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

Katie Quigley, Thomas Browne, Karl O' Connell, Paul Flynn, Ross D Evans, Michael P Mullen

ICAR 2023

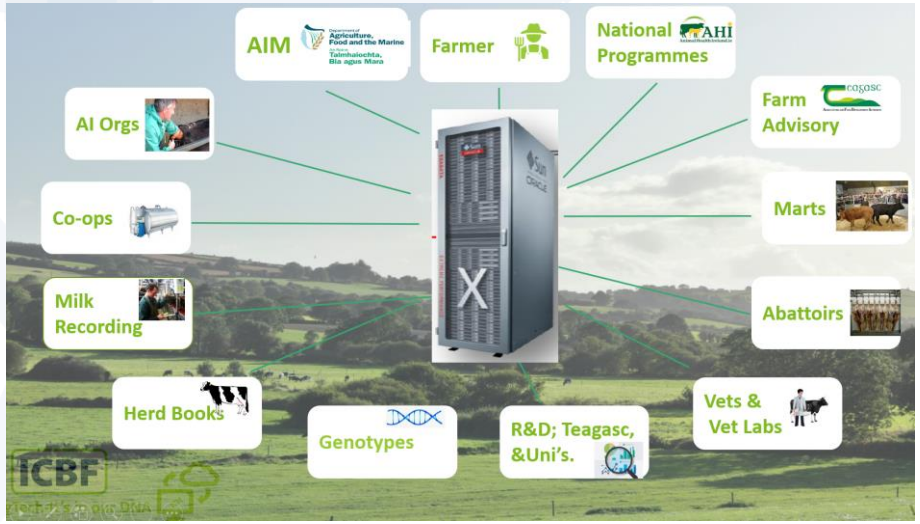
Date: 24th May 2023

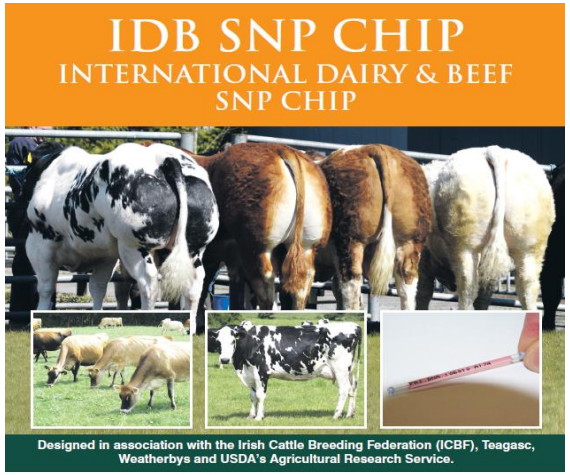
An Roinn Talmhaíochta, Bia agus Mara
Department of Agriculture, Food and the Marine

AgTech - it's in our DNA

Background

- Irish Cattle Breeding Federation (ICBF) - National cattle database
- >3 million genotypes
  - >2.6M with beef sire
  - >400K with dairy sire
- The International Dairy and Beef SNP Chip (IDB SNP Chip)
  - Parentage verification
  - Genomic evaluations for dairy and beef
  - Research
  - Calling of major genes





International Dairy and Beef SNP Chip (IDB)

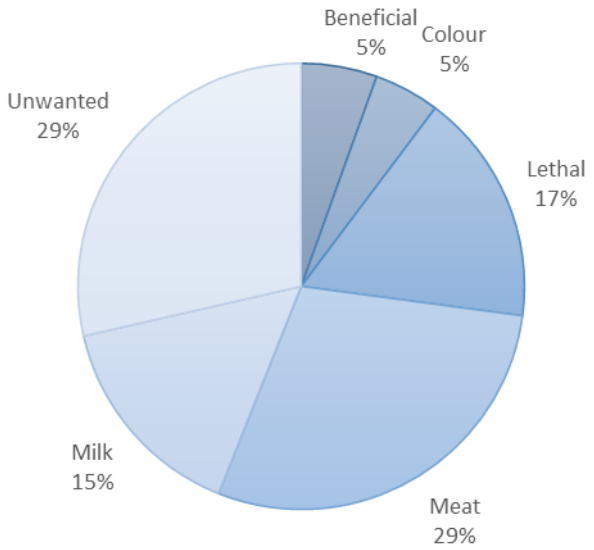
- Five iterations to date
- IDBv5 chip produced on Affymetrix/Thermofisher technology
  - 384 samples processed at a time = plate

IDB version	Provider	SNPs	Genotypes
IDBv1	Illumina Inc.	16k	30k
IDBv2	Illumina Inc.	16k	150k
IDBv3	Illumina Inc.	53k	1.2m
IDBv4	ThermoFisher	52k	400k
IDBv5	ThermoFisher	52k	1.1m

