Final report

Validation of Pooled DNA GEBV for Brahman Commercial Cow Fertility

Project code: P.PSH.1211

Prepared by: Yutao Li, Laercio Porto-Neto, Russell McCulloch, Sean McWilliam, Pamela Alexandre, Sigrid Lehnert and Toni Reverter
Commonwealth Scientific and Industrial Research Organisation

Date published: 29 April 2023

PUBLISHED BY
Meat & Livestock Australia Limited
PO Box 1961
NORTH SYDNEY NSW 2059

Meat & Livestock Australia acknowledges the matching funds provided by the Australian Government to support the research and development detailed in this publication.

Abstract

This project was undertaken to validate and estimate the cost-benefit for genomic predictions of Brahman cow fertility based on data (DNA and phenotypic information) from commercial Brahman cow herds. During the lifetime of the project, a total of 16,861 DNA samples were acquired. 3,442 of these samples were from cows which had 2nd joining pregnancy and lactation status (PLS) records. In order to estimate the power of genomic predictions built on pooled genotyping, 1,135 of these DNA samples were individually genotyped and 322 DNA pools were assembled, based on PLS phenotype classes. In addition, 2,418 individual bull genotypes were collected from MDH herd bulls. The genotype information from all animals was imputed to a high-density panel of >500K SNP. The impacts of different SNP densities (low vs HD), and different ways of deriving genomic relationship matrices (DNA-pooling vs individual genotypes) on the estimates of gEBV were evaluated across populations. The team developed new method (GA2CAT: Genomic Attribution to Categories with PLS being the focus of "categories") to genomically rank populations for traits defined by non-ordinal multi-class categories such as PLS. The method worked well when the categorical traits are distinguishable between classes and the method can be easily extended to deal with other phenotypes with non-ordinal multi-class features. Two executable programs for automatically generating GA2CAT values were developed and tested. We found that the genomic predictions using pooled DNA vs individually genotyped animals were very lowly correlated. The genomic ranking of candidate bulls using DNA pools of cows differed from rankings based on individual genotypes. We concluded that genomic prediction of a lowly heritable trait such as PLS needs to be based on individual genotypes. The preliminary results from a cost-benefit study showed that for a base herd of 1,500 animals, improving the re-conception rate by 2% on first calf heifers would yield an increase in the herd gross margin (GM) by $1,129.81 after taking into account the GM interest or $0.75 GM per adult equivalent.

Executive summary

Background

The productivity of cow herds is an important underpinning metric for industry profitability, and as the CRC for Beef Genetic Technologies demonstrated, a clear target for genetic improvement. In previous work, CSIRO and The University of Queensland have worked with Brahman producers and bull breeders to demonstrate that reproductive performance data from commercial Brahman cow herds can be used as phenotypic information to drive genomic predictions of daughter fertility in bulls. The current project was undertaken to move this novel application of genomics closer to commercial application by calculating the accuracy of the DNA-based predictions (typically denoted as GEBV for genomic-estimated breeding values) and by determining their cost-benefit to producers and bull-breeding operations. In addition, this project will be a major source of genotypic and phenotypic data for beef cow fertility which can be accessed and utilised by other R&D projects in the NLGC (for example Repronomics, Northern Genomics Project, Female Reproduction Phenobank), as well as possibly contributing valuable data to genetic evaluation systems used by the wider industry, such as BREEDPLAN.

Overall Project Objectives

- Determine the accuracy of pooled-DNA GEBV for commercial cow fertility within participating herds and for the Brahman breed more generally
- Calculate the cost-benefit of adopting the technology
- Build the "reference population" for Brahman cow fertility GEBV by collecting additional phenotypes and genotypes
- Develop methodology for integrating GEBV for commercial cow fertility into Brahman group BREEDPLAN services

