


Population structure and breed composition prediction in a multi-breed sheep population using genome-wide single nucleotide polymorphism genotypes

A. C. O'Brien^{1,2} , D. C. Purfield¹, M. M. Judge¹, C. Long³, S. Fair² and D. P. Berry^{1†}

¹Animal and Grassland Research and Innovation Centre, Teagasc, Moorepark, Fermoy P61 P302, Co. Cork, Ireland; ²Laboratory of Animal Reproduction, Department of Biological Sciences, Faculty of Science and Engineering, University of Limerick, Limerick V94 T9PX, Ireland; ³Sheep Ireland, Highfield House, Shinagh, Bandon P72 X050, Co. Cork, Ireland

(Received 25 February 2019; Accepted 21 August 2019; First published online 15 October 2019)

Knowledge of population structure and breed composition of a population can be advantageous for a number of reasons; these include designing optimal (cross)breeding strategies in order to maximise non-additive genetic effects, maintaining flockbook integrity by authenticating animals being registered and as a quality control measure in the genotyping process. The objectives of the present study were to 1) describe the population structure of 24 sheep breeds, 2) quantify the breed composition of both flockbook-recorded and crossbred animals using single nucleotide polymorphism BLUP (SNP-BLUP), and 3) quantify the accuracy of breed composition prediction from low-density genotype panels containing between 2000 and 6000 SNPs. In total, 9334 autosomal SNPs on 11 144 flockbook-recorded animals and 1172 crossbred animals were used. The population structure of all breeds was characterised by principal component analysis (PCA) as well as the pairwise breed fixation index (F_{st}). The total number of animals, all of which were purebred, included in the calibration population for SNP-BLUP was 2579 with the number of animals per breed ranging from 9 to 500. The remaining 9559 flockbook-recorded animals, composite breeds and crossbred animals represented the test population; three breeds were excluded from breed composition prediction. The breed composition predicted using SNP-BLUP with 9334 SNPs was considered the gold standard prediction. The pairwise breed F_{st} ranged from 0.040 (between the Irish Blackface and Scottish Blackface) to 0.282 (between the Border Leicester and Suffolk). Principal component analysis revealed that the Suffolk from Ireland and the Suffolk from New Zealand formed distinct, non-overlapping clusters. In contrast, the Texel from Ireland and that from New Zealand formed integrated, overlapping clusters. Composite animals such as the Belclare clustered close to its founder breeds (i.e., Finn, Galway, Lley and Texel). When all 9334 SNPs were used to predict breed composition, an animal that had a majority breed proportion predicted to be ≥ 0.90 was defined as purebred for the present study. As the panel density decreased, the predicted breed proportion threshold, used to identify animals as purebred, also decreased (≥ 0.85 with 6000 SNPs to ≥ 0.60 with 2000 SNPs). In all, results from the study suggest that breed composition for purebred and crossbred animals can be determined with SNP-BLUP using ≥ 5000 SNPs.

Keywords: principal component analysis, best linear unbiased prediction, breed diversity, fixation index, low-density panels

Implications

Knowledge of breed composition can be advantageous for a number of reasons. The deployment of approaches used in the present study could be readily utilised to accurately describe both population structure and breed composition. Subsequently, both population structure and breed composition can be used as a quality control measure in the genotyping process, to ensure the integrity of breed society flockbooks, to aid in the design of optimal (cross)breeding strategies,

and to group animals by breed to achieve accurate genotype imputation.

Introduction

Knowledge of population structure and breed composition of a population has many potential uses. Firstly, in populations of crossbred or composite animals which predominate in some countries (Dodds *et al.*, 2014; McHugh *et al.*, 2017), understanding the true breed composition of these animals could be useful to aid in designing optimal (cross)breeding

† E-mail: Donagh.berry@teagasc.ie

strategies to maximise the exploitation of inter-breed non-additive genetic effects. In the case of seedstock (i.e., purebred animals registered in a flockbook) animals, certainty of the breed composition of each animal could be helpful to breed societies to verify or authenticate the animal being registered, thereby maintaining the integrity of the flockbook (Dodds *et al.*, 2014). Furthermore, products (e.g., milk, meat, wool) from some breeds may command a higher market price (Judge *et al.*, 2017), while financial incentives exist for the preservation of some rare breeds. Therefore, knowledge of the breed composition of such animals is important to deliver consumer confidence in the authenticity of the product as well as ensuring the purity of the rare breed. In addition, accurate breed composition of individual animals is important for including as a correction factor in multi-breed genetic evaluations (McHugh *et al.*, 2017). Finally, being able to predict the breed composition of a batch of samples genotyped together can be valuable in detecting errors in the sample procurement or genotyping process, which can then be rectified prior to including in any downstream analysis.

In the absence of genomic information, the breed proportion of an animal is calculated as simply the average breed composition of both parents (Sölkner *et al.*, 2010). While this method is accurate where both parents are purebred (and parentage is verified), issues can arise where at least one parent is crossbred. The breed composition of the resulting progeny from the mating of a crossbred parent may deviate from the average contribution of both parents due to Mendelian sampling during gametogenesis (Sölkner *et al.*, 2010). While software packages exist to estimate breed composition from genomic data (e.g., ADMIXTURE; Alexander *et al.*, 2009), alternative approaches which are more amenable for seamless integration into pipelines used in genomic evaluations warrant consideration. Moreover, the repeatability of some of these software suites has recently been questioned (Crum *et al.*, 2018).

The objective of the present study was to accurately describe the population structure of 24 sheep breeds farmed in Ireland and quantify the breed composition of both purebred and crossbred animals using a single nucleotide polymorphism BLUP (SNP-BLUP) approach. Commonly used in genomic evaluations, the pipelines for SNP-BLUP generally already exist for deploying such an approach. The utilisation and accuracy of low-density panels to predict breed composition was also quantified.

Material and methods

Pedigree and genotype data

Pedigree-based breed composition information and genotype data were available on 12 875 adult and young sheep from 24 breeds as well as crossbreds. Of these sheep, 11 372 were recorded to be purebred, the vast majority of which were registered in the respective breed society flockbook. The remaining 1503 animals were crossbred animals which originated from a research flock. While the exact breed composition

of individual animals in the crossbred research flock was unknown, the breeds included in the breeding programme were known. For example, only Charollais rams were bred to crossbred ewes, while the breed composition of the crossbred ewes may have included Cheviot, Suffolk, and Texel.

