the 300-SNP panel originated from just the first three, six, nine, or 12 chromosomes, other than the true dam, a further nine, three, four, and two possible candidate dams were identified, respectively, additional to the true dams. Based on the likelihood approach, no matter how many chromosomes were represented in the 300-SNP panel, the correct sire was always the most likely sire identified, each with a probability >99%. In contrast for dams, when the 300-SNP panel was limited to just the first three chromosomes, nine of the 2,784 validated dams were not selected as the most likely dams and five of the validated dams identified as the most likely had a probability <99%. When SNPs were selected from the first six chromosomes, all the validated true dams were selected to be the most likely with only two of the properly identified dams having a probability of <99% (i.e., 52% and 75% probability of being the dam).

Imputation

Only the 300-SNP parentage panel was chosen for evaluating the accuracy of imputation to higher density; the 300 SNPs were those selected across the entire genome (i.e., all 26 chromosomes) or just the first three, six, nine, or 12 chromosomes. Little difference in mean imputation accuracy per SNP existed irrespective of how many chromosomes were represented with the correlation between the true and imputed genotypes varying from 0.57 to 0.58 across all five panels investigated (i.e., across the first three, six, nine, 12, or 26 chromosomes). The mean of the within-animal correlation between the true and imputed genotypes also varied from 0.57 to 0.58 across the five panels investigated.

Imputation accuracy per chromosome, however, varied per panel. When the 300 SNPs were chosen from just the first three chromosomes, the mean correlation between true and imputed genotypes of these chromosomes was 0.72 compared to a mean of 0.50 for the remaining 23 chromosomes. A similar trend was observed for the other SNP panels limited to a subset of chromosomes with a higher accuracy of imputation for the chromosomes with SNPs selected. When the 300 SNPs were chosen across all 26 chromosomes, the mean correlation between true and imputed genotypes of all chromosomes was 0.58. Thus, relative to when SNPs were chosen across the whole genome, the relatively poor imputation accuracy achieved on chromosomes with no SNP chosen was compensated by the greater imputation accuracy on chromosomes where more SNPs were chosen.

Discussion

Although developments in genomic solutions for genotyping is reducing the cost of acquiring many thousands of SNPs on individuals, there is still interest in ultra-low-density bespoke genotype panels for parentage testing (and possibly screening for polymorphisms in genes of major effect; Galloway *et al*., 2000; Nicol *et al*., 2009; Silva *et al*., 2011). Moreover, when assigning parents, the number of possible parent–offspring combinations can be extensive in large populations, and if high-density SNP panels are used, the computation can become unwieldy. The process of parentage assignment can be made more feasible if based on fewer SNPs, but the number of SNPs used should be small enough to be computationally possible in a feasible time period, yet also sufficiently accurate in the assignment. Moreover, the (international) transfer of either animals themselves or semen and embryos implies that the genotype of the possible parents may not always be freely available to the importer. Unease sometimes exists in the exchange of full genotype information among parties because of its potential use in genomic evaluations providing a possible advantage to the importing country by having improved genomic predictions. Exchange of a smaller subset of SNPs could help ameliorate such concerns, although earlier evidence suggests that it is possible to impute, with some degree of accuracy, a higher density genotype panel from a lower density genotype panel (Berry *et al*., 2014; Judge *et al*., 2016). The motivation therefore for the present study was to identify a small list of SNPs that could be used for accurate parentage assignment. A second motivation was to quantify if restricting these SNPs to a subset of chromosomes was useful for accurate parentage verification and discovery.

