frontiers in Genetics

METHODS
published: 15 March 2018
doi: 10.3389/fgene.2018.00084

SNP Data Quality Control in a National Beef and Dairy Cattle System and Highly Accurate SNP Based Parentage Verification and Identification

Matthew C. McClure[1]*, *John McCarthy*[1], *Paul Flynn*[2], *Jennifer C. McClure*[1], *Emma Dair*[1], *D. K. O'Connell*[1] and *John F. Kearney*[1]

[1] Irish Cattle Breeding Federation, Cork, Ireland, [2] Weatherbys Ireland, Kildare, Ireland

OPEN ACCESS

Edited by:
Riccardo Negrini,
Università Cattolica del Sacro Cuore,
Italy

Reviewed by:
Ignacio Aguilar,
National Institute of Agricultural
Research (INIA), Uruguay
Marco Milanesi,
Universidade Estadual Paulista Júlio
de Mesquita Filho (UNESP), Brazil

***Correspondence:**
Matthew C. McClure
mccluretiger@yahoo.com

Specialty section:
This article was submitted to
Livestock Genomics,
a section of the journal
Frontiers in Genetics

Received: 12 September 2017
Accepted: 27 February 2018
Published: 15 March 2018

Citation:
McClure MC, McCarthy J, Flynn P,
McClure JC, Dair E, O'Connell DK and
Kearney JF (2018) SNP Data Quality
Control in a National Beef and Dairy
Cattle System and Highly Accurate
SNP Based Parentage Verification and
Identification. Front. Genet. 9:84.
doi: 10.3389/fgene.2018.00084

A major use of genetic data is parentage verification and identification as inaccurate pedigrees negatively affect genetic gain. Since 2012 the international standard for single nucleotide polymorphism (SNP) verification in *Bos taurus* cattle has been the ISAG SNP panels. While these ISAG panels provide an increased level of parentage accuracy over microsatellite markers (MS), they can validate the wrong parent at ≤1% misconcordance rate levels, indicating that more SNP are needed if a more accurate pedigree is required. With rapidly increasing numbers of cattle being genotyped in Ireland that represent 61 *B. taurus* breeds from a wide range of farm types: beef/dairy, AI/pedigree/commercial, purebred/crossbred, and large to small herd size the Irish Cattle Breeding Federation (ICBF) analyzed different SNP densities to determine that at a minimum ≥500 SNP are needed to consistently predict only one set of parents at a ≤1% misconcordance rate. For parentage validation and prediction ICBF uses 800 SNP (ICBF800) selected based on SNP clustering quality, ISAG200 inclusion, call rate (CR), and minor allele frequency (MAF) in the Irish cattle population. Large datasets require sample and SNP quality control (QC). Most publications only deal with SNP QC via CR, MAF, parent-progeny conflicts, and Hardy-Weinberg deviation, but not sample QC. We report here parentage, SNP QC, and a genomic sample QC pipelines to deal with the unique challenges of >1 million genotypes from a national herd such as SNP genotype errors from mis-tagging of animals, lab errors, farm errors, and multiple other issues that can arise. We divide the pipeline into two parts: a Genotype QC and an Animal QC pipeline. The Genotype QC identifies samples with low call rate, missing or mixed genotype classes (no BB genotype or ABTG alleles present), and low genotype frequencies. The Animal QC handles situations where the genotype might not belong to the listed individual by identifying: >1 non-matching genotypes per animal, SNP duplicates, sex and breed prediction mismatches, parentage and progeny validation results, and other situations. The Animal QC pipeline make use of ICBF800 SNP set where appropriate to identify errors in a computationally efficient yet still highly accurate method.

Keywords: SNP, quality control, parentage, parentage prediction, ISAG200

INTRODUCTION

Since the 1960's bovine pedigree verification has been performed with various DNA technology, initially performed with blood groups (Stormont, 1967), then microsatellite markers (MS) (Davis and Denise, 1998), and now transitioning to single nucleotide polymorphisms (SNP) (Heaton et al., 2002). While the initial cost and availability of each new technology has hindered their adaption, their increasing ability to reduce pedigree errors cannot be ignored. A 10% pedigree error rate can have a 6–13% effect on the inbreeding coefficient, 11–18% reduction on breeding value trends, 2–3% loss in selection response (Banos et al., 2001; Visscher et al., 2002), and a downward basis on heritability estimates (Israel and Weller, 2000). While sire error rates have been estimated at >7% in national herds, dam errors and missing parental information can be substantial, especially in commercial herds (Harder et al., 2005; Sanders et al., 2006) and their effects are additive (Sanders et al., 2006).

While with all technology there is a need to balance cost with performance; for parentage validation the question has typically been how many markers are needed to obtain a high probability, but not necessarily 100%, that the reported parentage is correct. The International Society of Animal Genetic (ISAG) recommended parentage SNP panel of 100 SNP (ISAG100) has a reported parental exclusion probability (PE) of >0.999 and the ISAG200 panel (200 SNP) has a PE >0.9999999 (http://www.isag.us/). Many groups world-wide primarily, or only, use the ISAG100 or ISAG200 panel for initial bovine parentage validation, and some groups use less. While the PE values for the ISAG SNP panels appear sufficient for accurate parentage, Vandeputte (2012) notes that many reported PE values are overly optimistic, that increasing numbers of markers are needed to maintain the same PE value as the population size increases, and a marker set with a high PE value can still have a low probability of complete exclusion of all false parentage with large.

These issues with PE values could be one of the reasons why we and others have reported (McClure et al., 2015; Strucken et al., 2015) that using lower density parentage SNP panels like the ISAG100 and ISAG200 can result in false-positive validations and result in multiple parents being predicted when used in large population datasets. While there currently is no international standard for which or how many SNP to use for parentage outside of the ISAG set, we argue that a larger SNP set should be used such as the 800 SNP set (ICBF800) that ICBF has developed and uses for parentage validation and prediction (McClure et al., 2015). The ICBF800 was selected to be highly accurate for parentage use across multiple *Bos taurus* breeds and has also proven useful for sample quality control (QC).

As the Irish Cattle Breeding Federation's (ICBF) genotype database has rapidly increased since 2013, from ∼25,000 to currently having >1,000,000 animals genotyped, data quality has become a higher concern, especially as "once in a million" and "rare" types of errors are encountered. Most publications only perform SNP and individual genotype QC based on SNP QC via call rate (CR), minor allele frequency (MAF), and Hardy-Weinberg deviation (Turner et al., 2011). While Wiggans et al. (2011) briefly described additional QC steps they used, ICBF has developed their own QC pipeline to deal with other issues, some foreseen others unique, which we describe in full below. We hope that the full description of our QC pipelines and parentage process will be useful to other groups as they grapple with larger genotyping datasets, regardless of the species.

The main QC concerns for ICBF are (1) is this genotype ok and (2) does the genotype really belong to the listed animal. While most errors come about due to accidents or non-malicious actions, a very small amount could be due to intentional actions. We have tried to design a QC pipeline will identify both types of errors once enough data is collected. An example of potential errors includes:

A) Farmer

1) Calmer animal sampled as requested animal is dangerous
2) Same animal sampled >1 times but labeled differently
3) Wrong animal sampled

B) Laboratory

1) Genotype duplicate due to technician error on sample
2) Genotype assigned to wrong sample

C) AI center

1) Wrong label attached to AI straw
2) Wrong animal sampled
3) Unreported sex selected semen

D) Genotype format

1) Type (AA, AB, or BB) is missing, or low frequency
2) Format is wrong, e.g., mix of AB and ACTG format

MATERIALS AND METHODS

Animal Care and Use Committee approval was not obtained for this study because the data were obtained from the existing ICBF database, Bandon, Co. Cork, Ireland.

ICBF Animal and Genotype Database

The ICBF database was set up in 1998 and holds numerous records on all dairy and beef cattle in Ireland including, but not limited to, date of birth, sex, reported dam, reported sire or sire breed, parentage validation status and method, animal movement, date of death, pedigree based breed composition, milk recordings, carcass data, along with multiple other phenotypes. Eighty-five *B. taurus* breeds are represented in Ireland, 58 beef and 27 dairy, with 44 breeds having at least 1 purebred animals genotyped in the ICBF database and 61 breeds being represented in a purebred or crossbred genotyped animal (**Table S1**). Data is reported to the database from, but not limited to producers, Department of Agriculture, Food, and Marina (DAFM), marts, abattoirs, veterinarians, AI technicians, milk co-ops, and herd books. SNP genotype data is also reported to ICBF on Irish animals from commercial genotyping labs and via international collaborations. At submission, the ICBF database holds valid genotypes on >1.20 million individuals, while animals with pedigree and phenotype records stands at >40.5 million.

Genotyped animals represent a mixture of male and female Irish beef, dairy, purebred, crossbred, pedigree, and commercial animals from 59 *B. taurus* breeds who were genotyped on multiple SNP chips such as the Illumina 3K, LD, 50K, and HD; GeneSeek GGPLD and HD; and our custom International Beef and Dairy (IDB) chip (Illumina Inc., 2009, 2010, 2011a,b; Matukumalli et al., 2009; Neogen Corporation, 2012, 2013; Mullen et al., 2013). As the Illumina 3K has not been commercially available since September, 2011 (personal communication, André Eggen, 23/2/2015), and with its higher variability in genotyping accuracy (Wiggans et al., 2012) 3K genotypes were not used for the analysis or pipeline listed below.

The ICBF Genotype QC, Parentage, and Animal QC pipelines are described below.

ICBF PARENTAGE AND QC PIPELINES

Genotype QC Pipeline

Genotype quality is one of the easiest and most used QC tools used. All genotypes received at ICBF go through our Genotype QC pipeline (**Figure 1**) regardless if they were received from a genotyping lab or exchanged from another national evaluation center. At ICBF we use an animal call rate (CR) of ≥0.90 for a genotype to be used as genotype concordance rates falls below 99% when the CR is <90% (Cooper et al., 2013). The calculation of this individual CR does not include SNP that have CR <0.85 across our database.