Plate

Gene Titan

- ~165 major genes (MG) on IDBv5
  - Dairy and beef breeds
  - Classed into major gene categories



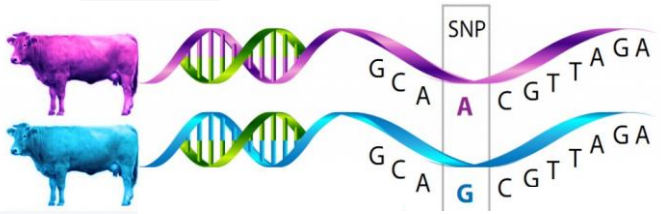
Current Process for Reporting Major Genes

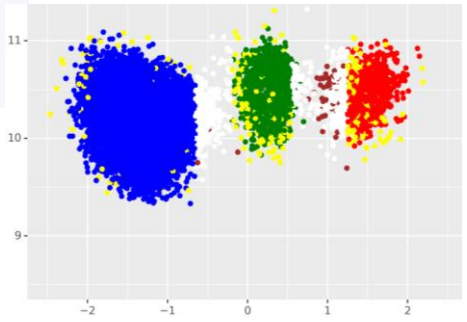
- Commercial service provider (Weatherby's Scientific) currently handles major gene analysis and reporting
- Any AI company, herdbook or individual breeder can submit request
- Manual process, involves checking individual animal on output from genotyping process
- Cost involved per major gene reported irrespective of royalty status

Major Gene Pipeline

Aim: *To develop an automated pipeline in ICBF to facilitate large scale routine reporting of major genes*

- Focused on IDBv5 genotypes
- QC metrics associated with SNP, genotype and plate
- QC metrics which are informative to aid genotyping process and improve reliability of genotypes
- QC metrics reviewed regularly



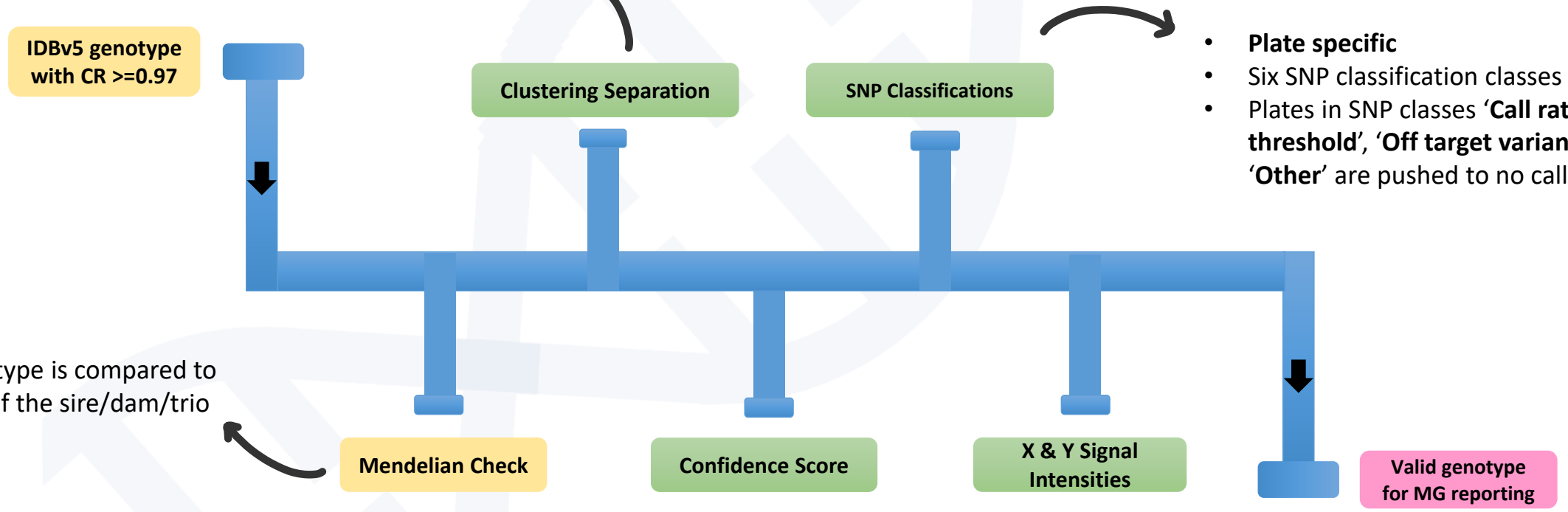


Pipeline Quality Control Metrics

Generic QC

Thermofisher QC

- **SNP and genotype specific**
- Contrast intensity thresholds for AA/AB/BB cluster