The importance of parentage assignment is well established in terms of more precise estimates of genetic parameters (Van Vleck, 1970) and thus the proper partitioning of variances into their causal components, as well as calculated expected responses to selection. Accurate parentage assignment, of course, also impacts the precision of genetic evaluations (Israel and Weller, 2000; Banos *et al*., 2001), although the impact of the extent of parentage mis-identification is a function of the heritability of the trait in question and also the quantity of available data for a given animal (Visscher *et al*., 2002). Numerous non-genomic approaches have been proposed as strategies to improve the assignment of parents to offspring. DNA technologies, however, if properly applied (e.g., the correct animals sampled, no sample mix-up, good quality genotypes on a sufficient number of DNA markers) can be extremely accurate. The rapid uptake of genome-wide-enabled selection in most species has revolutionised most livestock breeding sectors, and parentage verification or discovery is now routinely undertaken within the quality control steps prior to genomic evaluations. When attempting to verify the recorded parentage in the Irish sheep national database using up to 11,129 SNPs, Berry *et al*. (2016) reported sire-to-offspring and dam-to-offspring errors of 10.0% and 7.6%, respectively, which is similar to those reported in cattle (10.18% to 13.28%; Purfield *et al*., 2016) and goats (8.4% to 14.6%; Bolormaa *et al*., 2008). Therefore, parentage errors do exist and do affect genetic gain. Low-cost tools to correct recorded parentage could therefore advance genetic gain. Moreover, many commercial flocks worldwide do not record parentage, thereby not contributing to improving the accuracy of genetic evaluations. Having a robust, technically accurate, and inexpensive tool to assign parentage could further augment the rates of genetic gain.

International society of sheep genomics SNP panel

Of the 87-SNP ISGC panel, discordances existed between validated parent and offspring pairs for four SNPs (DU425259_620, DU518561_359, CZ920359_258, and DU364754_308). On further examination, the GenTrain score of CZ920359_258 and DU364754_308 for the entire dataset was 0.54 and 0.60, respectively, while that for both DU425259_620 and DU518561_359 was ≥0.87; the mean (minimum, maximum) GenTrain score of the remaining 83 SNPs was 0.87 (0.69, 0.95). A poor GenTrain score represents poor clustering of the genotypes, and SNPs with a GenTrain score of <0.55 are sometimes discarded prior to downstream genomic analyses (Zhao *et al*., 2015; Judge *et al*., 2016; Richardson *et al*., 2016). The call rate of DU364754_308 was also poor (77%), while that of CZ920359_258 was 97%; the call rates of both DU425259_620 and DU518561_359 were >99.5%. The data utilised by Berry *et al*. (2016) when comparing concordance rate between ovine SNP genotypes called on the same Illumina platform, called on different Illumina platforms and called on both an Illumina and Affymetrix platform was used to determine the concordance rate of these 4 SNPs. The concordance rate was 100% for all SNPs when comparing the OvineSNP50 BeadChip data generated in duplicate on 25 sheep (http://www.illumina.com/documents/products/datasheets/datasheet_ovinesnp50.pdf). When comparing genotypes from the Illumina OvineSNP50 BeadChip and a Custom Illumina BeadChip (developed in collaboration with the ISGC), different genotypes were obtained for one animal (out of 21 animals with a called genotype for the SNP on both panels) and two animals (out of 21 animals with genotypes for the SNP on both panels) for the DU425259_620 SNP and the DU364754_308 SNP, respectively; 100% concordance rate was observed for the DU518561_359 and CZ920359_258 SNPs.

The mean Affymetrix confidence score across the 84 animals used by Berry *et al*. (2016) was the greatest (i.e., worse)

at 0.025 for the DU425259_620 SNP compared to <0.005 for the other three SNPs. Moreover, the DU425259_620 was classified as "CallRateBelowThreshold" based on the Affymetrix probe set scoring with a call rate of 93% based on the 84 ovine samples used by Berry *et al*. (2016). Results therefore suggest that possibly four SNPs (i.e., DU425259_620, DU518561_359, CZ920359_258, and DU364754_308) could be removed from the recommended ISGC panel, or at the very least, discordance between parent and offspring genotypes tolerated more for these SNPs.