Next, we check for missing genotype classes, often BB, or mixed genotype classes, e.g., ABTG. While genotypes received should all be in a standard format, for instance Illumina AB allele or Top ACTG allele format, ICBF has received genotypes before that have had a mix of genotype formats or missing genotypes. If mixed or missing genotypes are found the genotype is invalidated.

Finally, we invalidate any genotype that has a genotype class (e.g., AA, AB, or BB) frequency below 20%. This last check came about as an analysis of 846,868 animals with ≥90% CR and no missing or mixed genotype classes revealed that <0.0001% of animals had a genotype (AA, AB, or BB) frequency lower than 20% (**Table 1**). As far as we know this is the first time the pattern of individual genotype frequencies has been analyzed. While highly inbred animals would have reduced AB frequency, this analysis strongly indicates any animal with a genotype class frequency <20% should be flagged for further analysis.

A genotype must pass all genotype QC quality checks to be used downstream for any other process.

Parentage Pipeline

Parentage: SNP Panel, Validation, Prediction, and Suggestion Pipeline

ICBF800 parentage SNP panel

The ICBF800 was developed to provide a set of highly informative SNP for parentage verification and prediction. As described in McClure et al. (2015), we identified >1 sire can be predicted at a 1% misconcordance rate using the ISAG100 or ISAG200 parentage SNP sets. In that study, we identified that ≥500 SNP with high MAF are needed to only predict 1 sire from a large database. To give a buffer between failing and verifying parents, 800 SNP (ICBF800) were selected based on ISAG200 membership, being part of the Illumina LD base content, clustering quality, and the SNP's MAF and CR in the Irish cattle population. The ICBF800 SNP set was reanalyzed in August 2016 after >500,000 cattle were genotyped and the breeds represented in the genotype database were more balanced (**Table 2**). The ICBF800 set identified in August 2016 is what ICBF currently uses, was used for the rest of this manuscript, and are identified in **Table S2**. ISAG200 and Illumina LD SNP that had clustering issues, call rates under 90%, or MAF <0.25 were excluded from the ICBF800 panel (**Table S2**, problem cluster examples in **Figure S1**). All non ISAG SNP chosen have MAF >0.42 across >800,000 animals and an average MAF of 0.36 in 145,664 purebred animals representing 25 breeds (**Tables S2**, **S3**).

Parentage analysis

All animals with valid SNP genotypes received by ICBF that pass the genotype QC process, are automatically sent through our Parentage Pipeline (**Figure 2**). Depending on the animal type (commercial, pedigree, or AI sire), parentage results, and owner or herd book requests an animal will exit the pipeline at different points. SNP based parentage validation and prediction are performed using the ICBF800 parentage SNP panel (**Table S2**) as it is more accurate than the ISAG100 or ISAG200 parentage panels as described below and in (McClure et al., 2015). A parentage is SNP validated or predicted if <4 mismatches (0.5% misconcordance rate) are recorded, a parentage fails if >12 mismatches (1.5% misconcordance rate) are recorded. Misconcordance counts of 5–12 mismatches, 0.5–1.5% misconcordance levels, are further analyzed by checking the misconcordance rate of all SNPs where both animals have a genotype. When all SNP are checked if the misconcordance rate is ≤1% the parentage validates, if >1% it fails.

Microsatellite imputation

Microsatellite (MS) imputation (McClure et al., 2013) is performed using a set of 921 SNP (**Table S4**) for pedigree animals whose parentage is not SNP validated or predicted but are MS genotyped. When requested a Genetic Relationship Matrix (GRM) via SVS version 8.8.0 software (Golden Helix, Montana, USA) is performed using the Illumina LD SNP panel to identify close genetic relatives to suggest potential non-genotyped parents.

Mating validation

When both parents SNP validate the mating is also validated by identifying any Mating SNP Misconcordances (MSM). MSMs are where the calf was AB and both sire and dam were homozygous for the same allele. If MSM rates >1% for the ICBF800 are identified all animals in the trio are manually checked.

Animal QC Pipeline

Genotypes that have passed through the Genotype QC pipeline are then sent through the Animal QC pipeline (**Figure 3**). All

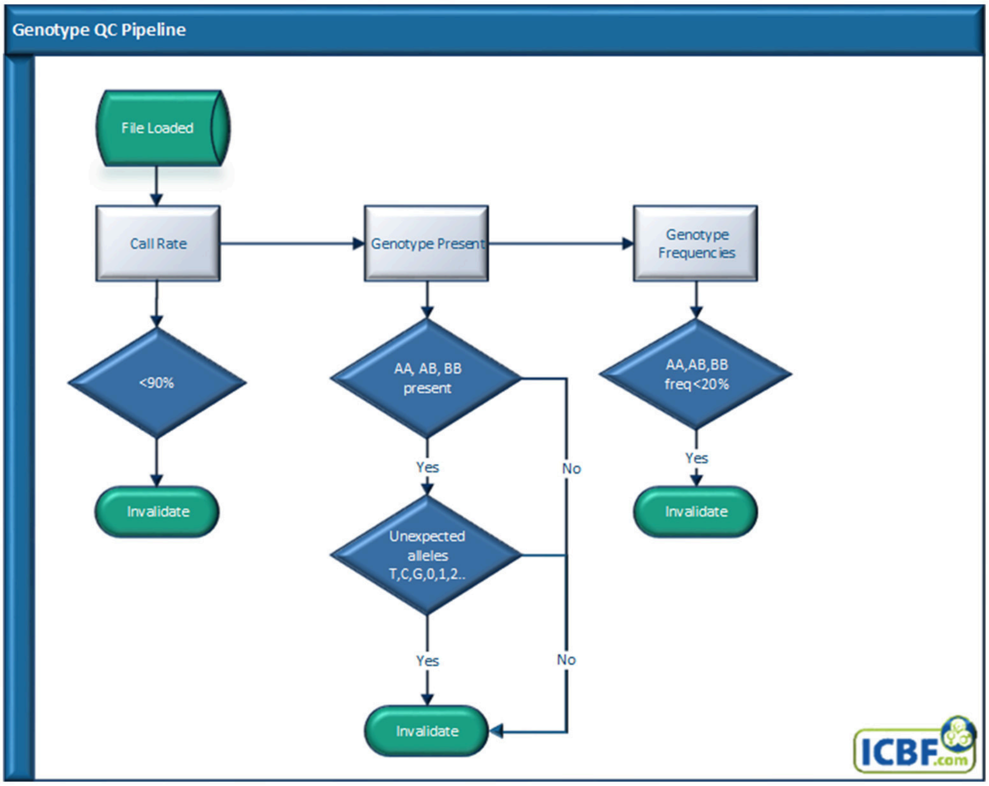

FIGURE 1 | Genotype quality control pipeline.

samples are fully sent through the entire Animal QC pipeline before a decision is made to invalidate a genotype. Overall a failure at any of the Animal QC checkpoint does not by itself cause a genotype to be invalidated but their combined results can.

SNP Duplicate Check

Excluding identical twins, two animals should not have the same genotype across large numbers of SNP. Considering that only 1.6% of bovine births are of identical twins (Mcclure et al., 2017) the identification of two animals with the same genotype is a likely indication of either a farm or lab error. The former being that the same animal was sampled twice, possibly with a decent amount of time between tissue collections when a different animal was requested, the latter usually occurring within a narrow time frame. Lab errors can often be detected by analyzing the genotypes of all animals processed during a set time frame, such as daily or weekly. Farm error identification requires analyzing all of the farm's DNA genotypes. Finally, you could have a farm level error combined with a mislabeling of the DNA collection device (usually labeled by a 3rd party), to detect these you would need to analyse all genotypes in the national database.

Comparing the full SNP genotype of an animal against hundreds of thousands or millions of records is possible but far too time consuming. To speed up this process, we use the ICBF800 to identify potential SNP duplicates and then compare all available SNP to confirm which potential duplicates are true duplicates. Additional QC information is used to determine who the genotype truly belongs to.

Animal With >1 Non-matching Genotypes

Animals can be genotyped more than once within and across countries for multiple legitimate reasons. Regardless of the number of times an animal is genotyped any SNP common across the SNP platforms and chip types should generate the same genotype, for instance Illumina and Affymetrix SNP chips have >99% concordance rates across platform and tissue types (Montgomery et al., 2005; Feigelson et al., 2007; Woo et al., 2007; McClure et al., 2009). If the concordance rate between two genotypes assigned to an animal is <99% then this indicates they are probably not from the same animal and one should be able to determine which genotype truly belongs to the individual. The ICBF800 panel is used to identify potential issues and then all available genotypes are used to confirm. Additional QC information is used to determine who the genotype truly belongs to.

Sex Prediction and Pseudoautosomal Region of Chromosome X Determination

Sex prediction is performed using chromosome Y (chrY) and chromosome X (chrX) SNP that are located in the non-pseudoautosomal (nPAR) region. Sex prediction using only chrY SNP is logically simpler to use, but not all commercial SNP chips

TABLE 1 | Count of animal's genotype frequency for AA, AB, and BB genotypes from 846,868 genotyped individuals with a genotype CR>0.9 and all 3 genotypes present.

Frequency	AA	AB	BB
0	–	–	–
0.005	–	–	–
0.01	–	–	–
0.02	–	–	–
0.03	1	–	–
0.04	–	–	–
0.05	–	–	–
0.06	1	–	–
0.07	–	–	–
0.08	2	–	–
0.09	–	–	–
0.1	–	–	–
0.15	–	1	–
0.2	2	68	1
0.25	1,977	2,745	6
0.3	6,20,489	57,372	44,438
0.35	2,23,499	5,64,809	1,53,114
0.4	885	78,550	6,37,289
0.45	2	1,38,927	12,002
0.5	–	4,385	9
>0.5	1	11	9

TABLE 2 | Percent of major breed represented the total ICBF genotype database by date.