- **Plate specific**
- Six SNP classification classes
- Plates in SNP classes '**Call rate below threshold**', '**Off target variant**' or '**Other**' are pushed to no call

- Genotype is compared to that of the sire/dam/trio

- **SNP and genotype specific**
- Indicates the confidence of the genotype call

- **SNP and genotype specific**
- Minimum X and Y signal intensity thresholds

Good plate vs bad plate

PASS

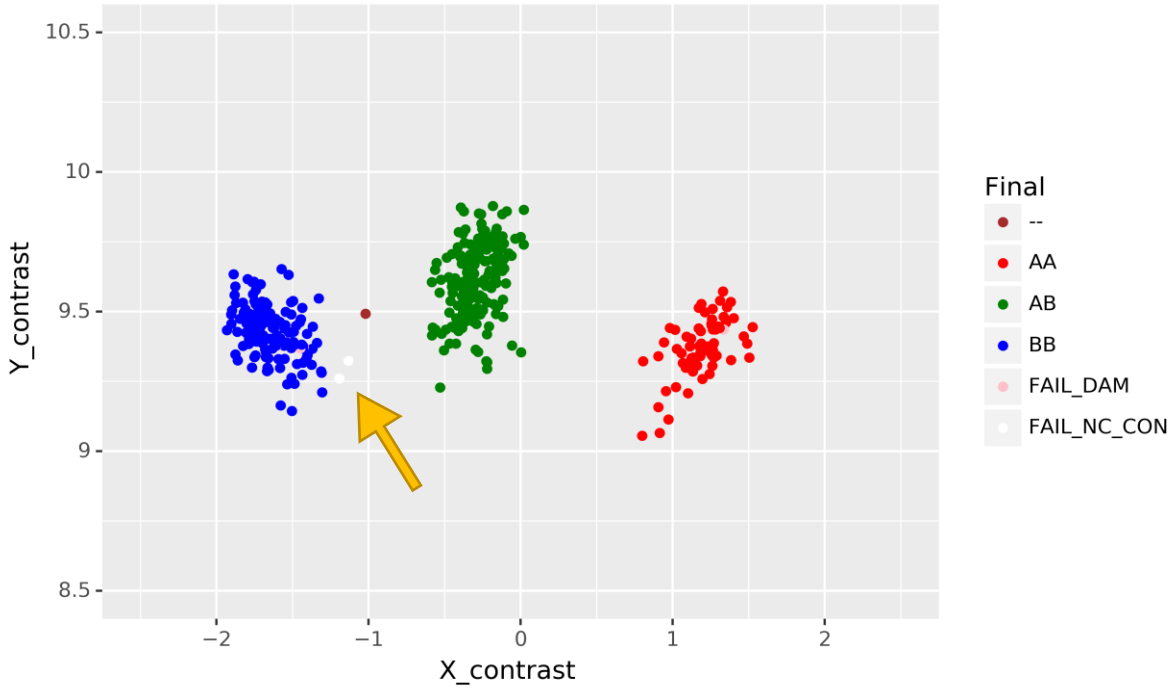


Plate classified as 'PolyHighResolution'