Single nucleotide polymorphism data of the 12 875 animals were available from one of three genotype panels: the Illumina OvineSNP50 Beadchip (51 135 SNPs; 3371 animals), a custom Illumina Infinium panel (14 918 SNPs; 9245 animals), and a custom AgResearch Ovine HD Beadchip (606 006 SNPs; 259 animals). Only the 9769 SNPs that were common to all three panels were retained. Animals were only retained if they had a call rate ≥ 0.90 ($n = 12\,316$). Only autosomal SNPs with a known genomic position, a call rate ≥ 0.95 , as well as an Illumina GenCall (GC) score ≥ 0.55 were retained. Inconsistency in the Mendelian inheritance pattern of each autosomal SNP was subsequently determined based on the proportion of genotypes per SNP that were opposing homozygotes in a validated parent–offspring pair; a total of 1492 sire–dam–offspring trios and 3418 parent–offspring pairs existed within the 12 316 genotyped animals. Single nucleotide polymorphisms were discarded where $>2\%$ of the parent–offspring autosomal genotypes did not conform to normal Mendelian inheritance. Finally, the extent to which each SNP genotype deviated from Hardy–Weinberg equilibrium was calculated within the five breeds with the largest genotyped populations (i.e., Belclare ($n = 1255$), Charollais ($n = 3004$), Suffolk ($n = 1997$), Texel ($n = 3365$), and Vendeen ($n = 758$)) separately. Single nucleotide polymorphisms that deviated from Hardy–Weinberg equilibrium ($P < 0.01 \times 10^{-7}$) in any one of the five main breeds were not considered further. Following edits, 9334 autosomal SNPs on 11 144 flockbook-recorded animals from 24 breeds as well as 1172 crossbred animals remained (Table 1). Information on all 9334 SNPs used in the present study is available in Supplementary Material Table S3. In all, 95.55% of SNPs were retained, while 98.00% and 77.98% of flockbook-recorded and crossbred animals passed quality control checks, respectively. To ensure that all sporadically missing genotypes were imputed, genotype imputation was undertaken using Flmpute V2.2 (Sargolzaei *et al.*, 2014), thus facilitating the requirements of software programmes used in the downstream analyses.

Population structure

Analysis of population structure was undertaken using principal component analysis (PCA) in the EIGENSOFT package (Price *et al.*, 2006). The population structure of all 11 144 flockbook-recorded animals was first determined. This was to ensure that the breed recorded in the pedigree was the true breed of the animal and that the animals of a recorded breed clustered together. Only animals that resided within the breed cluster agreeing with that recorded in the national database were used in the present study.

Of the animals validated from genotypes to be the breed recorded on the database, a maximum of 50 animals per breed were included in the PCA to describe breed structure

Table 1 The number of animals by pedigree recorded breed composition, included in the principal component analysis (PCA) for population structure and those included in the calibration or test population for single nucleotide polymorphism BLUP (SNP-BLUP).

Breed name	Flockbook-recorded	PCA	SNP-BLUP	
			Calibration	Test
Belclare	1 255	50	0	1 255
Beltex	181	50	181	0
Blackface*	65	65*	65	0
Bluefaced Leicester	30	30	30	0
Border Leicester	9	9	9	0
Charollais	3 003	50	500	2 503
Cheviot	38	38	38	0
Crossbred	1 172	0	0	1 172
EasyCare	14	14	0	0
Finn	28	28	28	0
Galway	93	50	93	0
Hampshire Down	14	14	14	0
Highlander	6	6	0	6
Lleyn	40	40	40	0
New Zealand Suffolk	78	50	0	0
New Zealand Texel	86	50	0	0
Primera	9	9	0	9
Rouge de l'Ouest	25	25	25	0
Shropshire	10	10	10	0
Suffolk	1 997	50	500	1 497
Swaledale	16	16	16	0
Texel	3 359	50	500	2 859
Vendeen	758	50	500	258
Zwartble	30	30	30	0

*For principle component analysis (PCA), the Blackface breed was divided into the Irish Blackface ($n = 35$) and the Scottish Blackface ($n = 30$).

in the present study. Where there were more than 50 animals per breed, the most marginally influential animals based on pedigree were identified and used for the analysis of population structure. For each breed separately, the marginal contribution of each animal to the current population of live animals of that breed was calculated. Animals were then ranked on the basis of their overall contribution and the 50 animals within each breed with the greatest marginal contribution were selected for inclusion in PCA.

Fixation index (F_{st}) values between each pair of breeds were calculated in ADMIXTURE version 1.3 (Alexander *et al.*, 2009). The number of animals per breed included in the PCA and used to estimate F_{st} values between breeds is summarised in Table 1.

Identifying calibration animals for breed composition analysis

As a result of the initial analysis of population structure, the Irish Blackface ($n = 35$) and the Scottish Blackface ($n = 30$) breeds clustered together and were subsequently combined as a single breed; the combined breed group will thus be referred to as the Blackface breed. For breeds with <200 animals genotyped (i.e., Beltex, Blackface, Bluefaced Leicester,

Border Leicester, Cheviot, Finn, Galway, Hampshire Down, Lleyn, Rouge de l'Ouest, Shropshire, Swaledale, and Zwartble), all genotyped animals were automatically included in the calibration population for subsequent breed prediction (Table 1). For the remaining breeds (i.e., Charollais, Suffolk, Texel, and Vendeen), 500 breed-verified animals per breed were randomly selected to represent a calibration population. A total of 2579 animals representing 17 breeds were included in the calibration population. The test population consisted of the remaining Charollais, Suffolk, Texel, and Vendeen animals as well as three composite breeds (i.e., Belclare, Highlander, and Primera) and the 1172 crossbred animals. The New Zealand Suffolk and the New Zealand Texel were excluded from breed composition prediction, while the EasyCare was not included in the test population as the breeds used to form the breed were not included in the present study.

Breed composition estimated using single nucleotide polymorphism best linear unbiased prediction

Variance components are a requirement of SNP-BLUP. The phenotypic variance of breed composition was estimated for the calibration population of each breed individually and was taken as the product of the proportion of the entire calibration population that were breed verified for the specific breed under analysis, multiplied by the proportion of the calibration dataset that were not breed verified for the specific breed under analysis, but were breed verified for another breed (i.e., the variance of a binary trait). Within the calibration population for a specific breed, the number of animals coded as purebred for that breed was approximately equal to the number of animals coded as non-purebred for that breed (but purebred for another breed). For example, 500 animals were classified as purebred Texel, while 507 other animals were classified as non-purebred Texel; each of the 16 remaining calibration breeds were represented in the non-purebred Texel element. The additive genetic effect was estimated from the phenotypic variance assuming a heritability of 0.999. The genetic variance per SNP was subsequently calculated for the calibration population of each breed individually as:

$$\sigma_{\text{SNP}_i}^2 = \frac{\sigma_a^2}{\sum_{i=1}^n 2p_i(1-p_i)}$$

where $\sigma_{\text{SNP}_i}^2$ is the estimated genetic variance common to all SNPs within that breed, σ_a^2 is the additive genetic variance for that breed, estimated previously, and p_i is the frequency of a given allele at position i which was summed across all n SNPs. Single nucleotide polymorphism effects were estimated using SNP-BLUP within the MiX99 Software suite (MiX99 Development Team, 2015) where all SNPs were simultaneously considered as random effects; an intercept term was included in all models. The linear mixed model fitted to each breed separately was:

$$Y_i = \mu + \sum_{j=1}^n X_j g_{ij} + e_i$$

where Y_i is the dependent variable which was either one (assumed purebred) or zero (purebred but not of the breed under investigation) to imitate the proportion of the breed under investigation in animal i ; μ is an intercept term; X_j is the allele substitution effect of locus j ; g_{ij} is the genotype of animal i at locus j and e_i is the residual term for animal i .