Breed[a]	6/2014	3/2015	3/2016	4/2017
AA[b]	4.41	7.22	9.51	9.58
AU	1	0.53	0.60	0.61
BA	0.04	0.59	0.77	0.81
BB	1.01	2.75	3.63	3.63
CH	9.09	19.2	19.94	21.48
HE	3.23	4.73	5.13	5.19
HO	68.02	30.08	13.89	13.63
JE	0.17	0.67	0.81	0.52
LM	9.72	22.93	30.67	31.5
MO	0.14	0.05	0.16	0.17
PI	0.54	0.17	0.21	0.21
PT	0.17	0.64	0.67	0.7
SA	0.04	0.92	1.70	1.72
SI	0.06	1.57	1.90	1.76
SM	2.34	6.77	8.12	7.75
Total[c]	99.98	98.82	97.7	99.26

[a] *The breed represents the animal's major breed component, so an animal that is 75% LM and 25% HO is counted as a LM individual.*

[b] *Breed abbreviations defined in **Table S1**.*

[c] *Total values do not add up to 100% as not all breeds are represented.*

Crossbred animals are counted by their main breed. F1 animals (50% breed 1 and 50% breed 2) are counted under which breed is alphabetically first.

contain chrY SNP. Sex prediction only using the heterozygosity rates of chrX SNP can be more challenging as care must be taken to avoid using PAR SNP and highly inbred females such, as L1 Dominette 01449 (Henderson et al., 2005), can look like males if not enough SNP are used. Also sex-selected semen (often female selected) can have odd results as the low chrY content can result in no chrY genotpyes due to low signal intensity caused by the low number of chrY containing semen cells.

As the Illumina LD base content is widely used across multiple commercial and custom bovine SNP chips ICBF uses 7 chrY SNP from the LD chip chrY sex prediction as in 4,901 HD genotyped animals (10% female) they were homozygous in all males and not present in the females. Those chrY SNP were BOVINEHD3100000048, BOVINEHD3100000099, BOVINEHD3100000103, BOVINEHD3100000210, BOVINEHD3100000517, BOVINEHD3100001188, and BOVINEHD3100001406.

SNP in the PAR of the X chromosome was determined by analyzing 467 SNP with unique locations on chrX (UMD3.1 assembly) in 606,122 IDBv3 genotyped animals. After filtering for MAF (<0.01) and CR (<0.90) the PAR region was determined via SNP with high male heterozygosity rates and genomic positions (**Figure 4**, **Table S5**). Male and female heterozygousity rates were determined in those animals to determine chrX sex prediction thresholds.

The current logic used by ICBF for sex prediction is as follows:

1) Predict sex with nPAR chrY SNP
    (a) Count nPAR chrY genotypes
    (b) If 0–1 genotypes = female
    (c) If 6–7 = male
    (d) If 2–5 = ambiguous sex
2) Predict sex with nPAR chrX SNP
    (a) Determine heterozygosity rate (# AB/ (#AA + #AB + #BB)) for nPAR SNP
    (b) If ≤5% het rate = male
    (c) If ≥15% female
    (d) If between 5 and 15% = ambiguous sex
3) Do chrX and chrY sex predictions match.
    (a) Yes—report sex predicted
    (b) No—manual check genotype results
4) If step 1 = ambiguous and step 2 = male or female (or inverse) then report non-ambiguous predicted sex
5) If step 1 and 2 are both ambiguous = report ambiguous sex predicted

Animals with non-matching chrX and chrY sex predictions could be Turner (X0) or Klinefelter's syndrome (XXY) animals (Berry et al., 2017) or they could be caused by genotyping a straw of sex selected semen.

Breed Composition Prediction

The breed composition of an animal is predicted via Admixture v2 (Alexander et al., 2009) in a supervised analysis. To maximize predictive analysis 36,819 SNP are used that map to the

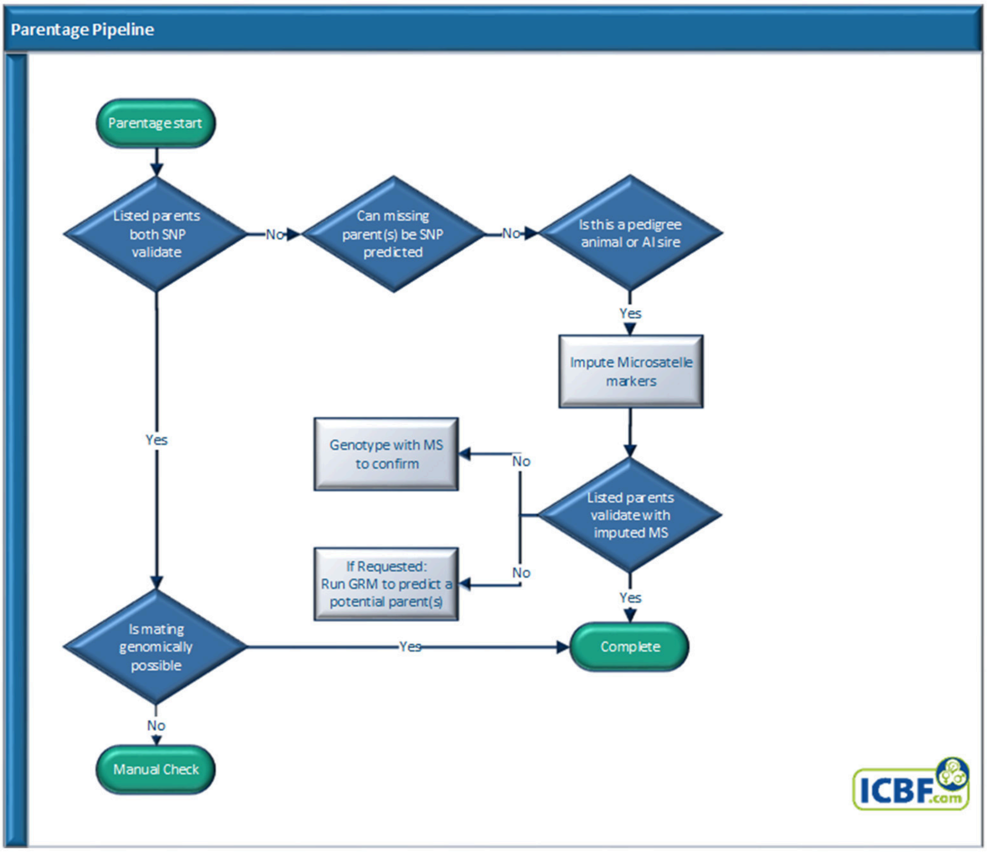
Parentage Pipeline
Parentage start
Listed parents both SNP validate
No
Can missing parent(s) be SNP predicted
No
Is this a pedigree animal or AI sire
Yes
Impute Microsatelle markers
Genotype with MS to confirm
No
Listed parents validate with imputed MS
No
If Requested: Run GRM to predict a potential parent(s)
Yes
Yes
Is mating genomically possible
Yes
Complete
No
Manual Check
ICBF.com

FIGURE 2 | Parentage pipeline.

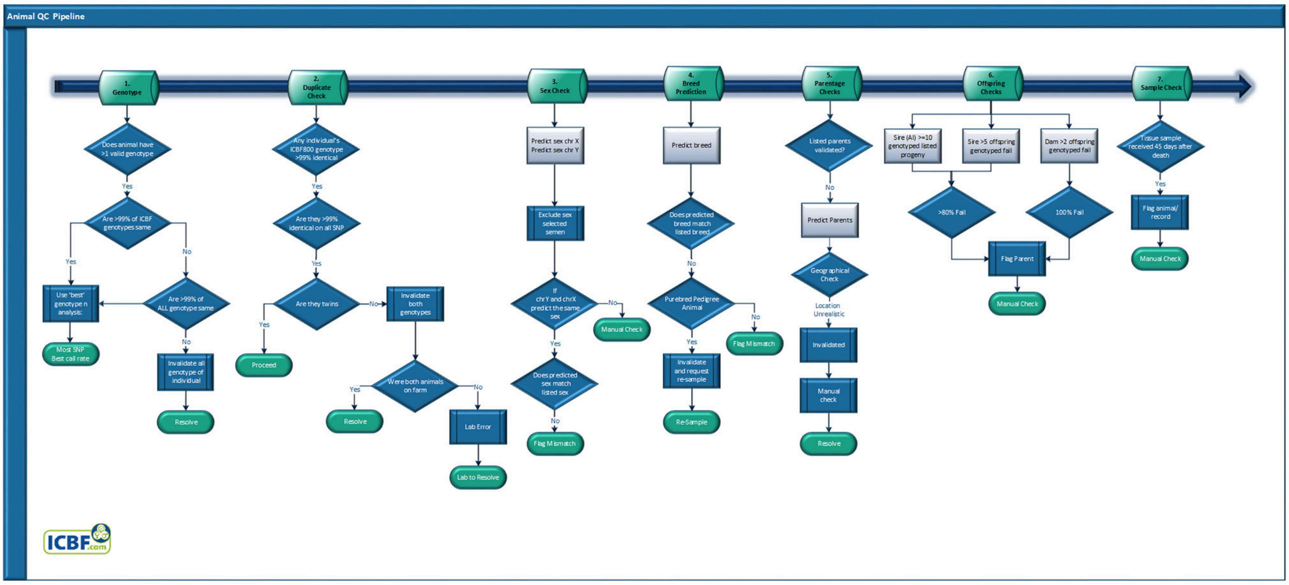
Animal QC Pipeline
1. Genotype
2. Duplicate Check
3. Sex Check
4. Breed Prediction
5. Parentage Checks
6. Offspring Checks
7. Sample Check
Does animal have >1 valid genotype
Yes
Are >99% of ICBF genotypes same
Yes
No
Use 'best' genotype n analysis:
Are >99% of ALL genotype same
No
Most SNP Best call rate
Invalidate all genotype of individual
Resolve
Any individual's ICBF800 genotype >99% identical
Yes
Are they >99% identical on all SNP
Yes
Are they twins
Yes
No
Invalidate both genotypes
Proceed
Were both animals on farm
Yes
No
Resolve
Lab Error
Lab to Resolve
Predict sex chr X Predict sex chr Y
Exclude sex selected semen
If chrY and chrX predict the same sex
No
Manual Check
Yes
Does predicted sex match listed sex
No
Flag Mismatch
Predict breed
Does predicted breed match listed breed
No
Purebred Pedigree Animal
Yes
No
Invalidate and request re-sample
Flag Mismatch
Re-Sample
Listed parents validated?
No
Predict Parents
Geographical Check
Location Unrealistic
Invalidated
Manual check
Resolve
Sire (AI) >=10 genotyped listed progeny
Sire >5 offspring genotyped fail
Dam >2 offspring genotyped fail
>80% Fail
100% Fail
Flag Parent
Manual Check
Tissue sample received 45 days after death
Yes
Flag animal/ record
Manual Check
ICBF.com

FIGURE 3 | Animal quality control pipeline.