Two samples outside cluster separation thresholds

✓ **Well-formed, well-separated, distinct clusters**

FAIL

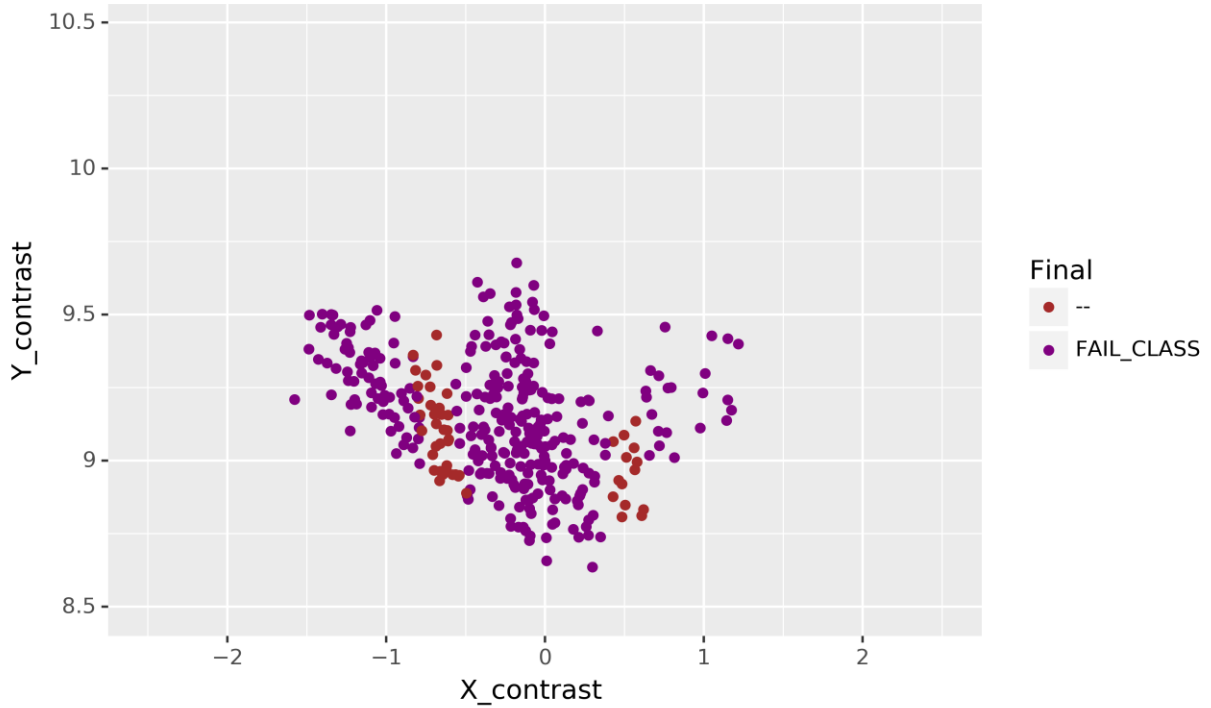


Plate classified as 'Other'

Poor cluster resolution

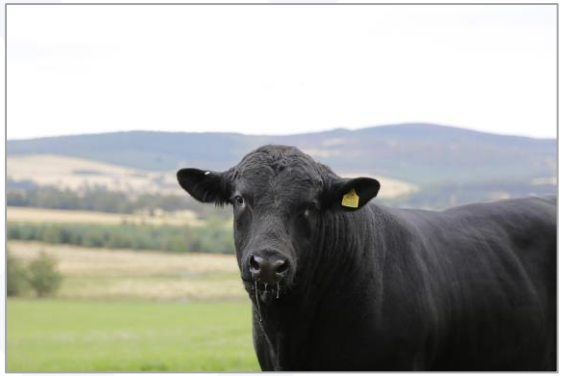
× **Merging, difficult to distinguish between clusters**

Reported Major Genes

- Pipeline is live since November 2022
- Live genotypes being released for 9

  Myostatin variants and Polled Celtic variant
- ~70 through validation pipeline – final checks

  before implementation.
- Extended to all non-royalty MG on IDBv5



Locus	Variant	Rs ID	Coordinates	Category
MSTN	*L64P*	*rs449270213*	2:6213889	Meat
MSTN	*F94L*	*rs110065568*	2:6213980	Meat
MSTN	*S105C*		2:6214012	Meat
MSTN	*nt419*		2:6215953	Meat
MSTN	*D182N*		2:6216072	Meat
MSTN	*Q204X*	*rs110344317*	2:6216138	Meat
MSTN	*E226X*		2:6216204	Meat
MSTN	*nt821del11*	*rs382669990*	2:6218379	Meat
MSTN	*C313Y*		2:6218499	Meat
POLLED	*Polled Celtic*		1:g.1706051_1706060del-ins1705834_1706045dup	Beneficial

Major Gene Pipeline Results

- >1 million genotypes (n=1,107,481) gone through the pipeline
- Pass Rate (PR) ranges from 91.5% to 99.1%
- Average pass rate for MG released to date is 95.75%

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 samples through the pipeline to date: 1,107,481