Practical implications of the study

The present study consisted of 4,660 parent–offspring pairs (i.e., 1,876 sire–offspring and 2,784 dam–offspring) validated using at least 10,933 SNPs, thereby providing a comprehensive database for validating the generated ultra-low-density SNP panels. Moreover, the presence of several closely related individuals as candidate parents also provided a very useful and practical dataset to evaluate the differentiation power of the various panels. Hill *et al*. (2008), also using a likelihood-based approach for parentage discovery, documented that the success rate of parentage assignment was dependent on both the number of potential candidate parents and the closeness of relatedness between the candidate parents and the focal animal. Double *et al*. (1997) modified their equation for parental exclusion to account for candidate sires that were related. The parental exclusion power of an SNP panel was less when related candidate sires were considered. In the present study, after considering only male animals that were at least 10 months of age when the focal animal was born, a total of 1,630,842 possible sire–offspring combinations were considered; for candidate dam–offspring combinations, this number was 33,722,321. Proportionally, therefore, the level of errors in parentage assignment in the present study with the developed SNP panels was miniscule. The strong editing criteria imposed on the candidate SNPs, including exceptionally high call rates and excellent reproducibility within and between platforms, ensure that accurate genotypes will be generated for all of the SNPs on the panel, thereby minimising redundancy. This also facilitated the use of a low genotype error rate in the likelihood-based parentage assignment approach. Likelihood approaches should be superior if genotyping errors exist but can also be more informative if the actual parent is not genotyped. Moreover, unlike the exclusion method, the likelihood approaches rank candidate parents in an objective way. For example, when using a 300-SNP panel, the delta statistic of the true sires, which were all ranked the most likely, was >30. This means that the allocated sire was 1×10^{13} times (i.e., e^{30}) more likely to be the sire than the next most likely candidate sire; the delta statistic for dams was not as pronounced even with a 500-SNP panel. Nonetheless, despite the strong family relationships among candidate sires that existed in the population, the strength exemplified in the likelihood-based statistic for sires provides a large degree of confidence in the parentage assignment.

The impact of parentage errors on genetic gain is a function

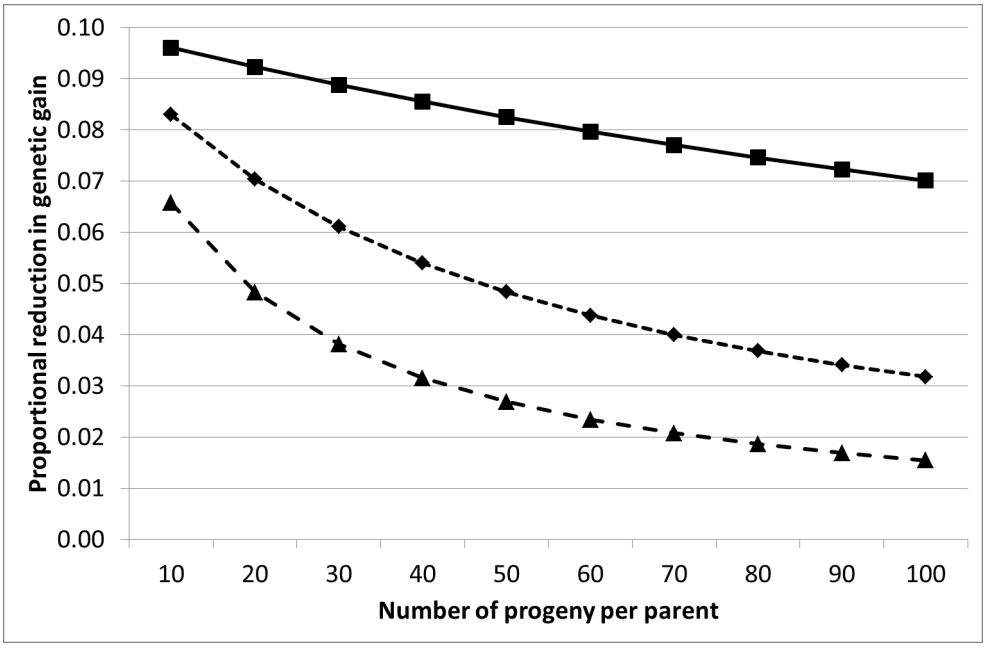

Figure 3. Proportional reduction in genetic gain (relative to no parentage errors) for a 10% sire parentage error for a trait with a heritability of 0.02 (squares), 0.10 (diamonds), and 0.25 (triangles) for different numbers of progeny per sire.