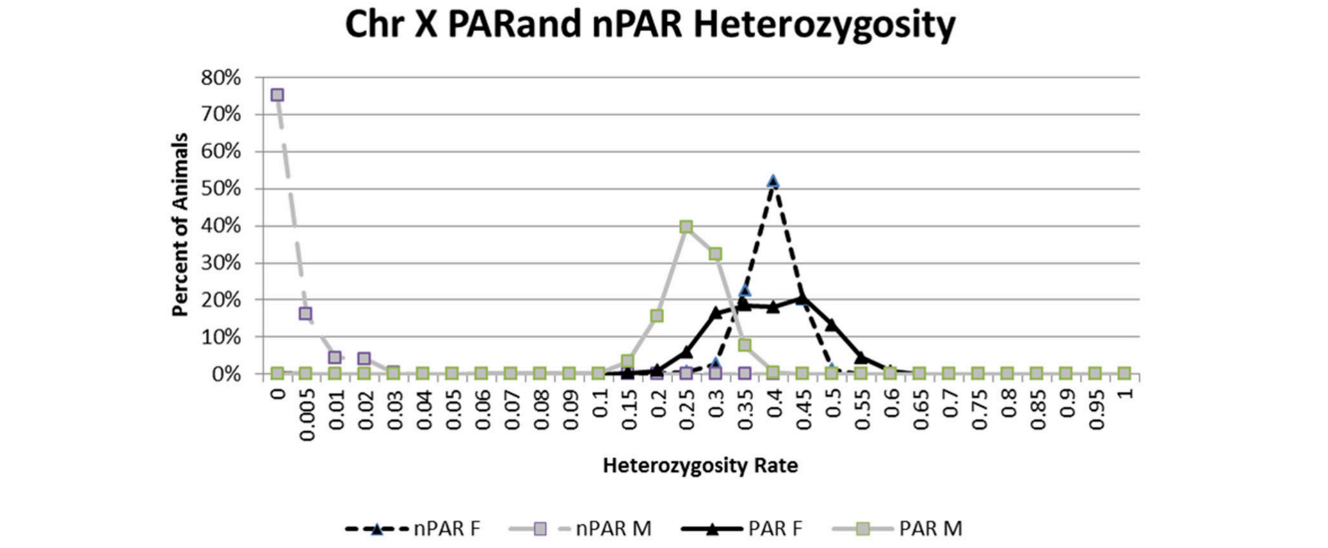

FIGURE 4 | The percent of 37,281 males and 568,641 females at each heterozygosity level for chromosome X SNP that are present pseudoautosomal region (PAR) and non-pseudoautosomal regions (nPAR) of chromosome X.

autosomes on the UMD3.1 assembly (Zimin et al., 2009) and are common across the 50k, HD, and IDBv3 panels. A set reference population is used that is comprised of 22,610 purebred animals from 14 breeds, Angus, Aubrac, Blonde D'Aquitane, Belgian Blue, Charolais, Friesian, Hereford, Holstein, Jersey, Limousin, Parthenaise, Salers, Shorthorn, and Simmental, with a minimum of 500 and a maximum of 2,000 animals per breed. Animals genotyped on lower density chips were not used in the reference population to maximize the number of SNP used because their inclusion would not have increased the number of reference breeds, nor greatly increased the animals in breeds with <2,000 animals. Breeds with >2,000 genotyped purebreds had their breed composition reference animals selected randomly when the reference population was initially defined. Breeds with lower numbers of purebred genotyped animals were tried but not used do to their low prediction accuracy, often <50% correlation to the animal's reported breed composition. While smaller number of reference animals per breed could have resulted in similar breed prediction accuracies we chose go with 500 to 2,000 reference animals per breed to ensure proper representation of any genomically diverse, yet pure, breed (e.g., see the PCA plot of Herefords and Holsteins in **Figure 5**). We didn't go above 2,000 per breed to keep the reference population relatively balanced, as it's currently designed any single breed only represents 2–8% of the total reference population.

Pedigree animals with a non-matching listed and predicted breed composition (e.g., 100% Limousin vs. 100% Angus) have their genotype invalidated and either the sample is regenotyped (if probable lab error) or a new tissue sample is requested (if probable farm error). For all non-Pedigree animals, a non-matching breed composition results in the genotype being flagged. Breed predictions are mainly used to help resolve SNP duplicate samples.

Parentage

National bovine pedigree errors of 4–13% have been reported (Visscher et al., 2002; Leroy et al., 2012) based upon farmer recorded pedigree information. In Ireland, a 7–9% listed pedigree error rate is seen on a national level. While one listed parent may be listed incorrectly, it is rare that both listed parents would be wrong. Even less likely, though possible, is that the true parent(s) do not reside on or close to the farm, excluding AI sires. While the true sire might be an intact young bull, or neighboring stock bull the sire is usually geographically located in the same area, excluding AI sires. ICBF takes into account the geographic location of the sire and dam, along with their breeds, to determine if the predicted parents for an animal are logical.

Offspring

As a genotype can appear correct, having passed SNP QC and the already listed Animal QC steps, it can still be wrongly assigned to an animal. For instance, if fraternal twins, of the same sex, are genotyped but the genotype is assigned to the other animal. In these cases, the error will only become apparent when their offspring are genotyped. If a dam has ≥2 genotyped offspring and all fail, if a stock bull has ≥5 genotyped offspring and 80%, or if an AI bull has ≥10 genotyped offspring and 80% the dam's/sire's genotype is flagged for a manual check. If its deemed that the animal's genotype should be invalidated then all of the parentage results based on that now invalid genotype are reset.

Additional Flag

While possible if a tissue sample is sent in to be genotyped >45 days after the animal's death is recorded ICBF records this as a warning flag. The genotype is not invalidated by this but this flag is taken into account for genotype resolution if needed, for instance for SNP duplicates**.** AI sires are excluded from this flag if an AI straw is the submitted sample.

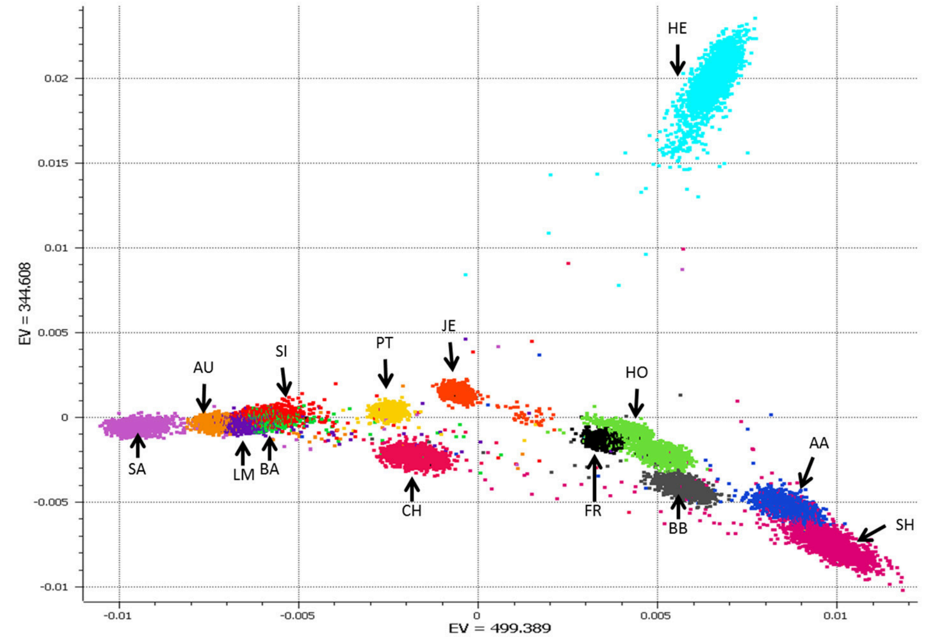

FIGURE 5 | Plot of principle components (PC) 1 and 2 from a PCA analysis of the 22,610 reference animals from 14 breeds. Breed abbreviation listed in **Table S1**.

APPLICATION RESULTS AND EFFECTIVENESS

Parentage SNP Panel

ICBF keeps a record of the number of SNP mismatches for all parentage validation checks against the listed parents. By 20/12/2016 775,390 parentage validation checks had been performed for 578,963 genotyped animals. 196,428 animals had parentage validation performed on both listed parents. Of them 6.10% of the listed parents failed with 13 or more SNP misconcordances, 0.02% of the listed parents fell into a "gray zone" with 5–12 SNP misconcordances, and 93.88% of the listed parents validated with 0–4 SNP misconcordances. Of the validated parents 94.40% had 0 SNP misconcordances, 5.30% had 1 SNP misconcordances, 0.26% had 2, 0.03% had 3, and <0.01% had 4.

Analysis of the 46 parentage doubtful "gray zone" animals using all available SNP resulted in 10 animals having a different result when all SNP were used vs. the ICBF800 (**Table S6**). These 10 animals had either7, 8, 9, or 10 SNP misconcordances on the ICBF800 panel. Animals with 5, 6, 11, or 12 SNP mismatches on the ICBF800 panel had the same parentage results for when all SNP were used. By having a 2-step parentage validation process for any animal with 0.5–1.5% misconcordance from ICBF800 parentage SNP, provides ICBF with an extremely accurate parentage validation process while minimizing computational requirements.

Fast Parentage SNP Prediction

For any animal who does not have both its sire and dam SNP validate parentage prediction is ran using the ICBF800 panel. This includes animals with a SNP failed listed parent, a SNP ungenotyped listed parent, or no listed parent. Essentially every animal with a valid genotype is checked to see if it could be the parent of the animal. To increase computing speed and keep high accuracy the following procedure is used.