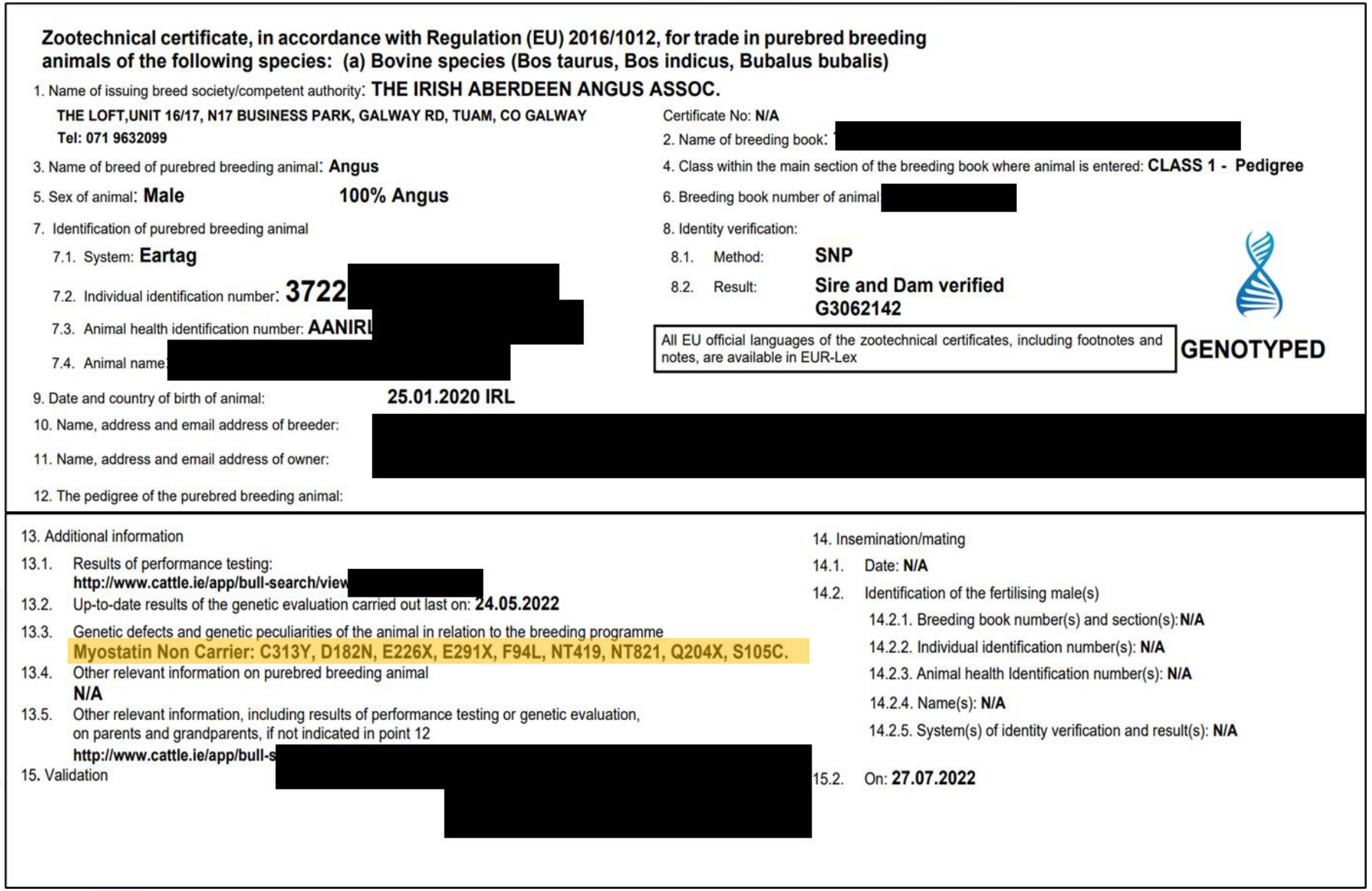
Breed Frequencies

- Frequencies of released MG in purebred populations

Breed	N	Polled Celtic	Myostatin								
			F94L	nt821	Q204X	C313Y	nt419	L64P	D182N	S105C	E226X
Limousin	7000	0.01	0.90	0.03	0.05						
Angus	5000	1.00		0.03							
Charolais	4900	0.01	0.14		0.14						
Hereford	3100	0.15									
Holstein	2800										
Simmental	1800	0.08			0.01						
Aubrac	800		0.84	0.04							
Dexter	700	0.06									
Saler	600	0.01		0.01							
Irish Moil	300	0.80	0.01								
Shorthorn	300	0.16	0.01	0.05							0.05
Blonde D'Aquitane	200		0.02								
Partenaise	200			0.91			0.01				0.05
Stabiliser	170	0.74	0.08	0.06							
Jersey	130										
Belgian Blue	120			1.00							

Reporting of MG

1) EU Zootechnical Certificates for all herd books in Ireland



Zootechnical certificate, in accordance with Regulation (EU) 2016/1012, for trade in purebred breeding animals of the following species: (a) Bovine species (Bos taurus, Bos indicus, Bubalus bubalis)

1. Name of issuing breed society/competent authority: **THE IRISH ABERDEEN ANGUS ASSOC.**
   THE LOFT,UNIT 16/17, N17 BUSINESS PARK, GALWAY RD, TUAM, CO GALWAY
   Tel: 071 9632099