The resulting allele substitution effects were then applied to the genotypes of all animals in the test population. Where a breed proportion was predicted to be less than zero, it was set to zero. Where the sum of all estimated breed proportions (across all 17 breeds) of an animal was greater than or less than one, the individual estimated breed composition was rescaled to be one as:

$$\text{Breed}_i^* = \frac{\text{Breed}_i}{\sum_i^n \text{Breed}_i}$$

where Breed_i^* was the rescaled proportion of Breed_i which was the sum of the breed prediction in the individual animal under investigation summed across all n breeds.

Development of the low-density genotype panels

The impact of the number of SNPs included in the breed composition prediction process was quantified by generating low-density panels with 2000, 3000, 4000, 5000, and 6000 SNPs. Each chromosome was first divided into segments of equal length; chromosome length was defined as the distance from the genomic position of the first SNP to that of the last SNP. The number of segments per chromosome was equal to the predefined number of SNPs for that chromosome based on the panel density under investigation (Supplementary Material Table S1). The SNP effects of the calibration population of each breed individually were obtained from SNP-BLUP undertaken using all 9334 SNPs. From this, the SD of the SNP effects across the calibration population of all 17 breeds was calculated. Single nucleotide polymorphisms were then ranked on an index based on the SD of the SNP effect; SNPs with a larger SD ranked higher on the index. A larger SNP effect SD across all 17 breed indicated the marker could better discriminate breeds. Subsequently, SNPs were chosen within each chromosomal segment by selecting the highest rank SNP in the segment. The numbers of SNPs that are common across each of the lower-density panels are in Supplementary Material Table S2. Information on the SNPs included in each of the lower-density panels is in Supplementary Material Table S4. For SNP-BLUP undertaken using the low-density panels, the SNP variances of the calibration population of each individual breed were re-estimated for each low-density panel individually, while the phenotypic, genetic and residual variance calculated for the calibration population of each breed individually remained the same.

Accuracy of breed composition prediction

Measures of breed composition prediction accuracy were only undertaken for animals where the unadjusted sum of the breed prediction per animal was $\geq 75\%$. Breed composition predicted using the full SNP dataset (i.e., 9334 SNPs) was considered the gold standard prediction and all low-density panels were compared to that. An animal was considered purebred when its breed percentage of a specific breed was predicted to be $\geq 90\%$ using 9334 SNPs. The accuracy of breed composition of crossbred animals predicted with the low-density panels was measured as the percentage of animals that were predicted to be within a $\pm 2.5\%$, $\pm 5.0\%$, and $\pm 7.5\%$ range of the breed proportion predicted using all 9334 SNPs.

Results

Population structure

The PCA was successful in separating out breed clusters based on the genomic data (Figures 1 and 2). The first, second, third and fourth principal components accounted for 35.66%, 21.93%, 18.75%, and 14.05% of the variability among breeds, respectively, totalling 90.39% of the variability. Both the Border Leicester and the Bluefaced Leicester formed distinct clusters in close proximity to each other, while both breeds clustered away from all other breeds. All breeds considered as mountain breeds (i.e., Irish and Scottish Blackface, Swaledale, and Cheviot) clustered in very close proximity (Figures 1 and 2). The close genetic relationship between the Irish Blackface and the Scottish Blackface was evident as the breeds produced overlapping and integrated clusters with the minimum F_{st} value (0.040) among all pairwise breed comparisons existing between these breeds; this indicates little to no genetic differentiation between the breed groups. The Shropshire and Hampshire Down breed clusters were also in close proximity (pairwise F_{st} between breeds of 0.151). The maximum breed pairwise F_{st} (0.282) was between the Border Leicester and Suffolk breeds (Table 2; Figure 3).

When the Bluefaced Leicester and Border Leicester were excluded from the analyses, the first, second, third and fourth principal components accounted for 35.37%, 18.95%, 14.64%, and 12.91% of the variation, respectively, totalling 81.87% of the variability. The Galway breed formed a more distinct cluster proximate to the remainder of the breeds when both the Bluefaced Leicester and Border Leicester were excluded from the PCA. The New Zealand Suffolk and Irish Suffolk clustered close together, although they did form distinct clusters (pairwise F_{st} of 0.106). In contrast, the New Zealand Texel and the Irish Texel which had a breed pairwise F_{st} value of 0.071 formed integrated, overlapping clusters. When the New Zealand Suffolk and Texel as well as the Irish Suffolk and Texel were solely included in the PCA (Supplementary Material Figure S1), the first, second, third and fourth principal components accounted for 25.07%, 10.29%, 7.00%, and 3.56% of the variation, respectively, totalling 45.92% of the variability.