First we create a table that contains the ICBF800 genotypes for all animals, this table is updated daily as new genotypes are received or bad genotypes invalidated. This table contains all animals that have a valid genotype and ≥600 genotypes of the ICBF800 panel each animal is represented only once. The number of SNP mismatches are calculated for each animal in the table, when over 12 mismatches are found the animal has "failed" and is excluded from the analysis. If all 800 SNP are analyzed for an animal and there are <12 SNP mismatches then a date of birth and a sex check is performed. If the predicted parent is less than 15 months older than the animal it is excluded, this removes any genotyped progeny of the animal while allowing young uncastrated bulls in the herd to be included. The logic of using 15 months is that the parent was ≥6 months old and 9 months for gestation. While cattle normally reach puberty at 9–10 months of age (Gasser et al., 2006; Rawlings et al., 2008), precocious puberty in heifers at 6 months of age has been observed (Wehrman et al., 1996) and there can be variability between an animal's true and recorded date of birth. The sex check ensures that the animal's dam is not predicted as its sire and vice versa this also allows an animal's sire and dam to be predicted together. Any predicted parent with 5–12 mismatches has all available SNP checked after the full table comparison is done.

Predictions are done in batches of 300, so that however many are requested it just peels off the next 300 and runs those. It was found that in the current setup the process was linear, so it was preferred to save off each 300 once done. This way if a prediction process is interrupted for any reason at most only 300 animal's predictions would be affected.

MS Imputation

While not perfect MS imputation is around 95% accurate (McClure et al., 2014) based upon its use in Ireland. Recently a 91% accuracy for MS imputation was obtained for Slovenia Brown Swiss (Obsteter, 2017) The difference in accuracy may be because ICBF has added over 1,500 more animals to the MS reference population that is publicly available in McClure et al. (2015). Animals whose listed parents fail via imputed MS data are directly MS genotyped to confirm or deny the imputed MS result. These animals are then routinely added to the reference MS imputation dataset as they have both SNP and MS genotypes. Imputed MS profiles have also been used as part of the QC pipeline to see if the imputed MS and genotyped MS profiles match. If not then sample is invalidated as the tissue for the SNP profile may not have come from the listed animal. As more animals are genotyped the amount of parentage verification via MS decreases.

GRM

When parentage cannot be determined via SNP or MS based methods, GRM becomes very useful for identifying close relatives and from that determining a most likely parent. We caution that GRM values can be inflated for inbred animals. GRMs are performed using Golden Helix SVS and the Illumina LD SNP panel. When needed, especially for highly inbred animals, higher SNP density panels are used. Results are used to suggest potential parents of the animal based on the listed pedigree of the identified closely related animals. Suggested parents are not viewed as being validated parents by ICBF, but provide a point of reference for breed societies, farmers, and potential animals to MS parentage check by other groups. Future work will see if GRM values can help identify who a SNP duplicate genotype truly belongs to when parentage validation, offspring validation, sex prediction, and breed composition prediction do not provide a clear answer.

Mating Validation

In May, 2017 ICBF had ~950,000 animals with a valid genotype, of them 195,299 trios existed with all individuals SNP genotyped and the offspring SNP validates to each parent separately. MSM, where the calf was AB and both sire and dam were homozygous for the same allele, were recorded. Of those 96.964% had 0 MSM, 3.000% had 1–3 MSM, 0.037% had 4–7 MSM, and 0.003% ($N = 5$) had 44 to 74 MSM.

For those 5 animals, from 5 different farms, with >40 MSM, we were able to resolve them using a combination of breed composition, progeny SNP validation, and GRM results along with ICBF's genotype tracking data. For each farm, the DNA kits were sent for the animal and dam on the same day and they were returned to the lab on the same day. It was determined that the cow and calf's DNA samples were swapped on the farm. Many of the cows had 2 genotyped progeny where 1 had failed SNP parentage. The failed calf SNP validated to the other calf based on the ICBF800 panel. One calf was listed as being 11% Shorthorn, 84% Limousine, and 5% unknown; its dam was listed as 22% Shorthorn, 68% Limousine, and 10% unknown; while the sire was listed as 100% Limousine. The breed predictions for the same animals were: calf as 36% Shorthorn, 55% Limousine, 9% unknown; dam as 20% Shorthorn, 75% Limousine, and 5% unknown; the sire was 100% Limousine. Given that the breed prediction for the calf matched the listed breed composition for the dam and vice versa this indicated the calf and dam's DNA were switched. For the animal's sire all of them had a listed sire that SNP failed validation and a predicted sire. Most of the dams had a listed sire that SNP failed, or didn't have a SNP genotype, and all had predicted sires. Essentially the calf's true sire was predicted as the dam's sire and vice versa. As the listed and predicted sires were older stock bulls or AI sires the predictions passed our QC checks.

SNP Duplicate and >1 Genotype for 1 Animal

Both processes use the ICBF800 parentage SNP set and the logic is similar. When an animal's genotype comes in, if it already has a genotype in the ICBF database the ICBF800 genotypes are compared. If <99% of them match then all available SNP are compared. If <99% of all SNP do not match then both genotypes are flagged as invalid until resolved.

To identify SNP duplicates the ICBF800 SNP genotypes are converted to a numeric string, for instance 001290021 where 0 = BB, 1 = AB, 2 = AA, 9 = missing, and one searches for exact matches of the entire string. Initially, ICBF did use the full string but quickly realized that random missing genotypes would cause some true SNP duplicates to be missed using an exact sting match. Therefore, we broke the 800 SNP string into 16 50-SNP non-overlapping blocks and performed exact string matches within each block. If a SNP duplicate was identified for any of the 16 blocks that pair had all SNP of the ICBF800 compared directly, and separately with all available SNP were compared, missing genotypes excluded. If >99% of the ICBF800 SNP were identical the genotypes were analyzed further.

This process worked well, but in February 2016, a SNP duplicate case was identified where random missing SNP caused not even one of the 16 50-SNP block to be exactly identical. This lead to using 40 20-SNP blocks to identify potential SNP duplicates. Checking all ~840,000 genotyped animals in March 2017, the process for identifying potential SNP duplicates took 1 min to run for 50-SNP blocks compared to 6 min for 20-SNP blocks. The 50-SNP block check found 459 potential SNP duplicates and the 20-SNP blocks found 610,172. Once potential SNP duplicates are found all ICBF800 SNP compared SNP by SNP with missing SNP excluded to identify those that are ≥99% identical. This took 13 min to run for the 459 potential SNP duplicates and 45 min for the 610,172 potential duplicates. Once set up the SNP duplicate check only needs to compare newly submitted genotypes so the time required will greatly decrease. The 20-SNP blocks did find 79 cases of true SNP duplicates that were not identified using the 50-SNP blocks.

For animals that have >1 genotype in the ICBF database the ICBF800 is used to see if ≥99% of the genotypes are identical.

If they are not then all available SNP are checked to see if they are ≥99% identical, if not then the genotypes are both flagged as invalid until resolved. For all cases where two genotypes were ≥99% identical for the ICBF800 they were also ≥99% identical for all available SNP. Similarly, no case has occurred where >1 genotype per animal was not ≥99% identical for ICBF800 and not also ≥99% for all available SNP.

The genotypes invalidated above are resolved if the results from the parentage validation or offspring validation can clearly identify one genotype as valid and the other as invalid. If not resolved then sex prediction, breed prediction, and date tissue sample submitted information is used. A check on if both animals were on the same farm ever and when tissue samples collected is also performed, if animals have never resided on the same farm then this points to a potential lab caused error. If both animals have resided on the same farm but one was not present when DNA was collected from the other this can be used to resolve whose genotype is valid and whose is invalid.

ChrX PAR Determination

The average male heterozygosity rate for the PAR region is 23.92% while the nPAR region is 0.23% compared to 36.95% and 37.06 in females respectively. **Figure 4** shows a clear separation of the nPAR heterozygosity rates for the 37,281 males with 7 chrY genotypes and the 568,641 females with 0 chrY genotypes. When we exclude animals with <90% CR for the chrX SNP, 99.95% of the males have <5% PAR SNP heterozygosity rate and 99.87% of the females 99.87% have ≥10%. The average male heterozygosity rate for the PAR region is 23.92% while the nPAR region is 0.23% compared to 36.95% and 37.06 in females respectively (**Figure 4**).

The chrX nPAR and PAR SNP on the IDBv3 are present on multiple commercial available SNP chips. Across the commercially available SNP chips in the ICBF database (LD, 50k, HD, GGPLDv1, GGPLDv2, GGPLDv3, GGPHDv1, GGPHDv2, IDBv1, IDBv2, and IDBv3) the average number of these nPAR SNP is 225 and PAR SNP is 49 (**Table S5**).

Sex Prediction

Sex prediction can be used as an easy QC check to tell if the DNA genotyped truly came from the reported animal, e.g., if a docile cow was sampled instead of a dangerous stock bull. In 612,722 IDBv3 genotyped animals for 7 chrY SNP 568,841 animals had 0 chrY SNP genotypes; 3,109 (1 SNP); 216 (2 SNP); 171 (3 SNP); 138 (4 SNP); 168 (5 SNP); 1515 (6 SNP); and 38,564 (7 SNP). Some research groups have notices chrY SNP genotypes appearing in female Holsteins and this may be caused by an historic event of some part of the chrY transferring to chrX or an autosome (George Wiggins, personal communication, 2016). For females with 1–4 chrY genotypes, 940 have listed dams that have been genotyped on a chip with chrY SNP. Of those 940 animals, which represent a mixture of purebred and crossbred beef and dairy, only 1.81% of their dams also have 0–4 chrY SNP genotypes called, the remaining dams have 0 chrY SNP called (**Table S7**). Therefore, it is very likely that these low levels of chrY genotypes in females is caused by either genotype errors or very low contamination levels.