of the heritability of the trait(s), in tandem with the number of progeny per individual (Visscher *et al*., 2002). The impact of lack of recorded parentage is expected to follow a similar trend and has an additive effect on genetic gain over and above pedigree errors (Sanders *et al*., 2006). Several traits constitute sheep breeding objectives (Santos *et al*., 2015), including those in Ireland that differ in estimated heritability. For example, the heritability of lamb survival is up to 0.09 (McHugh *et al*., 2017), the heritability of lameness is 0.10 (O'Brien *et al*., 2017), and the direct heritability of live weight is 0.25 (O'Brien *et al*., 2017). The proportional reduction in genetic gain for these three example traits with a 10% sire error (Berry *et al*., 2018) is illustrated in Figure 3. When progeny group size was <100, the impact of observed population norms for parentage errors can be quite substantial, and thus, the benefit of correcting such errors can accelerate genetic gain.

Phenotypic data included in genetic evaluations for many sheep populations are generally limited to selected flocks. In Ireland, these are almost exclusively purebred seedstock breeders. Having a low-cost tool to assign parentage would not only correct errors in these populations, concomitant with a reduction in labour requirements for parentage assignment during the busy lambing season, but could also expand the population size of phenotyped individuals through the recruitment of additional flocks. The knock-on effect (assuming no impact on heritability estimates) is more accurate genetic evaluations. Including more commercial flock data into genetic evaluations could also alleviate concerns of genotype-by-environment interactions where a large proportion of the phenotypic data used in sheep genetic evaluations are perceived to originate from flocks that may not always be reflective of commercial reality.

Assuming that the accelerated genetic gain achieved through parentage assignment outweighs the cost of such a large-scale genotyping strategy, there will be a return on investment. Moreover, the cost of genotyping is only borne by few, while the benefit is realised by many. Based on a very simple calculation, assuming an average lamb carcass weight of 20 kg at a value of €4/kg and a ratio of seedstock adults to commercial adults of 1:10, every incremental €10 in the price of genotyping (i.e., from sample procurement to genotyping) of all seedstock newborn (after all adults have been genotyped) would cost approximately €0.013 per kg carcass sold.

Conclusions

High accuracy of parentage verification and discovery can be achieved with low-density panels, but near-perfect discriminatory power would require approximately 250 to 300 carefully selected SNPs spread across at least six chromosomes. Limiting the selected SNPs to just a subset of chromosomes should not hinder international exchange of genotypes, and agreeing an internationally accepted standard set of SNPs informative across populations should facilitate the development of an international repository where bodies can submit the genotypes of identified animals for use in parentage verification/discovery at a global level. Irrespective, imputation accuracy to high density using 300 SNPs, even dispersed across the entire genome, was poor. Results suggest that, if using the ISGC panel of SNPs, leniency in the concordance rate should be given, in particular for the DU425259_620, DU518561_359, CZ920359_258, and DU364754_308 SNPs, and in particular, the Log-R ratio of discordances should be examined to ensure that any apparent discrepancy was not due to a deletion in that SNP.

Acknowledgements

Financial support from the Irish Department of Agriculture, Food and the Marine Stimulus Research Fund project OVIGEN is gratefully acknowledged.

References

Banos, G., Wiggans, G.R. and Powell, R.L. 2001. Impact of paternity errors in cow identification on genetic evaluations and international comparisons. *Journal of Dairy Science* **84**: 2523–2529.

Berry, D.P., McClure, M.C. and Mullen, M.P. 2014. Within-and across-breed imputation of high-density genotypes in dairy and beef cattle from medium and low-density genotypes. *Journal of Animal Breeding and Genetics* **131**: 165–172.

Berry, D.P., McHugh, N., Randles, S., Wall, E., McDermott, K., Sargolzaei, M. and O'Brien A.C. 2018. Imputation of non-genotyped sheep from the genotypes of their mates and resulting progeny. *Animal: An International Journal of Animal Bioscience* **12**: 191–198.

Berry, D.P., O'Brien, A., Wall, E., McDermott, K., Randles, S., Flynn, P., Park, S., Grose, J., Weld, R. and McHugh, N. 2016. Inter-and intra-reproducibility of genotypes from sheep technical replicates on Illumina and Affymetrix platforms. *Genetics Selection Evolution* **48**: 86.