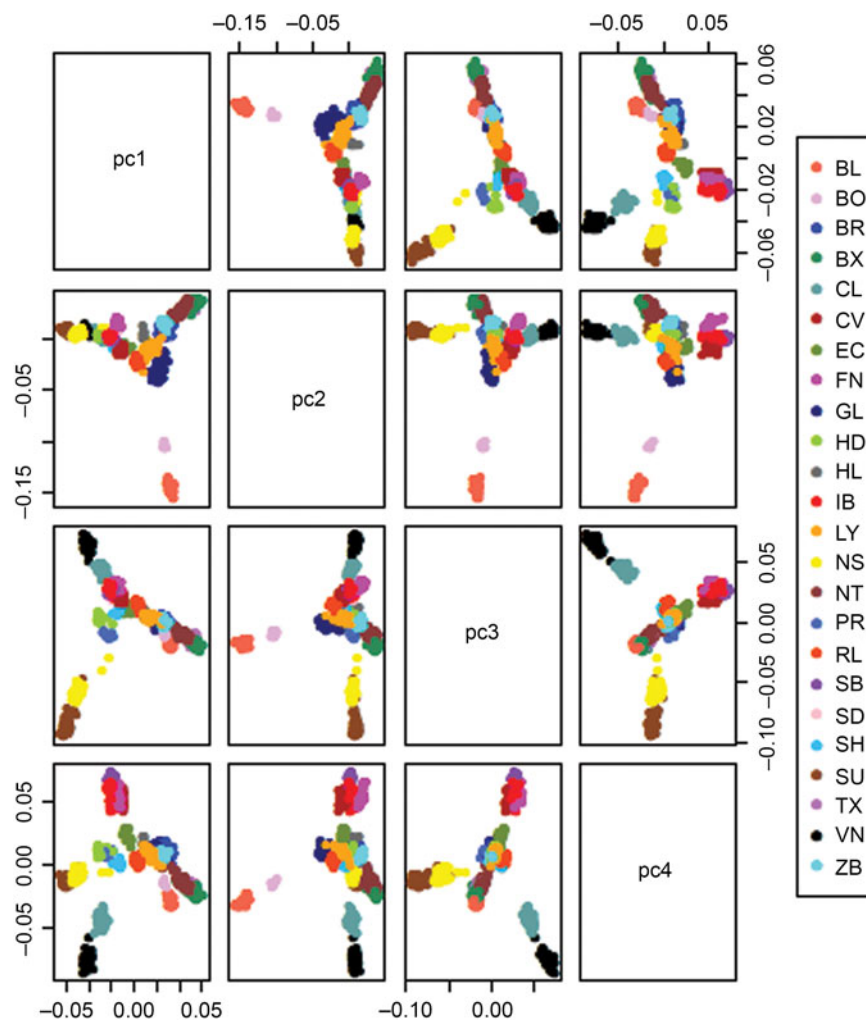


Figure 1 Principal component analysis of purebred animals distributed across the first four principal components. Breeds included in the analysis were Bluefaced Leicester (BL), Border Leicester (BO), Belclare (BR), Beltex (BX), Charollais (CL), Cheviot (CV), EasyCare (EC), Finn (FN), Galway (GL), Hampshire Down (HD), Highlander (HL), Irish Blackface (IB), Lleyn (LY), New Zealand Suffolk (NS), New Zealand Texel (NT), Primera (PR), Rouge de l'Ouest (RL), Scottish Blackface (SB), Swaledale (SD), Shropshire (SH), Suffolk (SU), Texel (TX), Vendeen (VN), and Zwartble (ZB).

The Belclare, a composite breed, clustered close to the breeds used in its development, namely the Finn, Galway, Lleyn, and Texel. Furthermore, low F_{st} values existed between the Belclare and Finn (0.119), between the Belclare and Galway (0.092), between the Belclare and Lleyn (0.066), and between the Belclare and Texel (0.076; Table 2; Figure 3). While not all founder breeds used to develop the Primera (i.e., Suffolk, Hampshire Down), Highlander (i.e., Texel, Finn) and EasyCare were included in the present study, the composite breeds did cluster in close proximity to the founder breeds (or breeds closely related to the founder breeds) that were included.

Breed composition

The percentage of Charollais, Suffolk, Texel, and Vendeen animals that were verified purebred based on the PCA but also deemed purebred by SNP-BLUP as ≥ 0.98 , ≥ 0.95 , ≥ 0.90 , ≥ 0.85 , and ≥ 0.80 of the respective breed is in Table 3. All Vendeen animals verified by the original PCA were predicted by SNP-BLUP to have a Vendeen proportion

≥ 0.85 , while 96.90% of Vendeens verified by the PCA had a predicted Vendeen proportion ≥ 0.90 . Of the Suffolks verified by the PCA, 98.33% were predicted by SNP-BLUP to have a Suffolk proportion ≥ 0.85 , the lowest percentage across all breeds. Only 90.45% of the Charollais animals verified by the PCA were predicted by SNP-BLUP to have a Charollais proportion ≥ 0.90 (Table 3). The percentage of Suffolk and Texel animals verified by the PCA to be purebred and predicted by SNP-BLUP to have a respective breed proportion ≥ 0.90 were 91.78% and 90.4%, respectively.

The breed proportion of the known founder breeds included in the three composite breeds was also quantified. Four breeds included in the present study were included in the formation of the Belclare. The estimated Finn proportion within the Belclare animals ranged from 0.018 to 0.196, while the Galway proportion ranged from 0.049 to 0.191; the Lleyn breed proportion in the Belclare animals ranged from 0.053 to 0.271, while the Texel proportion ranged from 0.167 to 0.489. A small proportion of Blackface was also predicted in the Belclare animals; the maximum proportion

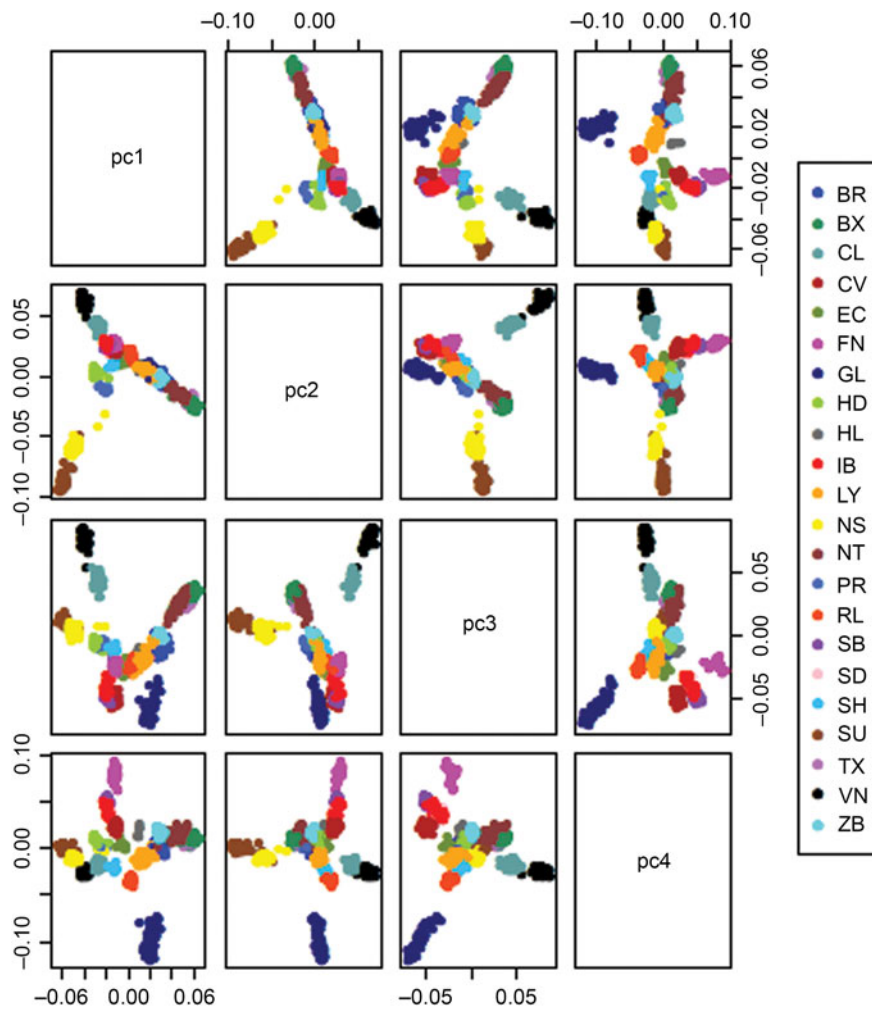


Figure 2 Principal component analysis of purebred animals distributed across the first four principal components. Breeds included in the analysis were Belclare (BR), Beltex (BX), Charollais (CL), Cheviot (CV), EasyCare (EC), Finn (FN), Galway (GL), Hampshire Down (HD), Highlander (HL), Irish Blackface (IB), Lleyl (LY), New Zealand Suffolk (NS), New Zealand Texel (NT), Primera (PR), Rouge de l'Ouest (RL), Scottish Blackface (SB), Swaledale (SD), Shropshire (SH), Suffolk (SU), Texel (TX), Vendeen (VN), and Zwartble (ZB).

predicted was 0.122. Within the six Highlander animals used in the present study, the predicted Finn proportion ranged from 0.116 to 0.155, while the predicted Texel proportion ranged from 0.095 to 0.158. Within the Primera animals, the predicted Hampshire Down proportion per animal ranged from 0.055 to 0.105, while the predicted Suffolk breed proportion ranged from 0.149 to 0.196.