Breed Prediction

Breed composition prediction is performed for animals whose listed breed composition is comprised of one of 14 reference breeds. Even when >10,000 genotypes are received on a daily basis, ADMIXTURE runs fast enough to be used daily in a production setting. As ADMIXTURE provides breed composition percent, future work will consider if this can be applied to help improve genomic breeding value predictions. Correlations of 96.6% have been obtained for of the main breed percent between the ICBF database and 710,000 animals with breed composition predicted (**Table 3**). Animals that are composed of breeds not in the reference population get predicted as a seemingly random mixture of the reference breed. Up to 24 breed composition prediction was tried with additional breeds having <100 genotyped purebreds (some only 5 animals). Breed prediction for these additional breeds was not accurate (data not shown) so currently breed prediction only uses 14 reference breeds. PCA analysis of the selected animals from the 14 breeds showed good separation between breeds (**Figure 5**).

As shown in **Figure 5**, most of the breed prediction reference animals fit into close breed specific clusters. There are examples of some animals that are plotted away from their breed cluster. As all reference animals are also ran back through breed composition prediction we can see how their predicted breed compares to the listed breed composition in the ICBF database. If you look at **Figure 5** you'll notice a red dot in the middle of the space between the HE, HO, and JE clusters. According to the ICBF database this animal is 100% Charolais, while according to both the PCA plot and the ADMIXTURE results it is predicted to be 50% Charolais and 50% Hereford (the light blue cluster at the top). ICBF will investigate these "stray" PCA results to see if the animal is possible not a true purebred and should be excluded from the reference population. Even with these "stray" animals it is impressive how accurate the breed predictions are, which we believe are because we chose to use a relatively large number of animals per breed.

The accuracy correlation should be looked at with 3 notes: (1) The ICBF database breed composition is based on 32 discrete parts thus a purebred is 32/32 and each part (1/32) is a 3.125% step. ADMIXTURE's predictions are continuous, so it can predict an animal to be 87.4356% Angus; (2) The reported animal's breed composition in the database is not perfect; (3) The database calculation assumes that a ½ AA ½ HE bred to a ½ HO ½ LM would generate a ¼ AA ¼ HE ¼ HO ¼ LM animal, as we know that is not true due to recombination and Mendelian sampling.

DISCUSSION AND CONCLUSION

While the ISAG100 and ISAG200 SNP panels do provide a good base for parentage validation via SNP they are not without their limitations, and only using them can result in parentage errors (Strucken et al., 2014, 2015; McClure et al., 2015; Buchanan, 2016). The number of parentage SNP used by each laboratory, breed society, or national valuation center will depend on cost and their level of acceptable risk for a parentage error. As the cost of SNP genotyping decreases the value of having a near

TABLE 3 | Breed Composition reference population size and the correlation between the predicted and listed breed compositions from 713,814 animals.

Parent validation[a]	AA[b]	AU	BA	BB	CH	FR	HE	HO	JE	LM	PT	SA	SH	SI	Ave
(A) ALL ANIMALS, *N* = 713,814															
NA	0.9410	0.9493	0.9021	0.9036	0.9432	0.6220	0.9258	0.8799	0.9321	0.9385	0.8971	0.9497	0.9083	0.9272	0.9014
Neither validated	0.9044	0.9367	0.8752	0.7871	0.8869	0.4478	0.8907	0.7967	0.9219	0.8886	0.8607	0.9290	0.8775	0.8854	0.8492
One validates	0.9607	0.9520	0.9240	0.9435	0.9647	0.7306	0.9488	0.9135	0.9518	0.9617	0.9111	0.9606	0.9293	0.9492	0.9287
Both validate	0.9713	0.9656	0.9196	0.9385	0.9660	0.7265	0.9612	0.9408	0.7627	0.9642	0.9346	0.9706	0.9462	0.9583	0.9233
(B) ANIMAL'S LISTED COMPOSITION HAS SOME OF THE 14 BREEDS, *N* = 711,579															
NA	0.9417	0.9501	0.9035	0.9043	0.9432	0.6225	0.9261	0.8830	0.9344	0.9385	0.9096	0.9497	0.9113	0.9285	0.9033
Neither validated	0.9057	0.9381	0.8773	0.7894	0.8870	0.4483	0.8911	0.8032	0.9250	0.8885	0.8812	0.9291	0.8823	0.8878	0.8524
One validates	0.9610	0.9524	0.9250	0.9437	0.9648	0.7308	0.9490	0.9153	0.9529	0.9617	0.9185	0.9606	0.9308	0.9499	0.9297
Both validate	0.9717	0.9659	0.9205	0.9386	0.9660	0.7267	0.9614	0.9408	0.7720	0.9642	0.9411	0.9706	0.9474	0.9585	0.9247
(C) ANIMAL'S LISTED COMPOSITION IS >50% OF THE 14 BREEDS, *N* = 707,546															
NA	0.9429	0.9511	0.9049	0.9049	0.9433	0.6245	0.9267	0.8836	0.9363	0.9386	0.9167	0.9499	0.9140	0.9303	0.9049
Neither validated	0.9075	0.9398	0.8792	0.7907	0.8872	0.4504	0.8922	0.8045	0.9273	0.8885	0.8921	0.9294	0.8864	0.8913	0.8547
One validates	0.9615	0.9531	0.9258	0.9440	0.9648	0.7318	0.9492	0.9155	0.9537	0.9617	0.9236	0.9607	0.9320	0.9508	0.9306
Both validate	0.9726	0.9662	0.9213	0.9389	0.9660	0.7291	0.9617	0.9410	0.7791	0.9643	0.9450	0.9707	0.9494	0.9590	0.9260
(D) ANIMAL'S LISTED COMPOSITION IS ONLY COMPRISED FROM THE 14 BREEDS, *N* = 224,428															
NA	0.9815	0.9833	0.9620	0.9528	0.9778	0.8313	0.9821	0.9498	0.9720	0.9758	0.9599	0.9816	0.9649	0.9742	0.9606
Neither validated	0.9677	0.9839	0.9611	0.9085	0.9493	0.7576	0.9631	0.9315	0.9725	0.9524	0.9624	0.9769	0.9572	0.9532	0.9427
One validates	0.9874	0.9823	0.9674	0.9686	0.9863	0.8442	0.9892	0.9496	0.9747	0.9848	0.9522	0.9852	0.9709	0.9826	0.9661
Both validate	0.9869	0.9841	0.9535	0.9554	0.9810	0.8887	0.9883	0.9755	0.8733	0.9790	0.9656	0.9832	0.9686	0.9783	0.9615

Correlation on 710,000 animals between the predicted and listed breed composition for 14 breeds.

[a] *Parent SNP validation possibilities: NA, results for all animals regardless of parentage validation status; Neither validated, neither parent validated, either parent not SNP genotyped or SNP failed; One validates, one parent SNP validates other SNP failed or not SNP genotyped; Both validate, both parents SNP validate.*

[b] *Breed abbreviations defined in **Table S1**.*

perfect pedigree could soon outweigh the cost of genotyping an animal with additional SNP. To ensure a higher level of parentage accuracy we recommend that at least 500 SNP are used and more is better (McClure et al., 2015). Some might argue that if one restricts parentage to only herd level, less SNP could be used for prediction, but this does not consider potential errors from fence jumping breeding stock or mis-recorded semen straws. In Ireland, the ICBF800 parentage SNP set have proven very effective for highly accurate parentage validation and prediction, while not being too computationally demanding. Pedigree errors based upon the listed parents in Ireland is runs between 6 and 8%, is the average rate when all animals regardless of breed, age, or pedigree status are analyzed. To date when all QC steps are used, >1 sire or dam has been predicted only due to identical twin animals. While identical twins will have unique methylation patterns (Kaminsky et al., 2009) a method to use this for parentage validation in livestock has not been developed. The ICBF800 is also useful for other processes: such as identification of SNP duplicates, if an animal's multiple genotypes match, GRM-lite, and parentage prediction as mentioned above. As the ICBF800 were selected based on their performance in a *B. taurus* population it is unknown how well they would perform in a *Bos indicus* or *B. taurus* x *indicus* population.

Overall the ICBF800 panel is more accurate than the current international bovine SNP panels; ISAG100 and ISAG200 (**Table 4**). On its own the ICBF800 panel is not perfect, but by using a 2-step process to further analyse any parentage result with 8–12 SNP misconcordance allows for a highly accurate parentage process with minimal computing requirements. We also recommend not using the SNP we identified having clustering issues for parentage analysis (**Table S2**). If one wishes to design their own parentage SNP panel we also recommend analyzing the SNP's clustering panel once a large number of animals are genotyped.

In the near future, national animal registration could occur via SNP genotypes. While an individual's date of birth can't currently be determined via its genotype its sex, parents, and breed composition can. SNP genotypes when combined with a robust QC pipeline and a tagging system that collects a tissue sample at the same time a national ID ear tag is applied would allow for unparalleled animal traceability. In theory, the animal and its products could be 100% identified if enough genotypes are used which would have applications in animal forensics, theft, and product marketing using a slight modification of the SNP duplication portion of the pipeline.

The ICBF800 works extremely well for parentage and when combined with the rest of our QC pipeline has practically removed the possibility of accidently validating or failing a pedigree incorrectly. While the ICBF800 works across multiple *B. taurus* breeds, a different set of >500 SNP could eventually

TABLE 4 | Comparison results of parentage validation test for 300,020 animals between the ICBF800 and the smaller ISAG100 and ISAG200 SNP panels.