Certificate No: **N/A**

2. Name of breeding book: [redacted]

3. Name of breed of purebred breeding animal: **Angus**

4. Class within the main section of the breeding book where animal is entered: **CLASS 1 - Pedigree**

5. Sex of animal: **Male** **100% Angus**

6. Breeding book number of animal [redacted]

7. Identification of purebred breeding animal

7.1. System: **Eartag**

7.2. Individual identification number: **3722** [redacted]

7.3. Animal health identification number: **AANIRL** [redacted]

7.4. Animal name: [redacted]

8. Identity verification:

8.1. Method: **SNP**

8.2. Result: **Sire and Dam verified**
G3062142

All EU official languages of the zootechnical certificates, including footnotes and notes, are available in EUR-Lex

GENOTYPED

9. Date and country of birth of animal: **25.01.2020 IRL**

10. Name, address and email address of breeder:

11. Name, address and email address of owner: [redacted]

12. The pedigree of the purebred breeding animal:

13. Additional information

13.1. Results of performance testing:
http://www.cattle.ie/app/bull-search/view [redacted]

13.2. Up-to-date results of the genetic evaluation carried out last on: **24.05.2022**

13.3. Genetic defects and genetic peculiarities of the animal in relation to the breeding programme
Myostatin Non Carrier: C313Y, D182N, E226X, E291X, F94L, NT419, NT821, Q204X, S105C.

13.4. Other relevant information on purebred breeding animal
N/A

13.5. Other relevant information, including results of performance testing or genetic evaluation, on parents and grandparents, if not indicated in point 12
http://www.cattle.ie/app/bull-s [redacted]

14. Insemination/mating

14.1. Date: **N/A**

14.2. Identification of the fertilising male(s)

14.2.1. Breeding book number(s) and section(s): **N/A**

14.2.2. Individual identification number(s): **N/A**

14.2.3. Animal health Identification number(s): **N/A**

14.2.4. Name(s): **N/A**

14.2.5. System(s) of identity verification and result(s): **N/A**

15. Validation

15.2. On: **27.07.2022**

Reporting of MG

2) Herd-profiles on ICBF website

Animal Number:		**Genotype Received:**	01-JUL-22
Animal Name:		**Call Rate:**	.99475
Breed:	LM	**Chip Type:**	IDBV5
Birth Date:	30-MAR-22	**Genotype Valid:**	Yes
Death Date:			
Sire:			
Dam:			

Show 10 rows. Showing 1 to 10 of 10 entries — Hide filters — Excel PDF Print — First Previous 1 Next Last

Major Gene | Type | Code | Quality Check | Result

Major Gene	Type	Code	Quality Check	Result
Myostatin C313Y	Meat	MYO_C313Y	PASS	NO COPY
Myostatin D182N	Meat	MYO_D182N	PASS	NO COPY
Myostatin E226X	Meat	MYO_E226X	PASS	NO COPY
Myostatin F94L	Meat	MYO_F94L	PASS	SINGLE COPY
Myostatin L64P	Meat	MYO_L64P	PASS	NO COPY
Myostatin NT419	Meat	MYO_NT419	PASS	NO COPY
Myostatin NT821DEL11	Meat	MYO_NT821	PASS	NO COPY
Polled Celtic	Beneficial	POLL_C	PASS	DOUBLE COPY
Myostatin Q204X	Meat	MYO_Q204X	PASS	NO COPY
Myostatin S105C	Meat	MYO_S105C	PASS	NO COPY

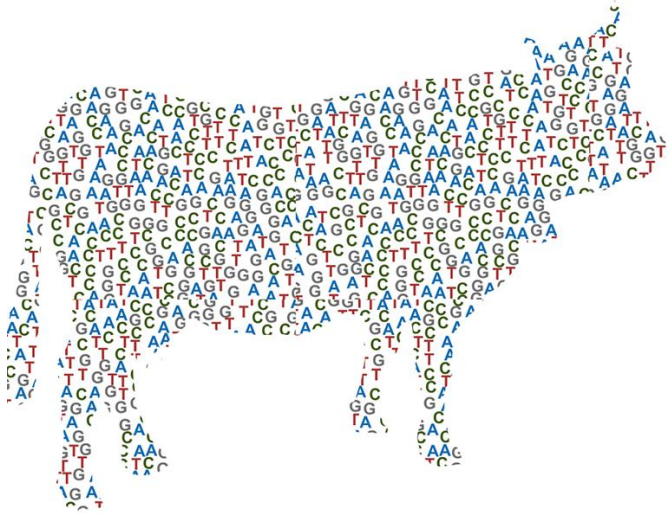
Showing 1 to 10 of 10 entries — First Previous 1 Next Last

Analysis Disclaimer: This analysis has been prepared in good faith on the basis of information provided to it.
No representation or warranty expressed or implied is made or given as to the accuracy, reliability, completeness or correctness of the analysis.
No liability is accepted for any losses (whether direct or indirect), damages, costs or expenses whatsoever, incurred or arising from any use of or reliance on the analysis or the information contained in it by any person.

