Development of an automated quality control pipeline to facilitate the reporting of major gene genotypes

K. Quigley[1], T. Browne[1], K. O' Connell[1], P. Flynn[2], R.D. Evans[1] and M.P. Mullen[3]

[1]Irish Cattle Breeding Fedaration, Link Rd, Ballincollig, Co. Cork. P31 D452, Ireland
[2]Weatherbys Scientific, Unit F1 M7 Business Park Newhall, Naas Co. Kildare W91 VX86, Ireland
[3]AgriGenomics Ltd., Cloonmweelaun, Menlough, Ballinasloe, Co. Galway, Ireland

Abstract

The Irish Cattle Breeding Federation (ICBF) national cattle database stores in excess of 3.1 million genotypes, from both dairy and beef herds consisting of both purebred and crossbred animals. The reporting of genomic mutations with large effect termed major genes, is of benefit to breeders and industry, providing valuable information on both desirable and undesirable major genes segregating within herds. Pre genomics, the major gene status of an animal, for example for some genetic diseases, was only discovered following the birth of an affected calf. Genotyping allows the identification and management of animals and their major gene status before the birth of any progeny. Recently, the ICBF have developed an automated pipeline to facilitate the largescale reporting of a panel of major gene genotypes. The pipeline consists of a series of additional quality control steps to increase reliability in the final genotype call for automated reporting. Current quality control steps in the pipeline include a manifest call rate of >=97 %, custom confidence scores (variant and genotype specific), SNP classification categories (plate specific), custom clustering separation (variant and genotype specific), minimum X and Y signal intensities, heterozygosity threshold check, and sire/dam/trio (where available) Mendelian checks. The development of this major gene pipeline will provide additional information to industry to aid breeding decisions and an opportunity to develop mating strategies where useful. .

Introduction

The Irish Cattle Breeding Federation (ICBF) national cattle database formed in 1998, stores in excess of 3.1 million genotypes, from both dairy and beef herds consisting of both purebred and crossbred animals. Services such as genetic/genomic evaluations, parentage verification, gender verification, breed composition and more recently major gene status, are provided to farmers and can be used to support breeding decisions on farm.

Such services are enabled by using the ICBF custom genotyping platform, termed The International Dairy and Beef SNP Chip (Mullen *et al.,* 2013), of which there are five iterations to date. The latest iteration of this platform (IDBv5), consists of 51,421 SNPs including the International Society of Animals Genetics (ISAG) recommended 200 SNPs for parentage verification, the International Committee for Animal Recording (ICAR) 554 SNPs for parentage discovery, an updated ICBF 800 parentage SNP panel (McClure *at al.,* 2018), a large number of genome wide polymorphisms for genomic evaluation and research purposes, and approximately 190 major genes which can be further categorised based on their effect; Beneficial, Meat, Milk, Colour, Unwanted and Lethal. These major genes underly a broad range of genetic conditions observed in many cattle breeds such as conditions that are favourable to breeders, conditions underlying pigmentation, conditions affecting milk and meat productivity, conditions

that affect the animal before it can make an economic return, and conditions that result in embryonic lethality (McClure & McClure, 2016). Leveraging genomic data will help to monitor and manage genetic conditions known to be segregating within Irish cattle populations, including the identification of carrier animals.

The major gene status of animals is of interest to cattle breeders due to both the desirable and undesirable effects of genetic conditions on production and performance, ultimately impacting the profitability of farming enterprises (Cole *et al.,* 2016). Prior to the genomics era, a major gene carrier was only identified after an affected calf was produced, due to the phenotypic resemblance between both carrier and normal animals for many genetic conditions (Cieploch *et al.,* 2017). Since the advent of genotyping, major gene status can be determined before the animal reaches sexual maturity, allowing breeders to make more informed breeding decisions, for example, to develop mating strategies where carriers have been identified and reduce the risk of producing affected offspring or conversely to increase the frequency of a desirable major gene in their herd (McClure *et al.,* 2013).