		SNP panel used	
	Count[a]	ISAG100[b] (%)	ISAG200 (%)
Sire	292462	0.249	0.022
Dam	27330	0.040	0.004
Total[c]	319792	0.231	0.021

*ISAG200 SNP with clustering issues in **Table S3** were excluded from the analysis.*

[a] *Count of SNP parentage checks analyzed, by sire, dam, or total.*

[b] *Percent of animals that had a different parentage verification result for the 100 SNP panel when compared to the ICBF800 panel at the 1% misconcordance level.*

[c] *19,774 animals had both sire and dam SNP checked.*

be determined to work better for specific or rare breeds. In fact, the current set of 800 SNP used by ICBF is no longer the best set of 800 SNP to use based on the MAF criteria we original used (**Table S2**). This is caused by changing nation-wide MAF as more animals from more breeds are genotyped. In 2014, 69% of the genotyped animals were Holstein; while by March 2016 Holsteins represented <14% of the genotyped animals. **Table 2**, shows the major breed component of animals genotyped at ICBF across multiple years. Currently 25 breeds have ≥10 genotyped purebred animals in the ICBF database, while 21 breeds have <10. Even with changing nation-wide MAF the non-ISAG200 ICBF800 SNP have an average MAF of 0.484 with a minimum MAF of 0.429 across 852,087 animals on April 5th, 2017. When only purebred animals are considered within breeds with >500 animals genotyped ($N = 15$ breeds), the minimum MAF for the non-ISAG200 ICBF800 SNP within breed ranges from 0.027 to 0.144 and the average MAF ranges from 0.375 to 0.408. For the ISAG200 SNP the minimum MAF within these purebreds ranges from 0.017 to 0.145 and the average MAF ranges from 0.310 to 0.402 (Summary in **Table 5**, by SNP in **Table S3**).

While the SNP part of the QC pipeline results in a "black or white" answer, the Animal QC portion has all parts run and the combined output is used to determine if a genotype should be invalidated. For instance, a young animal could pass all the QC checks except sex. This could simply be because the famer recorded its sex wrong by accident. The full Animal QC analysis also provides greater evidence that a genotype truly does not belong to the listed animal if an inquiry ever arises. While sex prediction can be done using only chrX or chrY SNP, we do recommend that both be used to increase sex prediction accuracy and to also identify Klinefelter's and Turner syndrome animals.

Only after samples have passed the full QC pipeline does ICBF use the data for parentage analysis, genetic disease/trait status, and SNP imputation. The IDBv3 has >200 diagnostic probes for genetic diseases and traits (http://www.icbf.com/?page_id=2170). FImpute (Sargolzaei et al., 2011) is used to impute all animals with a valid genotype to 50 k density for genomic breeding value estimation and to IDBv3 density for genetic disease/trait status. By using only valid genotypes and having a highly accurate pedigree via the ICBF800, ICBF can help farmers maximize their genetic gain while minimizing their genetic disease risk.

TABLE 5 | MAF summary of parentage SNP across 852,087 animals and for 25 breeds.

		ALL	PURE[a]	CROSS[b]	AA[c]	AU	BA	BB	CH	DX	FR	HE	HO	JE	KE	LH	LM	MH	MO	MY	NR	PI	PT	RM	SA	SH	SI	SP	ST
SNP Panel[d]	Count[e]	8,52,087	1,46,094	7,05,993	19,077	2,189	1,354	2,103	34,538	21	710	1,06,77	16,795	1,315	11	10	39,148	13	204	16	21	595	1,091	59	3,341	2387	9,900	63	26
ICBF800	Max	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
	Min	0.267	0.255	0.269	0.040	0.053	0.091	0.058	0.073	0.000	0.082	0.017	0.082	0.017	0.000	0.000	0.051	0.000	0.069	0.000	0.000	0.114	0.132	0.000	0.027	0.050	0.035	0.024	0.038
	Ave	0.470	0.459	0.471	0.349	0.380	0.390	0.368	0.404	0.334	0.387	0.337	0.398	0.327	0.308	0.215	0.393	0.311	0.366	0.359	0.348	0.388	0.388	0.312	0.349	0.340	0.366	0.328	0.377
Non ISAG	Max	0.500	0.500	0.500	0.500	0.500	0.500	0.499	0.500	0.500	0.500	0.500	0.500	0.499	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.499	0.500	0.500
	Min	0.428	0.345	0.435	0.040	0.109	0.091	0.105	0.144	0.000	0.083	0.028	0.124	0.034	0.000	0.000	0.051	0.000	0.069	0.000	0.000	0.130	0.132	0.000	0.027	0.072	0.094	0.024	0.058
	Ave	0.484	0.470	0.485	0.354	0.385	0.394	0.371	0.408	0.337	0.392	0.338	0.396	0.332	0.312	0.219	0.401	0.312	0.364	0.364	0.349	0.392	0.390	0.312	0.352	0.342	0.373	0.333	0.377
ISAG	Max	0.499	0.500	0.498	0.499	0.499	0.500	0.500	0.498	0.500	0.499	0.497	0.498	0.500	0.500	0.500	0.499	0.500	0.500	0.500	0.500	0.499	0.499	0.500	0.500	0.499	0.500	0.500	0.500
	Min	0.267	0.255	0.269	0.042	0.053	0.109	0.058	0.073	0.000	0.082	0.017	0.082	0.017	0.000	0.000	0.052	0.000	0.071	0.000	0.000	0.114	0.145	0.017	0.038	0.050	0.035	0.024	0.038
	Ave	0.427	0.422	0.428	0.334	0.363	0.374	0.356	0.389	0.327	0.374	0.334	0.402	0.310	0.296	0.201	0.370	0.308	0.371	0.340	0.345	0.376	0.381	0.313	0.339	0.332	0.346	0.315	0.376

[a] *Value when only purebred animals are analyzed, including breeds with<10 purebred animals genotyped.*

[b] *Values when only crossbred animals are analyzed.*

[c] *Breed abbreviations defined in **Table S1**.*

[d] *ISAG includes all ISAG200 SNP except those excluded for QC issues in **Table S2**.*

[e] *Number of genotyped animals.*

In addition to using ≥500 SNP for parentage validation, we also strongly suggest each laboratory, breed society, or national valuation center put in place a QC pipeline. The QC pipeline we describe above can be used as an initial foundation for one to build a custom QC pipeline one. One advice we have for any QC pipeline is to include logic checks for any situation one can think of regardless of how "rare" you may think it could occur. It is far easier to deal with a problem genotype early than down the road after pedigrees have been validated, breeding values estimated, and breeding decisions made. The SNP and Animal QC process developed at ICBF has been extremely useful to improve and guarantee the accuracy of the data and any report based on it, from parentage to genomic breeding values. We hope that other can use our QC process to help improve their own systems as they increase their amount of genetic and pedigree data.

AUTHOR CONTRIBUTIONS

MM and JK: developed the improved SNP parentage process; MM, JM, ED, JCM, DO, and JK: all identified aspects of the quality control pipeline; MM and JM: developed the quality control pipeline; MM and PF: identified SNP clustering issues; MM: wrote the manuscript with input from all authors.

FUNDING

This project was funded solely using internal funds from ICBF.

ACKNOWLEDGMENTS

We wish to thank all of the Irish farmers that have had their animals genotyped and provided data to ICBF, without them this work would not have been possible. SNP parentage selection work was originally published as SNP Selection for Nationwide Parentage Verification and Identification in Beef and Dairy Cattle in the ICAR Technical Series 19 on page 185, accessible at: http://www.icar.org/wp-content/uploads/2015/11/ICAR-Technical-Series-19-Krakow-2015-Proceedings.pdf.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00084/full#supplementary-material

Figure S1 | SNP cluster issues for 16 SNP based on Illumina GenomeStudio 2.0 SNP cluster plots of 4,561 animals from multiple breeds. SNP ID **(Left** to **Right** and **Top** to **Bottom)** are: ARS-BFGL-NGS-3547, ARS-BFGL-NGS-11469, ARS-BFGL-NGS-43361, ARS-BFGL-NGS-53975, ARS-BFGL-NGS-62906, ARS-BFGL-NGS-66558, ARS-BFGL-NGS-76191, ARS-BFGL-NGS-103099, ARS-BFGL-NGS-112652, ARS-BFGL-NGS-118188, ARS-USMARC-Parent-DQ837645-rs29015870, BTA-100621-no-rs, BTB-00147175, BTB-01834338, Hapmap47324-BTA-55159, UA-IFASA-9571.

Table S1 | List of *Bos Taurus* breeds present (purebred or crossbred) in Ireland and if they are represented in the ICBF genotype database from a genotyped purebred or crossbred individual.

Table S2 | List of SNP from the Ilumina LD base content with ICBF800, ISAG100, and ISAG200 SNP identified.

Table S3 | Parentage SNP MAF across 852,087 animals and for 25 breeds.

Table S4 | SNP. 923 SNP used by ICBF for microsatellite impuation.

Table S5 | Chromosome X PAR and nPAR location based on IDBv3 SNP data in 38,564 males and 568,841 females.

Table S6 | Full parentage results for animals with 0.5–1.5% ICBF800 parentage SNP misconcordances.

Table S7 | Inheritance analysis of patterners of 625,384 chromosome Y SNP containing females.

REFERENCES

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664. doi: 10.1101/gr.094052.109

Banos, G., Wiggans, G. R., and Powell, R. L. (2001). Impact of paternity errors in cow identification on genetic evaluations and international comparisons. *J. Dairy Sci.* 84, 2523–2529. doi: 10.3168/jds.S0022-0302(01)74703-0

Berry, D. P., Wolfe, A., O'donovan, J., Byrne, N., Sayers, R. G., Dodds, K. G., et al. (2017). Characterization of an X-chromosomal non-mosaic monosomy (59, X0) dairy heifer detected using routinely available single nucleotide polymorphism genotype data. *J. Anim. Sci.* 95, 1042–1049. doi: 10.2527/jas.2016.1279

Buchanan, J. W. (2016). "Evaluation of SNP-based parentage panels in commercial beef cattle, populations," in *Plant and Animal Genome XXIV Conference, 2016* (San Diego, CA).