Breed composition predicted using low-density genotype panels

The unadjusted sum of the predicted breed composition was <0.75 for 2 (Texel), 5 (4 Belclare, 1 Texel), and 20 (18 Texel, 1 Belclare, 1 Charollais) animals when SNP panels contained 4000, 3000, and 2000 SNPs, respectively; these animals were not subsequently included in the further analysis. As the panel density decreased, the predicted breed proportion threshold, used to identify animals as purebred, also decreased. When ≥ 5000 SNPs were included on the panel, $\geq 90\%$ of all animals deemed purebred using the 9334 SNPs, had a respective breed proportion ≥ 0.85 . At 4000

SNPs, $\geq 90\%$ of Charollais and Suffolk animals had a respective predicted breed proportion ≥ 0.80 , while a breed proportion threshold of 0.85 was sufficient to recognise $\geq 90\%$ of purebred Texel and Vendeen animals. When the panel density was reduced to 3000 SNPs, $\geq 90\%$ of Charollais, Suffolk, and Texel animals deemed purebred using 9334 SNPs had a respective breed proportion ≥ 0.75 , while over 90% of purebred Vendeen animals had a breed proportion ≥ 0.80 . The threshold required to recognise $\geq 90\%$ of animals, previously deemed purebred, using 2000 SNPs, was 0.60 for Charollais and Suffolk animals and 0.65 for Texel and Vendeen.

Using 9334 SNPs, the mean predicted breed proportion (SD in parenthesis) of Charollais, Cheviot, Suffolk, and Texel in the 1172 crossbred animals was 0.375 (0.075), 0.081 (0.064), 0.118 (0.058) and 0.109 (0.089), respectively. The percentage of crossbred animals, whose breed composition was within a $\pm 2.5\%$, $\pm 5.0\%$, and $\pm 7.5\%$ range of the breed composition, predicted using the 9334 SNPs is in Table 4. As the density of the panel reduced, the number of animals

Table 2 Mean fixation index (F_{st}) values among breeds. Breeds include Bluefaced Leicester (BL), Border Leicester (BO), Belclare (BR), Beltex (BX), Charollais (CL), Cheviot (CV), EasyCare (EC), Finn (FN), Galway (GL), Hampshire Down (HD), Highlander (HL), Irish Blackface (BF), Lleyn (LY), New Zealand Suffolk (NS), New Zealand Texel (NT), Primera (PR), Rouge de l'Ouest (RL), Scottish Blackface (SB), Swaledale (SD), Shropshire (SH), Suffolk (SU), Texel (TX), Vendeen (VN), and Zwartble (ZB).

	BL	BO	BR	BX	CL	CV	EC	FN	GL	HD	HL	IB	LY	NS	NT	PR	RL	SB	SD	SH	SU	TX	VN	
BO	0.213																							
BR	0.185	0.219																						
BX	0.211	0.248	0.080																					
CL	0.206	0.242	0.102	0.131																				
CV	0.198	0.234	0.099	0.130	0.103																			
EC	0.231	0.266	0.135	0.161	0.139	0.136																		
FN	0.240	0.274	0.119	0.158	0.131	0.132	0.172																	
GL	0.183	0.223	0.092	0.118	0.120	0.115	0.150	0.155																
HD	0.235	0.271	0.133	0.162	0.121	0.130	0.168	0.156	0.150															
HL	0.211	0.247	0.104	0.127	0.122	0.119	0.157	0.131	0.127	0.152														
IB	0.191	0.228	0.084	0.115	0.080	0.071	0.121	0.106	0.103	0.107	0.101													
LY	0.166	0.201	0.066	0.090	0.089	0.084	0.122	0.125	0.084	0.122	0.098	0.069												
NS	0.228	0.264	0.129	0.156	0.115	0.125	0.166	0.157	0.144	0.134	0.148	0.105	0.116											
NT	0.204	0.240	0.082	0.076	0.121	0.119	0.154	0.146	0.113	0.152	0.111	0.105	0.085	0.146										
PR	0.246	0.279	0.146	0.172	0.141	0.145	0.183	0.173	0.160	0.164	0.164	0.125	0.135	0.140	0.160									
RL	0.190	0.230	0.102	0.129	0.108	0.112	0.150	0.148	0.106	0.145	0.130	0.098	0.089	0.139	0.122	0.159								
SB	0.209	0.247	0.106	0.136	0.104	0.094	0.140	0.127	0.125	0.128	0.122	0.040	0.092	0.126	0.126	0.145	0.120							
SD	0.239	0.276	0.136	0.164	0.132	0.127	0.169	0.156	0.153	0.157	0.152	0.096	0.122	0.155	0.153	0.176	0.148	0.119						
SH	0.231	0.265	0.133	0.161	0.128	0.135	0.171	0.169	0.140	0.151	0.153	0.118	0.118	0.151	0.149	0.174	0.137	0.141	0.169					
SU	0.246	0.282	0.146	0.174	0.133	0.144	0.179	0.170	0.162	0.150	0.164	0.120	0.134	0.106	0.165	0.163	0.158	0.142	0.170	0.167				
TX	0.207	0.242	0.076	0.054	0.126	0.124	0.156	0.153	0.113	0.158	0.121	0.111	0.082	0.154	0.071	0.169	0.125	0.131	0.158	0.156	0.170			
VN	0.230	0.265	0.128	0.157	0.076	0.125	0.163	0.148	0.144	0.140	0.145	0.100	0.114	0.135	0.145	0.160	0.133	0.124	0.152	0.146	0.152	0.153		
ZB	0.214	0.252	0.102	0.109	0.129	0.127	0.161	0.153	0.128	0.162	0.133	0.111	0.102	0.154	0.107	0.172	0.131	0.133	0.158	0.159	0.172	0.106	0.154	