The current process in place for the reporting of major genes in Ireland is handled by a commercial service provider (Weatherby's Ireland), where any breeder, artificial insemination (AI) company or herdbook can make a request. For the service provider, this is a manual process whereby the genotype is analysed on output from the genotyping process. There is also a cost for the breeder, irrespective of the royalty status of the major gene requested. With this in mind, the ICBF aimed to develop an automated quality control pipeline which would facilitate the largescale routine reporting of major gene genotypes. The metrics applied as part of the pipeline are described herein..

Methods

Major Gene Reporting

The International Dairy and Beef (IDB) is the ICBF custom genotyping platform, of which there are five iterations to date. The current iteration termed IDBv5, is a ThermoFisher Applied Biosystems™ Axiom™ Genotyping array. The major gene pipeline consists of a series of quality control metrics associated with each SNP, genotype and genotyping plate (n=384 samples). The pipeline is initially focused on genotypes derived from the IDBv5 platform. The aim of the pipeline is to improve the reliability of the final genotype call for automated reporting. At the time of submission, there were a total of 3,124,175 genotypes in the database, of which 1,114,739 are IDBv5 genotypes (CR>=0.97) and are eligible for the major gene pipeline.

Current quality control metrics included in the major gene pipeline are described in Table 1. Generic quality control thresholds applied to genotypes include; an animal call rate of >=0.97 and a Mendelian check (Parents and Trio). Three additional metric thresholds which are outputs of the Thermofisher genotyping process are also applied to each SNP and genotype; Clustering separation X contrast values, Confidence Score and minimum X and Y signal intensity values. Thresholds applied are specific to each SNP and genotype. One quality control metric (SNP classification) is applied to each genotyping plate.

Results and discussion

Genotypes from the major gene pipeline, released since November 2022, include major genes of immediate interest to breeders and industry, namely Myostatin and Polled status. With regards to Myostatin, there are nine variants routinely reported including *L64P*, (Dierks *et al.*, 2014) *F94L*, *nt419*, *Q204X*, *E226X*, *C313Y* (Grobet *et al.*, 1998), *nt821del11* (Grobet *et al.*, 1997) *S105C* and *D182N* (Dunner *et al.*, 2003) (Table 2).

Table 1 Description of QC metrics included in the major gene pipeline.*

Metric	Specific to	Description
Animal Call Rate (ACR)	Genotype	ACR threshold of >=97%. The call rate is defined as the proportion of SNPs with a genotype call for each individual i.e. the number of called SNPs/ the total number of SNPs.
Mendelian Check	Genotype	To detect scenarios where the genotype of the individual is not consistent with the transmission pattern expected according to Mendel's law of inheritance using comparisons to both parents individually and as a trio where available.
Cluster Separation	SNP and Genotype	Thresholds applied to contrast values based on the clustering resolution of each genotype class. Clusters should be well separated and distinct from each other, be well-formed and have no visible cluster abnormalities.
Confidence Score	SNP and Genotype	The confidence score is described as 1 minus the posterior probability of the genotype belonging to the assigned genotype cluster. It can range between zero and one, with lower confidence scores indicating more confident genotype calls.
SNP Classification	Plate	Each genotyping plate is classified into one of the six SNP classification categories – PolyHighResolution, MonoHighResolution, NoMinorHom, CallRateBelowThreshold, Off Target Variant and Other. This metric analyses the performance of the AA, AB and BB clusters, and their relationship to each other. **PolyHighResolution** - SNPs with well separated, distinct genotyping clusters and >2 occurrences of the minor allele. **MonoHighResolution** - SNPs with one distinct and well-formed genotyping cluster - all genotyped samples are monomorphic/ homozygous. **NoMinorHom** - SNPs with well separated, distinct genotyping clusters with no minor homozygous genotypes i.e. One cluster is homozygous and one is heterozygous (for biallelic SNPs). **OffTargetVariant (OTV)** - SNP sites whose sequences are significantly different from the sequences of the hybridisation probes. **CallRateBelowThreshold** - SNP call rate is below the threshold, but other QC metrics are acceptable. **Other** – At least one QC metric is not meeting the required threshold.
Signal Intensity	SNP and genotype	Minimum thresholds applied to X and Y intensity values to identify and exclude low intensity genotypes.