Result: The animal will carry 'No Copy', 'Single Copy' or a 'Double Copy' of the gene variant

Summary

- Pipeline overlays additional QC metrics to improve the confidence and reduce errors/miscalling of major gene genotypes
- Why is it of benefits to breeders?
    - Valuable information on both desirable and undesirable genes segregating within herds
    - Informed breeding decisions, carry out selective matings
- Downstream benefits
    - Herd book major gene management
    - Farmers – Mating decision support tool (ICBF - Sire Advice)
    - Genomic evaluations – increased accuracy where dubious SNP calls are censored and imputed



Our Farmer & Government Representation

An Roinn Talmhaíochta, Bia agus Mara
Department of Agriculture, Food and the Marine

Irish Creamery Milk Suppliers Association

Feirmeoirí Aontaithe na hÉireann – The Irish Farmers' Association – IFA

Our AI & Milk Recording Organisations

DoveaGENETICS

MUNSTER BOVINE
Collectively Improving Farm Life

PROGRESSIVE GENETICS

TIPPERARY COOPERATIVE

Our Herdbooks

Pasthenais

Irish Angus Cattle Society

The Irish Aberdeen-Angus Association

Irish Simmental

Jersey Cattle Association of Ireland

Irish Limousin Cattle Society

IHFA – Irish Charolais Cattle Society

Irish Shorthorn Society Ltd.

Salers Cattle Society of Ireland

Irish Hereford Breed Society

Irish Aubrac

MRI Cattle Society of Ireland
Meuse Rhine Issel -- Milk & Muscle!

AYRSHIRES

EASBB

The Irish Moiled Cattle Society

Blonde D'Aquitaine

Irish Belgian Blue

Droimeann Cattle Society

ISP

Montbeliarde

Ireland's native dairy breed

Norwegian Red Cattle Society

icbf Acknowledging Our Members

Validation

- Phenotypic identification (hetero, homo, non-carrier)
- Genotype confirmation by targeted sequencing (Sanger)
  - Identify control samples using ICBF database (>3M)
  - Pilot QC ongoing for MSTNs, PC and PF
  - Genotype the validated control samples
- Bioinformatic validation of the array probe designs
  - Alignment analysis – BLAST
  - Original publication (OMIA)

Definitions of metrics

Animal CR – 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.

Cluster separation: Clustering separation values are specific to each SNP and each genotype. Approx 300k genotypes are manually inspected to determine thresholds for AA, AB and BB clusters.

SNP Classification categories – Based on QC metrics: Call rate, Fishers Linear Discriminant (FLD), Heterozygous Strength Offset (HetSO), Homozygote Ratio offset (HomRo), nMinorAllele.

X and Y signal intensity - X and Y raw signal intensity thresholds of 500-700. Low intensities indicate poor confidence in the genotype.

Heterozygosity check – For rare alleles e.g. S105C, where plates with >50% AB/BB are pushed to no calls.

Confidence Score

- The confidence score can be described as 1 minus the posterior probability of the point belonging to the assigned genotype cluster.
- It can range between zero and one, with lower confidence scores indicating more confident genotype calls.
- TF default is 0.15, ICBF set custom thresholds for each genotype for each SNP.
- Takes into consideration:
  - All variations of CS and X contrast thresholds to see impact across individual plates, 12 plate (~4k) batches
  - SNP classifications
  - Review pass rates
  - Check breeds (impact of background)
  - Breed frequency
  - Sample type (ear/hair)

TF SNP Classification Categories

SNP Classification Category	Description
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).
OTV	Off target variant refers to sample cases where the 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 good
Other	Indicates an issue with a SNP - one or more QC metrics are below the thresholds. Expect lower quality genotypes.