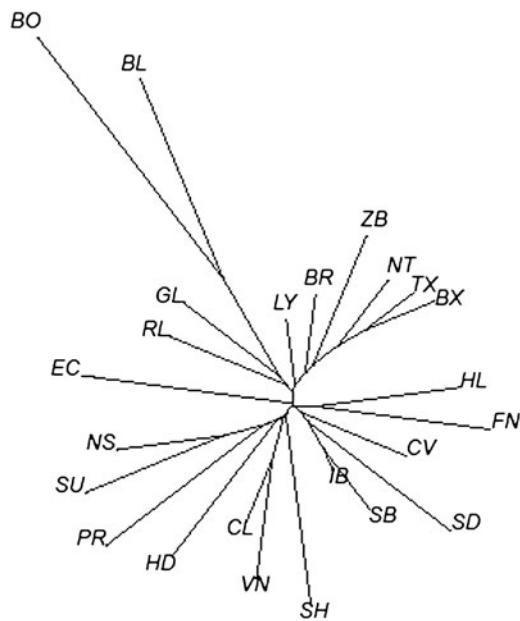


Figure 3 Genetic distance between breeds based on pairwise F_{st} estimates. Breeds included in the analysis were Bluefaced Leicester (BL), Border Leicester (BO), Belclare (BR), Beltex (BX), Charollais (CL), Cheviot (CV), EasyCare (EC), Finn (FN), Galway (GL), Hampshire Down (HD), Highlander (HL), Irish Blackface (IB), Lleyn (LY), New Zealand Suffolk (NS), New Zealand Texel (NT), Primera (PR), Rouge de l'Ouest (RL), Scottish Blackface (SB), Swaledale (SD), Shropshire (SH), Suffolk (SU), Texel (TX), Vendeen (VN), and Zwartble (ZB).

Table 3 The percentage of animals that were verified purebred based on principal component analysis and identified by single nucleotide polymorphism BLUP (SNP-BLUP) as ≥ 0.98 , ≥ 0.95 , ≥ 0.90 , ≥ 0.85 and ≥ 0.80 of the respective breed

Verified breed	Test animals	SNP-BLUP breed composition				
		≥ 0.98	≥ 0.95	≥ 0.90	≥ 0.85	≥ 0.80
Charollais	2503	56.17	61.17	90.45	97.56	99.60
Suffolk	1497	68.14	71.81	91.78	96.53	98.33
Texel	2859	55.09	58.41	90.84	98.92	99.86
Vendeen	258	75.19	80.62	96.90	100.00	100.00

that were within a $\pm 2.5\%$, $\pm 5.0\%$, and $\pm 7.5\%$ range also decreased. For example, when the breed composition of the 1172 crossbred animals was predicted using 6000 SNPs, the Texel breed proportion in 83.45% of animals was within a $\pm 2.5\%$ range of that predicted using 9334 SNPs; this reduced to 50.60% of animals when 3000 SNPs were used to predict breed composition. The percentage of animals whose breed proportion was within $\pm 2.5\%$, $\pm 5.0\%$, or $\pm 7.5\%$ of the predicted breed composition using the 9334 SNPs was consistently poorest in Charollais. When breed proportion was predicted using 6000 SNPs, 93.94% and 92.66% of crossbred animals were within $\pm 2.5\%$ for the Cheviot and Suffolk proportions, respectively. In contrast, only 73.29% of crossbred animals were within $\pm 2.5\%$ of the Charollais proportion predicted using 9334 SNPs. For $\geq 90\%$ of crossbred animals,

Table 4 The percentage of 1172 crossbred animals whose breed composition predicted using the low-density panels were within a $\pm 2.5\%$, $\pm 5.0\%$ and $\pm 7.5\%$ range of the breed composition predicted using the entire single nucleotide polymorphism dataset

	Range (%)	Single nucleotide polymorphism panel density				
		6000	5000	4000	3000	2000
Charollais	± 2.5	73.29	62.03	50.94	37.88	18.52
	± 5.0	95.82	90.27	84.81	67.15	35.24
	± 7.5	99.83	98.55	96.50	85.84	51.28
Cheviot	± 2.5	93.94	90.70	86.95	71.42	57.17
	± 5.0	100.00	100.00	99.91	97.87	88.99
	± 7.5	100.00	100.00	100.00	99.91	98.04
Suffolk	± 2.5	92.66	86.43	71.16	56.23	30.89
	± 5.0	100.00	99.74	97.01	89.08	57.59
	± 7.5	100.00	100.00	99.74	97.95	78.24
Texel	± 2.5	83.45	71.93	60.49	50.60	28.50
	± 5.0	99.15	95.56	89.93	78.07	47.78
	± 7.5	100.00	99.91	98.38	91.55	61.69

the predicted breed proportion of Charollais, Cheviot, Suffolk, and Texel was within a $\pm 5.0\%$ range of the breed composition predicted using 9334 SNPs, when ≥ 5000 SNPs were used. Nonetheless, it was necessary to broaden the range to $\pm 7.5\%$ when 4000 SNPs were included on the panel (Table 4).

Discussion

The objective of the present study was to describe the population structure of 24 sheep breeds commonly farmed in Ireland and develop pipelines to accurately predict the breed composition of both flockbook-recorded and crossbred animals using a SNP-BLUP approach. The utilisation and accuracy of low-density panels to predict breed composition was also quantified. The inclusion of breed composition as a quality control check when importing a new batch of genotypes into central databases can be of considerable benefit to ensure that the genotypes correspond to the correct animals aiding the identification of possible sample mix-ups somewhere in the chain. Another potential benefit of breed prediction is the ability to accurately assign animals to a (majority) breed for imputation purposes. Previous work by O'Brien *et al.* (2019) in sheep revealed that where small reference populations exist, imputation accuracy can be improved when only animals from that breed were used in the reference population. Similarly, when undertaking within-breed genome-based studies, the ability to verify that the recorded purebred is truly purebred is important to avoid undetected population stratification affecting the results (McGovern *et al.*, 2019; Twomey *et al.*, 2019). Other uses of knowing the breed composition of an animal include the monitoring of individual society flockbooks to ensure that animals being registered to the flockbook are breed verified (Dodds *et al.*, 2014), as well as aiding in the development of optimal (cross)breeding strategies. Further,

financial incentives may often exist for the preservation of a select number of native breeds. Similarly, a premium price is often paid for products with perceived superior quality; Angus beef, Merino wool, and Irish Mayo Connemara hill lamb are examples of such products.

Population structure

Population structure has been demonstrated in native (Welsh) sheep (Beynon *et al.*, 2015), in a sample of European sheep (Lawson Handley *et al.*, 2007) and in a diverse selection of sheep (Kijas *et al.*, 2012). Of the 24 breeds included in the present study, 13 have not previously been included in a breed characterisation study. Corroborating results of previous studies by Kijas *et al.* (2012) and Lawson Handley *et al.* (2007), the population structure of breeds in the present study stratified somewhat by geography. This was true for two French breeds (Charollais and Vendéen) and two traditional English breeds (Bluefaced Leicester and Border Leicester). While the Rouge de l'Ouest originates in France, genetic similarity between it and the other French breeds was not evident although this may be due to the relatively small number of genotyped Rouge de l'Ouest sheep compared with the Charollais and Vendéen. Both the Beltex and Texel clustered close to the Zwartble breed despite being phenotypically quite diverse. The Texel is characterised by its white, wool-free head and limbs, black hooves and a short wide face with a black nose. In contrast, the Zwartble has a black/brown fleece, head and limbs, with a distinctive white blaze on the face and white socks on two to four legs. Nonetheless, both breeds originated in close proximity in north of the Netherlands; the Texel originated in the Wadden Islands, while the Zwartble originated in Friesland. Genetic similarity between the Texel and Beltex breeds is evident both from the PCA and from the low F_{st} of 0.054 between the breeds. The Beltex breed was developed in Belgium by selectively breeding Texel sheep for hypertrophy.