Cooper, T. A., Wiggans, G. R., and Vanraden, P. M. (2013). Short communication: relationship of call rate and accuracy of single nucleotide polymorphism genotypes in dairy cattle. *J. Dairy Sci.* 96, 3336–3339. doi: 10.3168/jds.2012-6208

Davis, G. P., and Denise, S. K. (1998). The impact of genetic markers on selection. *J. Anim. Sci.* 76, 2331–2339. doi: 10.2527/1998.7692331x

Feigelson, H. S., Rodriguez, C., Welch, R., Hutchinson, A., Shao, W., Jacobs, K., et al. (2007). Successful genome-wide scan in paired blood and buccal samples. *Cancer Epidemiol. Biomark. Prev.* 16, 1023–1025. doi: 10.1158/1055-9965.EPI-06-0859

Gasser, C. L., Grum, D. E., Mussard, M. L., Fluharty, F. L., Kinder, J. E., and Day, M. L. (2006). Induction of precocious puberty in heifers I: enhanced secretion of luteinizing hormone. *J. Anim. Sci.* 84, 2035–2041. doi: 10.2527/jas.2005-636

Harder, B., Bennewitz, J., Reinsch, N., Mayer, M., and Kalm, E. (2005). Effect of missing sire information on genetic evaluation. *Arch. Tierz* 48, 219–232. doi: 10.5194/aab-48-219-2005

Heaton, M. P., Harhay, G. P., Bennett, G. L., Stone, R. T., Grosse, W. M., Casas, E., et al. (2002). Selection and use of SNP markers for animal identification and paternity analysis in U.S. beef cattle. *Mamm. Genome* 13, 272–281. doi: 10.1007/s00335-001-2146-3

Henderson, D., Thomas, M., and Da, Y. (2005). Bovine genomics from academia to industry. *Comp. Funct. Genomics* 6, 174–180. doi: 10.1002/cfg.467

Illumina Inc. (2009). *Genome-Wide Dna Analysis Beadchips*. Available online at: https://www.illumina.com/content/dam/illumina-marketing/documents/products/brochures/datasheet_omni_whole-genome_arrays.pdf (Accessed March 25, 2009).

Illumina Inc. (2010). *Bovinehd Genotyping Beadchip*. Available online at: https://www.illumina.com/Documents/products/datasheets/datasheet_bovineHD.pdf (Accessed Jun 12, 2011).

Illumina Inc. (2011a). *Bovineld Genotyping Beadchip*. Available online at: https://www.illumina.com/documents/products/datasheets/datasheet_bovineLD.pdf (Accessed August 31, 2012).

Illumina Inc. (2011b). *Goldengate Bovine3k Genotyping Beadchip*. Available online at: https://www.illumina.com/Documents/products/datasheets/datasheet_bovine3k.pdf (Accessed May 16, 2012).

Israel, C., and Weller, J. I. (2000). Effect of misidentification on genetic gain and estimation of breeding value in dairy cattle populations. *J. Dairy Sci.* 83, 181–187. doi: 10.3168/jds.S0022-0302(00)74869-7

Kaminsky, Z. A., Tang, T., Wang, S. C., Ptak, C., Oh, G. H., Wong, A. H., et al. (2009). DNA methylation profiles in monozygotic and dizygotic twins. *Nat. Genet.* 41, 240–245. doi: 10.1038/ng.286

Leroy, G., Danchin-Burge, C., Palhiere, I., Baumung, R., Fritz, S., Meriaux, J. C., et al. (2012). An ABC estimate of pedigree error rate: application in dog, sheep and cattle breeds. *Anim. Genet.* 43, 309–314. doi: 10.1111/j.1365-2052.2011.02253.x

Matukumalli, L. K., Lawley, C. T., Schnabel, R. D., Taylor, J. F., Allan, M. F., Heaton, M. P., et al. (2009). Development and characterization of a high density SNP genotyping assay for cattle. *PLoS ONE* 4:e5350. doi: 10.1371/journal.pone.0005350

McClure, M. C., Mccarthy, J., Flynn, J., Weld, R., Keane, M., O'connel, K., et al. (2015). "SNP selection for nationwide parentage verification and identification in beef and dairy cattle," in *Proceedings, International Committee For Animal Recording Technical Series, June 2015,*eds Z. Kowalski, N. Petreny, M. Burke, P. Bucek, L. Journaux, M. Coffey, C. Hunlun, and D. Radzio (Krakow; Rome: ICAR), 175–181.

McClure, M. C., McClure, J. C., and Mccarthy, J. (2017). Rate of bovine heteropaternal superfecundation in the irish national herd: twins with different sires. *Anim. Genet.* 48, 721–723. doi: 10.1111/age.12619

Mcclure, M. C., Mckay, S. D., Schnabel, R. D., and Taylor, J. F. (2009). Assessment Of DNA extracted from FTA(R) cards for use on the illumina iselect beadchip. *BMC Res. Notes* 2:107. doi: 10.1186/1756-0500-2-107

McClure, M. C., Mullen, M. P., Kearney, J. F., Cromie, A. R., Treacy, M., Flynn, P., et al. (2014). *Application of a Custom SNP Chip: Microsatellite Imputation, Parentage SNP Imputation, Genomic Evaluations, and Across-Breed Nation-Wide Genetic Disease Prevalence with the International Beef and Dairy SNP Chip*. Berlin: ICAR.

McClure, M. C., Sonstegard, T. S., Wiggans, G. R., Van Eenennaam, A. L., Weber, K. L., Penedo, M. C. T., et al. (2013). Imputation of microsatellite alleles from dense SNP genotypes for parentage verification across multiple *Bos taurus* and *Bos indicus* breeds. *Front. Genet.* 4:176. doi: 10.3389/fgene.2013.00176

Montgomery, G. W., Campbell, M. J., Dickson, P., Herbert, S., Siemering, K., Ewen-White, K. R., et al. (2005). Estimation of the rate of SNP genotyping errors from DNA extracted from different tissues. *Twin Res. Hum. Genet.* 8, 346–352. doi: 10.1375/twin.8.4.346

Mullen, M. P., Mcclure, M. C., Kearney, J. F., Waters, S. M., Weld, R., Flynn, P., et al. (2013). *Development of a Custom SNP Chip for Dairy and Beef Cattle Breeding, Parentage, and Research*. INTERBULL, August 23–25, 2013 Nantes. 47.

Neogen Corporation (2012). *Geneseek Genomic Profiler For Dairy Cattle.* Available online at: http://www.neogen.com/Geneseek/pdf/Catalogs/DairyGenomicProfiler.pdf (Accessed August 31, 2012).

Neogen Corporation (2013). *Geneseek Genomic Profiler Bovine HD*. Available online at: http://www.neogen.com/Corporate/PR2013/2013-02-07.pdf (Accessed March 1, 2013).

Obsteter, J. (2017). *Parentage Verification Using Imputed Microsatellite and SNP Data In Slovenian Brown Swiss Population*. Puerto Varas: Interbull Bulletin.

Rawlings, N., Evans, A. C., Chandolia, R. K., and Bagu, E. T. (2008). Sexual maturation in the bull. *Reprod. Domest. Anim.* 43(Suppl. 2), 295–301. doi: 10.1111/j.1439-0531.2008.01177.x

Sanders, K., Bennewitz, J., and Kalm, E. (2006). Wrong and missing sire information affects genetic gain in the angeln dairy cattle population. *J. Dairy Sci.* 89, 315–321. doi: 10.3168/jds.S0022-0302(06)72096-3

Sargolzaei, M., Chesnais, J. P., and Schenkel, F. S. (2011). Fimpute-an efficient imputation algorithm for dairy cattle populations. *J. Dairy. Sci.* 94:421.

Stormont, C. (1967). Contribution of blood typing to dairy science progress. *J. Dairy Sci.* 50, 253–260. doi: 10.3168/jds.S0022-0302(67)87401-0

Strucken, E. M., Gudex, B., Ferdosi, M. H., Lee, H. K., Song, K. D., Gibson, J. P., et al. (2014). Performance of different SNP panels for parentage testing in two East Asian cattle breeds. *Anim. Genet.* 45, 572–575. doi: 10.1111/age.12154

Strucken, E. M., Lee, S. H., Lee, H. K., Song, K. D., Gibson, J. P., and Gondro, C. (2015). How many markers are enough? Factors influencing parentage testing in different livestock populations. *J. Anim. Breed. Genet.* 133:13–23. doi: 10.1111/jbg.12179

Turner, S., Armstrong, L. L., Bradford, Y., Carlson, C. S., Crawford, D. C., Crenshaw, A. T., et al. (2011). Quality control procedures for genome-wide association studies. *Curr. Protoc. Hum. Genet.* Chapter 1, Unit1.19. doi: 10.1002/0471142905.hg0119s68

Vandeputte, M. (2012). An accurate formula to calculate exclusion power of marker sets in parentage assignment. *Genet. Sel. Evol.* 44:36. doi: 10.1186/1297-9686-44-36

Visscher, P. M., Woolliams, J. A., Smith, D., and Williams, J. L. (2002). Estimation of pedigree errors in the UK dairy population using microsatellite markers and the impact on selection. *J. Dairy Sci.* 85, 2368–2375. doi: 10.3168/jds.S0022-0302(02)74317-8

Wehrman, M. E., Kojima, F. N., Sanchez, T., Mariscal, D. V., and Kinder, J. E. (1996). Incidence of precocious puberty in developing beef heifers. *J. Anim. Sci.* 74, 2462–2467. doi: 10.2527/1996.74102462x

Wiggans, G. R., Cooper, T. A., Vanraden, P. M., Olson, K. M., and Tooker, M. E. (2012). Use of the illumina Bovine3k beadchip in dairy genomic evaluation. *J. Dairy Sci.* 95, 1552–1558. doi: 10.3168/jds.2011-4985

Wiggans, G. R., Vanraden, P. M., and Cooper, T. A. (2011). The genomic evaluation system in the united states: past, present, future. *J. Dairy Sci.* 94, 3202–3211. doi: 10.3168/jds.2010-3866

Woo, J. G., Sun, G., Haverbusch, M., Indugula, S., Martin, L. J., Broderick, J. P., et al. (2007). Quality assessment of buccal versus blood genomic DNA using the Affymetrix 500 K GeneChip. *BMC Genet.* 8:79. doi: 10.1186/1471-2156-8-79

Zimin, A. V., Delcher, A. L., Florea, L., Kelley, D. R., Schatz, M. C., Puiu, D., et al. (2009). A whole-genome assembly of the domestic cow, *Bos taurus*. *Genome Biol.* 10:R42. doi: 10.1186/gb-2009-10-4-r42

Conflict of Interest Statement: PF was employed by company Weatherbys Scientific.

The other authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.