*More details can be found in Thermo Fisher Scientific Inc (2020).

Same variants are positioned in exonic regions of *MSTN*, located on BTA2 (Positions based on assembly ARS UCD 1.2 of the *Bos taurus* genome build). Additionally of interest is the Polled Celtic variant, a complex rearrangement positioned between IFNAR2 and OLIG1 (Aldersey *et al.*, 2020; Allais-Bonnet *et al.*, 2013; Medugorac *et al.*, 2012) (Table 3).

Myostatin has been the subject of interest to cattle breeders for some time, primarily due to its effect on carcass performance but also due to its negative impact on calving difficulty (Purfield *et al.*, 2019; Bellinge *et al.*, 2005; Casa *et al.*, 1999). Additionally, Polledness in cattle resulting in the absence of horns, is a favourable trait for many breeders, alleviating the cost associated with dehorning and averting associated injuries, safety and welfare concerns (Aldersey *et al.*, 2020; Allais-Bonnet *et al.*, 2013; Medugorac *et al.*, 2012).

Table 2. RS IDs (where available), coordinates and OMIA references for Myostatin variants routinely released since November 2022 as part of the ICBF major gene pipeline (Positions based on assembly ARS UCD 1.2 of the Bos taurus genome build).

Variant	Rs ID	Coordinates	Amino Acid Change	OMIA
L64P	rs449270213	2:6279187	p.Leu64Pro	000683-9913
F94L	rs110065568	2:6279278	p.Phe94Leu	000683-9913
S105C		2:6279310	p.Ser105Cys	000683-9913
nt419		2:6281243	-	-
D182N		2:6281368	p.Asn182Asp	000683-9913
Q204X	rs110344317	2:6281434	p.Gln204X	000683-9913
E226X		2:6281500	p.Glu226X	000683-9913
nt821del11	rs382669990	2:6283674	p.Glu275ArgfsX14	000683-9913
C313Y		2:6283794	p.Cyc313Try	000683-9913

Table 3. RS IDs (where available), coordinates and OMIA references for the Polled Celtic variant routinely released as part of the ICBF major gene pipeline - (Positions based on assembly ARS UCD 1.2 of the Bos taurus genome build).

Variant	Rs ID	Coordinates	Amino Acid Change	OMIA
Polled Celtic	-	1:2429327_2429336del-2429109_2429320dupins	-	000483-9913

Since the implementation of the pipeline in November 2022, 1,107,481 genotypes have been released for the nine Myostatin polymorphisms and the Polled Celtic variant. Of this number, the sample pass rate ranges from 91.5 % for S105C to 99.1 % for C313Y, with an average sample pass rate of 95.6 % (Table 4)

Communication of results to farmers, herd-books and AI companies is through herd profiles on the ICBF website, where herd owners can access major gene reports for each animal that meets the criteria (Figure 1). Additionally, results may be reported on EU Zootechnical certificates where results are stated in a section dedicated to genetic defects and genetic peculiarities.

This pipeline provides valuable information on the major gene status of animals which herdowners may incorporate to aid breeding decisions. Moreover, the reporting of results to cattle herd-books ensures the monitoring and management of major genes

Table 4. Animal Pass rates for the 10 major genes routinely released as part of the ICBF major gene pipeline since November 2022.

Locus	MG variant	Total passed*	Pass rate (%)
MSTN	L64P	1,059,810	95.7
MSTN	F94L	1,038,833	93.8
MSTN	nt419	1,037,607	93.7
MSTN	S105C	1,012,891	91.5
MSTN	D182N	1,083,609	97.8
MSTN	Q204X	1,091,075	98.5
MSTN	E226X	1,044,293	94.3
MSTN	nt821del11	1,091,882	98.6
MSTN	C313Y	1,097,347	99.1
POLLED	Polled Celtic	1,046,889	94.5

*Total passed samples of total samples through the pipeline (n=1,107,481).