The overlapping PCA clusters and low F_{st} value (0.040) between the Irish and Scottish Blackface breeds were expected, as it has become the norm for Blackface breeders in Ireland to import Perth-type Scottish Blackface rams from Scotland thereby reducing the genetic diversity between these two 'breeds'. As both the Shropshire and the Hampshire Down were separately formed in the late and early 1800s, respectively, in the United Kingdom by crossing the Cotswold and Southdown with other traditional English breeds, it was expected that these breeds would cluster in close proximity.

Although the Irish Suffolk and New Zealand Suffolk are technically considered the same breed, results from the present study reveal that they are in fact two distinctive sub-populations, clustering separately albeit in close proximity. Similarly, Kijas *et al.* (2012) reported that the Australian and Irish Suffolk formed non-overlapping clusters. The separate clusters suggests that while both originated from the traditional English Suffolk breed, breeders in the respective countries have been selecting for different attributes since the introduction of the Suffolk in 1891 and 1913 in Ireland and New Zealand, respectively. Ireland and New Zealand operate both terminal and maternal/dual purpose breeding

indices. Moderate to strong correlations exist between the Irish Terminal and New Zealand Terminal Sire Overall Index (0.66) and between the Irish Replacement index and New Zealand Dual Purpose Overall Index (0.86; Santos *et al.*, 2015). A similar trend of distinct non-overlapping clusters was not observed for the Irish Texel and New Zealand Texel; this may be due to the much later introduction of the Texel in New Zealand in the 1990s.

Breed prediction

Prediction of breed composition using genomic information has been performed in sheep (Dodds *et al.*, 2014) and cattle (Kuehn *et al.*, 2011; Frkonja *et al.*, 2012) using SNP data. The accuracy by which breed composition can be predicted is dependent on the number of informative SNPs considered (Judge *et al.*, 2017) as well as the method of prediction. Methods previously used to predict breed composition in cattle include regression (Kuehn *et al.*, 2011), ADMIXTURE (Alexander *et al.*, 2009), and SNP-BLUP (Strucken *et al.*, 2017) while the regression method and a genomic selection method have previously been applied to sheep (Dodds *et al.*, 2014). The accuracy of predicting breed composition (depicted as the regression of predicted breed composition on the recorded breed composition) reported for the regression method varies from 0.737 (Kuehn *et al.*, 2011) in cross-bred cattle to 0.979 in sheep (Dodds *et al.*, 2014). Similarly, the genomic selection method has been reported to vary in prediction accuracy from 0.869 to 0.985 in sheep (Dodds *et al.*, 2014). Strucken *et al.* (2017) reported that SNP-BLUP more accurately predicted the dairy breed proportion in East African cattle than ADMIXTURE where the evaluated SNP panels contained ≤ 1400 SNPs. Alternative methods such as multinomial logistic mixed models could also be used to predict breed composition thus avoiding rescaling the breed composition to sum to one. While a multi-variate method may be computationally more savvy assuming zero covariances amongst breeds; it is mathematically equivalent to running several univariate models as undertaken in the present study.

Many other studies have used the regression of the predicted breed composition on the pedigree recorded breed composition as a metric for prediction accuracy (Kuehn *et al.*, 2011; Dodds *et al.*, 2014). Nonetheless, if clusters exist in the data, then the relationship between predicted and pedigree recorded breed proportion may be exaggerated (i.e., Simpson's paradox). Secondly, errors may exist in the actual recorded breed composition which may bias the estimates of accuracy.

In the present study, SNP-BLUP using 9334 SNPs was more successful in identifying purebred Vendéen animals (that had been breed verified) compared to Charollais, Suffolk and Texel breed-verified animals. Previous estimates of the effective population size (N_e) of Charollais and Texel were 357 and 316, respectively, while the N_e of Suffolk animals ranged from 185 to 300 (Kijas *et al.*, 2012; Purfield *et al.*, 2017). A previous N_e estimates of Vendéen was considerably lower at 195 (Purfield *et al.*, 2017) which may contribute to the greater performance of SNP-BLUP in identifying purebred

Vendean animals. Furthermore, 48.44% of Vendean animals in the test population had at least one parent in the calibration population, while only 13.26%, 13.62%, and 6.43% of Charollais, Suffolk, and Texel animals in the test population, respectively, had at least one parent in the calibration population. The stronger relationship between the Vendean test and calibration populations could also contribute to the greater performance of SNP-BLUP within the purebred verified Vendean animals.

Lower-density panels

The utility of low-density panels for the breed composition prediction has been reported in cattle (Kuehn *et al.*, 2011; Frkonja *et al.*, 2012), although no such study exists in sheep. Kuehn *et al.* (2011) and Frkonja *et al.* (2012) concluded that a minimum of 3000 and 4000 SNPs, respectively, were required for accurate breed composition prediction in cattle. Only 300 SNPs were required by Judge *et al.* (2017) to accurately predict the breed proportion of Angus or Hereford cattle, although the objective of Judge *et al.* (2017) was to predict the breed proportion of a single breed rather than the breed composition of several breeds. The number of SNPs required to predict the breed composition of (purebred and crossbred) cattle is fewer than the 5000 SNPs required in the present study. Greater diversity exists within sheep breeds which could contribute to the greater number of SNPs required for prediction of breed composition. Previous estimates of N_e for Northern European sheep breeds ranged from 100 (Wiltshire) to 795 (Finn; Kijas *et al.*, 2012; Purfield *et al.*, 2017) while estimates of N_e in Hereford, Simmental and Holstein-Friesian cattle ranged from 64 to 127 (McParland *et al.*, 2007). Estimates of genetic diversity (i.e., pairwise F_{st} estimates) between Irish cattle breeds has previously been reported to range between 0.049 (Limousin and Charollais) and 0.165 (Jersey and Hereford; Kelleher *et al.*, 2017). This range in pairwise F_{st} value was smaller than that reported for sheep in the present study (0.040 to 0.282) and indicates greater genetic diversity between sheep breeds which may contribute to a greater number of SNP required for accurately predicting breed composition in sheep. The number of breeds included in the analysis could also influence the number of SNPs required to estimate breed composition.