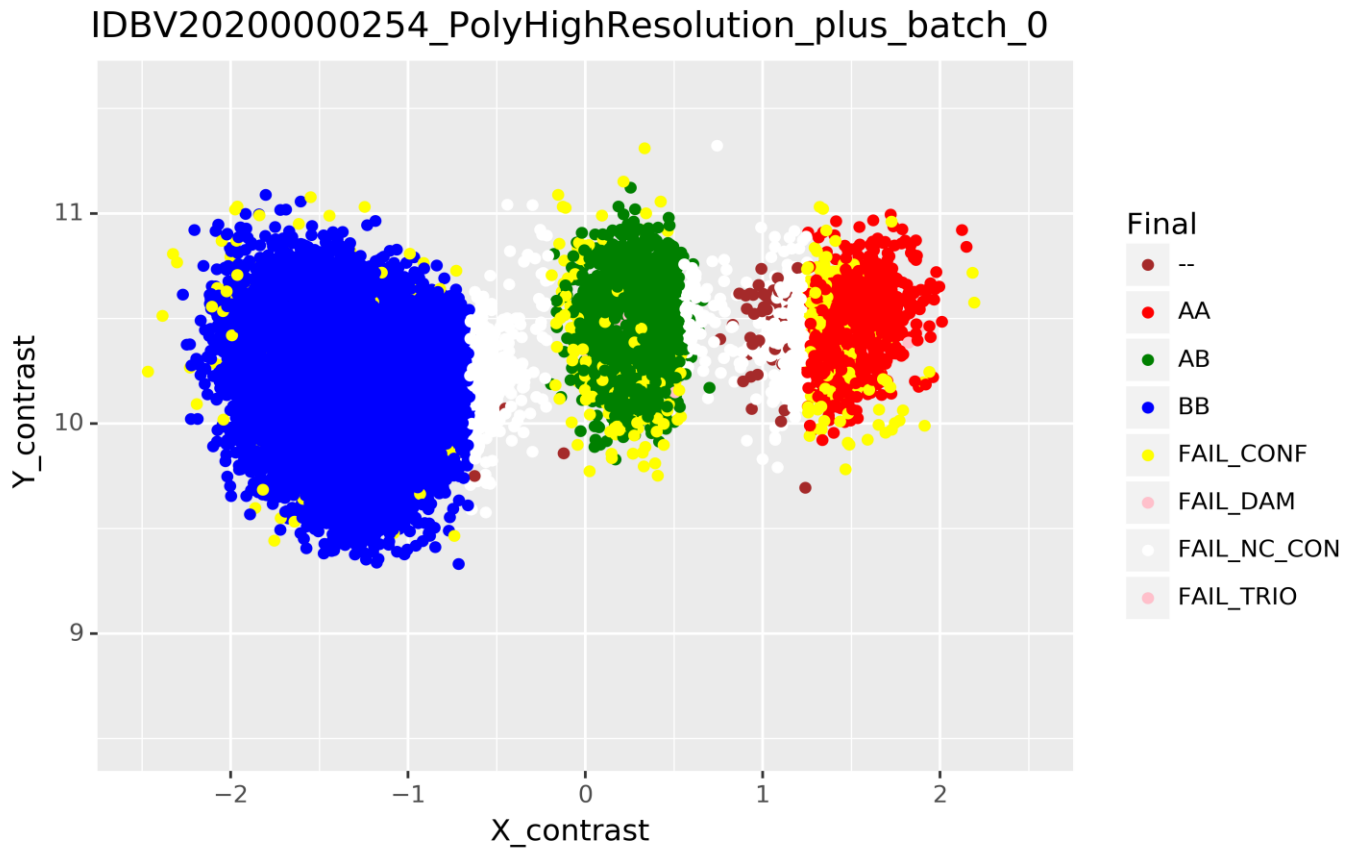
Concordances – v5

	Average % (Min-Max)	Median (%)	#MG 100%
IDBv5 – MG (107/165) concordance	99.05 (73-100)	100	53 (15 mono)
IDBv5 manifest wide concordance	99.77 (40-100)	100	NA

Failed samples based on QC metrics

	FAIL_CONF	FAIL_CLASS	FAIL_NC	FAIL_NC_CON	FAIL_DAM	FAIL_NC_XY	FAIL_SIRE	FAIL_TRIO
S105C	11635		1147	41257	16		1	
nt419	7507	41880	993	44	120		466	5
E226X	8198	34146	1483	2315	1			10
L64P	11616	1981	319	1				1
D182N	13882		405	2533				2
F94L	35686	2535	1065	2559	845		8	515
Q204X	2942		221	1791	20		146	88
nt821del11	9025		97	147	214		100	222
C313Y	2068		223	1	5			3
Polled Celtic	37883	3228	1182	5650	239	17	2	450

SNP and genotype specific metrics



- White samples – pushed to no calls where samples lie outside cluster separation thresholds
- Yellow samples – pushed to no calls as they lie outside the confidence score thresholds

Drivers and Challenges for new uptakes of new sources and uses of data recording

- Data leveraged from one genotype
- Major gene reporting - benefits to breeders
  - Valuable information on both desirable and undesirable genes segregating within herds
  - Informed breeding decisions
- Herd book major gene management
- Farmers – Mating decision support tool (ICBF - Sire Advice)
- Genomic evaluations