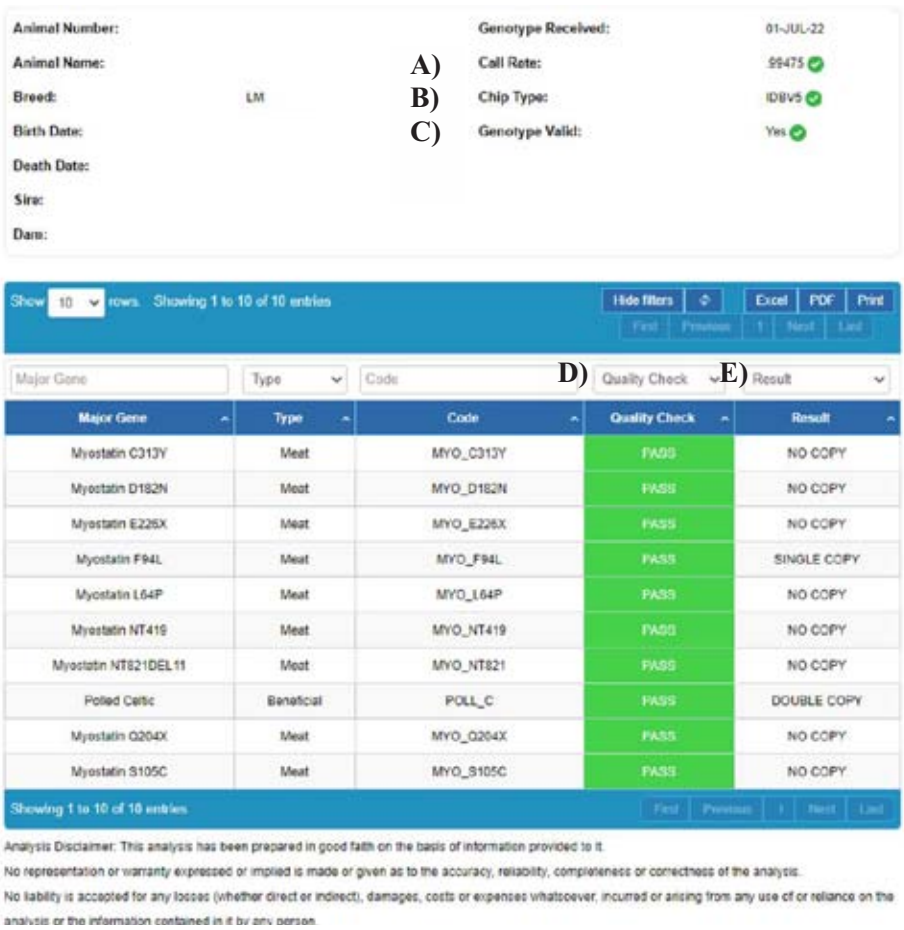

Figure 1 An example of a major gene profile which herd owners can access through their herd profile for eligible animals. A) The Call Rate threshold is 97%, i.e. every animal entering the major gene pipeline must have a call rate of >=97%. If the call rate is <97%, the animal will not have results available for any major gene. B) The Chip Type is the platform the animal was genotyped on. Major Genes are currently being reported for animals genotyped on IDBv5. C) Each individual SNP, genotype and plate go through a series of quality control metrics that improve the confidence of each genotype reported. A genotype passing all QC metrics is deemed Genotype Valid. D) Quality Check: Indicates whether the genotype has passed background checks and is valid for the major gene pipeline. E) The Results are reported as no copy, single copy and double copy.

segregating within cattle breeds. Ongoing and future work includes expanding the major genes incorporated into the pipeline for routine release, including Polled Friesian, lethals and colour related major genes.

References

Aldersey JE, Sonstegard TS, Williams JL, Bottema CDK. (2020). Understanding the effects of the bovine POLLED variants. Animal Genetics. 51(2):166-176.