Composite breeds

The breed composition of composite breeds was also successfully predicted using SNP-BLUP. The breed composition of the initial Belclare was described as 45% Lleyn, 32% high fertility commercial line, 18% interbred Finn \times Galway, and 5% Galway (Rasali *et al.*, 2005). The Texel was later introduced (Hanrahan, 2002). The range of Galway proportion in the Belclare animals in the present study was 0.049 to 0.191 which is similar to the original Galway proportion in the initial formation of the Belclare, while the range of Lleyn proportion in Belclare animals (0.053 to 0.271) was lower than that of the initial composite breed. The breed composition of the modern Highlander (<https://www.focusgenetics.com>) has been


suggested as 25% Romney, 25% Texel, and 50% Finn. The Texel proportion estimated in the present study (0.095 to 0.158) was comparable to the breed composition description of the Highlander, while the Finn proportion (0.116 to 0.155) observed was lower than that previously suggested.

Conclusion

The breeds that were genetically most diverse from each other were the Suffolk and Border Leicester (pairwise F_{st} of 0.282). The Irish Blackface and Scottish Blackface formed overlapping, integrated clusters in the PCA and had the lowest pairwise F_{st} between breeds (0.040); these breeds could subsequently be considered as the same breed genetically. Accurate breed composition was achievable using ≥ 5000 SNPs. Nevertheless, as the number of SNPs on the panel reduced, the threshold required to identify purebred animals must also reduce (≥ 0.90 threshold for 9334 SNPs compared to ≥ 0.85 threshold for 5000 SNPs).

Acknowledgements

This research was part of the OviGen project (14/S/849) funded by the Department of Agriculture, Food and Marine, Ireland. Funding from the H2020 SMARTER project (Grant agreement ID: 772787) is also gratefully acknowledged.

 A. C. O'Brien 0000-0002-9991-8967

Declaration of interest

Authors declare no conflict of interest.

Ethics statement

Ethics committee approval was not obtained because the data used in the present study were managed by a third party, Sheep Ireland.

Software and data repository resources

The data were not deposited in an official repository.

Supplementary material

To view supplementary material for this article, please visit <https://doi.org/10.1017/S1751731119002398>.

References

- Alexander DH, Novembre J and Lange K 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* 19, 1655–1664.
- Beynon SE, Slavov GT, Farré M, Bolormaa S, Waddams K, Davies B, Haresign W, Kijas J, MacLeod IM, Newbold J, Davies L and Larkin DM 2015. Population structure and history of the Welsh sheep breeds determined by whole genome genotyping. *BMC Genetics* 16, 65.
- Crum TE, Schnabel RD, Decker JE, Regitano LCA and Taylor JF 2018. CRUMBLER: a tool for the prediction of ancestry in cattle. Available at: <https://www.biorxiv.org/content/10.1101/396341v1.article-info>
- Dodds KG, Aurvay B, Newman SN and McEwan J 2014. Genomic breed prediction in New Zealand sheep. *BMC Genetics* 15, 92.

- Frkonja A, Gredler B, Schnyder U, Curik I and Sölkner J 2012. Prediction of breed composition in an admixed cattle population. *Animal Genetics*, 43, 696–703.
- Hanrahan JP 2002. Response to divergent selection for ovulation rate in Finn sheep. In Proceedings of the 7th World Congress on Genetics Applied to Livestock Production, 19th to 23rd August 2002, Montpellier, France, INRA, pp. 673–676.
- Judge MM, Kelleher MM, Kearney JF, Sleator RD and Berry DP 2017. Ultra-low-density genotype panels for breed assignment of Angus and Hereford cattle. *Animal* 11, 938–947.
- Kelleher MM, Berry DP, Kearney JF and McParland S 2017. Inference of population structure of purebred dairy and beef cattle using high-density genotype data. *Animal* 11, 15–23.
- Kijas JW, Lenstra JA, Hayes B, Boitard S, Porto Neto LR, San Cristobal M, Servin B, McCulloch R, Whan V, Gietzen K and Paiva S 2012. Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. *PLoS Biology* 10, e1001258.
- Kuehn LA, Keele JW, Bennett JW, Mc Daneld TG, Smith TPL, Snelling WM, Sonstegard TS and Thallman RM 2011. Predicting breed composition using frequencies of 50,000 markers from the US Meat Animal Research Centre 2,000 Bull Project. *Journal of Animal Science* 89, 1742–1750.
- Lawson Handley L-J, Byrne K, Santucci F, Townsend S, Taylor M, Bruford MW and Hewitt GM 2007. Genetic structure of European sheep breeds. *Heredity* 99, 620–631.
- McGovern SP, Purfield DC, Ring SC, Carthy TR, Graham DA and Berry DP 2019. Candidate genes associated with the heritable humoral response to *Mycobacterium avium* subspecies *paratuberculosis* in dairy cows have factors in common with gastrointestinal diseases in humans. *Journal of Dairy Science* 102, 1–15.
- McHugh N, Pabiou T, Wall E, McDermott K and Berry DP 2017. Impact of alternative definitions of contemporary groups on genetic evaluations of traits recorded at lambing. *Journal of Animal Science* 95, 1926–1938.
- MiX99 Development Team 2015. Biometrical genetics, Natural Resources Institute Finland (Luke), Jokioinen, Finland.
- McParland S, Kearney JF, Rath M and Berry DP 2007. Inbreeding trends and pedigree analysis of Irish dairy and beef cattle populations. *Journal of Animal Science* 85, 322–331.
- O'Brien AC, Judge MM, Fair S and Berry DP 2019. High imputation accuracy from informative low-density to medium-density single nucleotide polymorphism genotypes is achievable in sheep. *Journal of Animal Science* 97, 1550–1567.
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA and Reich D 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 38, 904–909.
- Purfield DC, McParland S, Wall E and Berry DP 2017. The distribution of runs of homozygosity and selection signatures in six commercial meat sheep breeds. *PLoS ONE* 12, e0176780.
- Rasali DP, Shrestha JNB and Crow GH 2005. Development of composite sheep breeds in the world: a review. *Canadian Journal of Animal Science* 86, 1–24.
- Santos BFS, McHugh N, Byrne TJ, Berry DP and Amer PR 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 JP and Schenkel FS 2014. A new approach for efficient genotype imputation using information from relatives. *BMC Genomics* 15, 478.
- Struckem EM, Al-Mamun HA, Esquivelzeta-Rabell C, Gondro C, Mwai OA and Gibson JP 2017. Genetic tests for estimating dairy breed proportion and parentage assignment in East African crossbred cattle. *Genetic Selection Evolution* 49, 67.
- Sölkner J, Frkonja A, Raadsma HW, Jonas E, Thaller G, Gootwine E, Seroussi E, Fuerst C, Egger-Danner C and Gredler B 2010. Estimation of Individual Levels of Admixture in Crossbred Populations from SNP Chip Data: examples with Sheep and Cattle Populations. *Interbull Bulletin* 42, 62–66.
- Twomey AJ, Berry BP, Evan RD, Doherty ML, Graham DA and Purfield DC 2019. Genome-wide association study for endo-parasite phenotypes using imputed whole-genome sequence data in dairy and beef cattle. *Genetics Selection Evolution* 51, 15.