Bolormaa, S., Ruvinsky, A., Walkden-Brown, S.W. and van der Werf, J.H.J. 2008. DNA-based parentage verification in two Australian goat herds. *Small Ruminant Research* **80**: 95–100.

Clarke, S.M., Henry, H.M., Dodds, K.G., Jowett, T.W.D., Manley, T.R., Anderson, R.M. and McEwan, J.C. 2014. A high throughput single nucleotide polymorphism multiplex assay for parentage assignment in New Zealand sheep. *PLOS ONE* **9**: e93392.

Dodds, K.G., Tate, M.L. and Sise, J.A. 2005. Genetic evaluation using parentage information from genetic markers. *Journal of Animal Science* **83**: 2271–2279.

Double, M.C., Cockburn, A., Barry, S.C. and Smouse, P.E. 1997. Exclusion probabilities for single-locus paternity analysis when related males compete for matings. *Molecular Ecology* **6**: 1155–1166.

Galloway, S.M., McNatty, K.P., Cambridge, L.M., Laitinen, M.P.E., Juengel, J.L., Jokiranta, T.S., McLaren, R.J., Luiro, K., Dodds, K.G., Montgomery, G.W., Beattie, A.E., Davis, G.H. and Ritvos O. 2000. Mutations in an oocyte-derived growth factor gene (BMP15) cause increased ovulation rate and infertility in a dosage-sensitive manner. *Nature Genetics* **25**: 279–283.

Hayes, B.J., Bowman, P.J., Chamberlain, A.J. and Goddard, M.E. 2009. Invited review: genomic selection in dairy cattle: progress and challenges. *Journal of Dairy Science* **92**: 433–443.

Heaton, M.P., Leymaster, K.A., Kalbfleisch, T.S., Kijas, J.W., Clarke, S.M., McEwan, J.C., Maddox, J.F., Basnayake, V., Petrik, D.T., Simpson, B., Smith, T.P., Chitko-McKown, C.G. and International Sheep Genomics Consortium. 2014. SNPs for parentage testing and traceability in globally diverse breeds of sheep. *PLOS ONE* **9**: e94851.

Hill, W.G., Salisbury, B.A. and Webb, A.J. 2008. Parentage identification using single nucleotide polymorphism genotypes: application to product tracing. *Journal of Animal Science* **86**: 2508–2517.

Israel, C. and Weller, J.I. 2000. Effect of misidentification on genetic gain and estimation of breeding values in dairy cattle populations. *Journal of Dairy Science* **83**: 181–187.

Judge, M.M., Kearney, J.F., McClure, M.C., Sleator, R.D. and Berry, D.P. 2016. Evaluation of developed low-density genotype panels for imputation to higher density in independent dairy and beef cattle populations. *Journal of Animal Science* **94**: 949–962.

McHugh, N., Pabiou, T., Wall, E., McDermott, K. and Berry, D.P. 2017. Impact of alternative definitions of contemporary groups on genetic evaluations of traits recorded at lambing. *Journal of Animal Science* **95**: 1926–1938.

Nicol, L., Bishop, S.C., Pong-Wong, R., Bendixen, C., Holm, L.E., Rhind, S.M. and McNeilly, A.S. 2009. Homozygosity for a single base-pair mutation in the oocyte-specific GDF9 gene results in sterility in Thoka sheep. *Reproduction* **138**: 921–933.

Nielsen, R., Paul, J.S., Albrechtsen, A. and Song, Y.S. 2011. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics* **12**: 443–451.

O'Brien, A.C., McHugh, N., Wall, E., Pabiou, T., McDermott, K., Randles, S., Fair, S. and Berry, D.P. 2017. Genetic parameters for lameness, mastitis and dagginess in a multi-breed sheep population. *Animal: An International Journal of Animal Bioscience* **11**: 911–919.

Purfield, D., McClure, M. and Berry, D.P. 2016. Justification for setting the individual animal genotype call rate threshold at eighty-five percent. *Journal of Animal Science* **94**: 4558–4569.