Allais-Bonnet A, Grohs C, Medugorac I, Krebs S, Djari A, *et al.* (2013). Novel Insights into the Bovine Polled Phenotype and Horn Ontogenesis in Bovidae. PLOS ONE. 8(5): e63512.

Bellinge, R.H.S., Liberles, D.A., Iaschi, S.P.A., O'Brien, P.A. and Tay, G.K. (2005). Myostatin and its implications on animal breeding: a review. Animal Genetics. 36(1).

Casas E, Keele JW, Fahrenkrug SC, Smith TPL, Cundiff LV, Stone RT. (1999). Quantitative analysis of birth, weaning, and yearling weights and calving difficulty in Piedmontese crossbreds segregating an inactive myostatin allele. Journal of Animal Science.77:1686–92.

Ciepłoch, A., Rutkowska, K., Oprządek, J., Poławska, E. (2017). Genetic disorders in beef cattle: a review. Genes & Genomics. 39:461-471.

Cole JB, Null DJ, VanRaden PM. (2016). Phenotypic and genetic effects of recessive haplotypes on yield, longevity, and fertility Journal of Dairy Science. 99(9):7274-7288.

Dierks, C., Eder, J., Glatzer, S., Lehner, S. and Distl, O. (2015). A novel myostatin mutation in double-muscled German Gelbvieh. Animal genetics [online], 46(1), pp.91–2.

Dunner S, Miranda ME, Amigues Y, Cañón J, Georges M, Hanset R, Williams J, Ménissier F. (2003). Haplotype diversity of the myostatin gene among beef cattle breeds. Genetics Selection Evolution. Jan-Feb;35(1):103-18.

Grobet, L., Martin, L.J.R., Poncelet, D., Pirottin, D., Brouwers, B., Riquet, J., Schoeberlein, A., Dunner, S., nissier, F., Massabanda, J., Fries, R., Hanset, R. and Georges, M. (1997). A deletion in the bovine myostatin gene causes the double muscled phenotype in cattle. Nature Genetics. 17(1), pp.71–74.

Grobet, L., Poncelet, D., Royo, L.J., Brouwers, B., Pirottin, D., Michaux, C., Ménissier, F., Zanotti, M., Dunner, S. and Georges, M. (1998). Molecular definition of an allelic series of mutations disrupting the myostatin function and causing double muscling in cattle. Mammalian Genome. 9(3), pp.210–213.

McClure, M., Mullen, M., Waters, S., Kearney, F., McClure, J., and Flynn, P., Weld, R. (2016). P6001 Effectively managing bovine genetic disease risk via genotyping the Irish national herd. Journal of Animal Science. 94. 148.

McClure, MC., McCarthy, J., Flynn, P., McClure, JC., Dair, E., O'Connell, DK., Kearney, JF. (2018). SNP Data Quality Control in a National Beef and Dairy Cattle System and Highly Accurate SNP Based Parentage Verification and Identification. Frontiers in Genetics. 15;9:84.

McClure, M., McClure, J. (2016). Genetic Disease and Trait Information for IDB Genotyped Animals in Ireland. Irish Cattle Breeding Federation (ICBF).

Medugorac I, Seichter D, Graf A, Russ I, Blum H, Göpel KH, Rothammer S, Förster M, Krebs S. (2012). Bovine polledness--an autosomal dominant trait with allelic heterogeneity. PLoS One. 7(6)

Mullen, M.P., McClure, M.C., Kearney, J.F., Waters, S.M., Weld, R., Flynn, P., Creevey, C.J., Cromie, A.R. and Berry, D.P. (2013). Development of a custom SNP chip for dairy and beef cattle breeding, parentage and research. Interbull Bulletin No. 47.

Purfield, D.C., Evans, R.D. & Berry, D.P. (2019). Reaffirmation of known major genes and the identification of novel candidate genes associated with carcass-related metrics based on whole genome sequence within a large multi-breed cattle population. BMC Genomics 20, 720.

Thermo Fisher Scientific Inc (2020). Axiom™ Genotyping Solution Data Analysis Guide [online] Revision C.0 (18 April 2023)