Richardson, I.W., Berry, D.P., Wiencko, H.L., Higgins, I.M., More, S.J., McClure, J., Lynn, D. and Bradley, D.G. 2016. A genome-wide association study for genetic susceptibility to *Mycobacterium bovis* infection in dairy cattle identifies a susceptibility QTL on chromosome 23. *Genetics Selection Evolution* **48**: 19.

Rupp, R., Mucha, S., Larroque, H., McEwan, J. and Conington, J. 2016. Genomic application in sheep and goat breeding. *Animal Frontiers* **6**: 39–44.

Sanders, K., Bennewitz, J. and Kalm, E. 2006. Wrong and missing sire information affects genetic gain in the Angeln dairy cattle population. *Journal of Dairy Science* **89**: 315–321.

Santos, B.F.S., McHugh, N., Byrne, T.J., Berry, D.P. and Amer, P.R. 2015. Comparison of breeding objectives across countries with application to sheep indexes in New Zealand and Ireland. *Journal of Animal Breeding and Genetics* **132**: 144–154.

Sargolzaei, M., Chesnais, J.P. and Schenkel, F.S. 2014. A new approach for efficient genotype imputation using information from relatives. *BMC Genomics* **15**: 478.

Silva, B.D., Castro, E.A., Souza, C.J., Paiva, S.R., Sartori, R., Franco, M.M., Azevedo, H.C., Silva, T.A., Vieira, A.M., Neves, J.P. and Melo, E.O. 2011. A new polymorphism in the growth and differentiation factor 9 (GDF9) gene is associated with increased ovulation rate and prolificacy in homozygous sheep. *Animal Genetics* **42**: 89–92.

Spelman, R.J., Hayes, B.J. and Berry, D.P. 2013. Use of molecular technologies for the advancement of animal breeding: genomic selection in dairy cattle populations in Australia, Ireland and New Zealand. *Animal Production Science* **53**: 869–875.

Van Vleck, L.D. 1970. Misidentification in estimating the paternal sib correlation. *Journal of Dairy Science* **53**: 1469–1474.

Visscher, P., Woolliams, J., Smith, D. and 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.

Wellmann, R., Preuß, S., Tholen, E., Heinkel, J., Wimmers, K. and Bennewitz, J. 2013. Genomic selection using low density marker panels with application to a sire line in pigs. *Genetics Selection Evolution* **45**: 28.

Zhao, F., McParland, S., Kearney, F., Du, L. and Berry, D.P. 2015. Detection of selection signatures in dairy and beef cattle using high-density genomic information. *Genetics Selection Evolution* **47**: 49.

Appendix 1. Number of single nucleotide polymorphism (SNPs) selected per chromosome for each panel density

Chromosome	Panel density									
	87	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
1	5	7	16	22	28	34	39	45	51	56
2	11	11	15	20	25	30	36	41	46	51
3	7	9	14	18	23	27	32	37	41	46
4	4	4	7	10	12	15	17	19	22	24
5	3	4	7	9	11	13	15	18	20	22
6	4	4	7	10	12	14	17	19	21	24
7	4	4	6	8	10	12	14	16	18	20
8	2	3	6	7	9	11	13	15	17	18
9	4	4	6	8	10	12	14	15	17	19
10	2	3	5	7	9	11	12	14	16	18
11	4	4	4	5	7	8	10	11	12	14
12	2	3	5	6	8	10	11	13	15	16
13	3	3	5	7	8	10	12	14	15	17
14	3	3	4	5	7	8	9	10	12	13
15	4	4	5	7	8	10	12	13	15	16
16	0	2	4	6	7	9	10	12	13	15
17	6	6	4	6	7	9	10	12	13	15
18	5	5	4	6	7	8	10	11	13	14
19	3	3	4	5	6	7	9	10	11	12
20	2	2	3	4	5	6	7	8	9	10
21	3	3	3	4	5	6	7	8	9	10
22	1	2	3	4	5	6	7	8	9	10
23	2	2	4	5	6	8	9	10	11	12
24	1	2	3	3	5	5	6	7	8	10
25	2	2	3	4	5	6	6	7	8	9
26	0	1	3	4	5	5	6	7	8	9