Methodology

- The use of animals in research was approved by CSIRO's Queensland Animal Ethics Committee;
- MDH Pty Ltd provided access to commercial Brahman cow herds for phenotype collection at pregnancy test muster and DNA sample collection by Tissue Sampling Units (TSU) at Iffley Station, Queensland for this project. Candidate herd bull tail hair samples were provided by Gipsy Plains Brahmans, Cloncurry, Queensland.
- Pregnancy and lactation status (PLS) was recorded for 3,442 Brahman cows following their second joining, and tissue samples for DNA were collected for the study;
- All bull samples collected were individually genotyped with 50K Bovine SNP chip. For cow samples, both DNA pooling and individually genotyping were applied.
- Imputation of 50K genotypic data to high-density (> 500 K SNP) panel using the high-density genotypes of 861 BeefCRC Brahman cattle as the refence was applied to all samples collected during the project life.
- Hybrid-GRM (combining genotypes from both individual DNA samples and pooled DNA samples) and Principal Component Analysis were applied to investigate and display genomic relationships between bulls and reference cows of difference resources;
- Genomic ranking of Brahman bulls using either pooled genotypes or individually genotyped cows as the reference population was performed to all candidate bulls using both conventional GBLUP models (genomic best linear unbiased prediction) and a newly

developed method, genomic attributions to Pregnancy and lactation status (GA2PLS) and later changed to genomic attributions to a categorical trait (GA2CAT).

- A cost-benefit analysis was conducted to investigate the suitability of economic models to estimate the economic impact of implementing our genomic predictions.
- The genomic estimates from using the new novel genomic prediction method were compared with the EBV estimates for Days to Calving (DTC) for the Repronomics data sets. The method was also compared with the most commonly used machine learning methods (Random Forest and Supporting Vector Machine).

Results/key findings

- A new method was developed to genomically allocate animals into phenotypic categories. The testing and validation of the GA2CAT algorithm have been performed comprehensively using the existing MDH animals (cows and bull populations) and external populations that have the traits of non-ordinal multi-class categories. The results showed that the method worked well when the categorical traits are distinguishable between classes with a strong genomic background (e.g. breed types) and the method can be easily extended to deal with other phenotypes with non-ordinal multi-class features;
- Two programs, a UNIX Shell script and a FORTRAN95 source code, have been developed for the efficient implementation of the GA2CAT algorithm. They have been independently tested with identical results. The programs can be used for rapidly generating GA2CAT values for candidate bulls of any population;
- The genomic prediction of 1,451 MDH and Gipsy Plains bulls from 2020, 2021, and 2022 seasons were performed to estimate their contributions to their female progeny's reproductive performance using GA2CAT. The rankings of the animals were provided to the farms on time to assist their sire selection decision;
- There were very low correlations between the genomic predictions using pooled DNAs vs individually genotyped animals. The genomic ranking of candidate bulls using DNA pools of cows of the same phenotype of PLS had different outcomes when using individually genotyped cows. The reproductive trait – pregnancy and lactation status is a lowly heritable trait which makes the development of accurate genomic predictions challenging. As a precautionary measure, we recommend the use of individual genotypes to make genomic predictions of bulls' reproductive performance;
- A preliminary study on the cost:benefit analysis using the BREEDCOW+ herd modelling software showed that for a base herd of 1500 animals, improving the re-conception rate by 2% on first calf heifers would yield an increase in the herd gross margin (GM) by $1,129.81 after taking into account of the GM interest or $0.75 GM per adult equivalent.

Benefits to industry

- The new metric – weighted average of GA2CAT values for different categories, is easy to derive and most importantly it provides commercial bull-buyers a single 'ranking' value for each candidate bull so that they can readily make the decisions about which bulls to choose;
- The population variation of GA2CAT values reflects the degree of relatedness between testing and reference populations: as would be expected, GA2CAT predictions are valid for the most closely related cattle populations;

- A newly developed executable computer program for routine calculation of GA2CAT values for candidate bulls will promote adoption by commercial producers wanting to improve the reproductive performance of their herds.

Future research and recommendations

- As a low heritability and imbalanced multi-class trait, the genomic selection for 2nd joining pregnancy and lactation status presents serious challenges. The new metric "GA2CAT" is recommended to apply in conjunction with individual bull's BBSE (Bull Breeding Soundness Examination) results when selecting candidate bulls for mating;
- Due to 2019 flood, the project could only collect the candidate bulls from Gipsy Plains during the 2021-2022 and 2022-2023 seasons. As result, we have not been able to conduct a systematic validation of the accuracies of genomic prediction by our new method GA2CAT for all individual bulls that sired the MDH cows (with 2nd joining pregnancy and lactation status records) from 2019-2022. A future study is needed to evaluate the accuracy of the new prediction tool;
- Further research is required to identify a new phenotype that is more heritable than the phenotype of pregnancy and lactation status.

Table of contents

1. Background

Genomic prediction of female reproductive performance

Female reproductive traits directly impact the profitability of commercial beef herds. Female reproductive performance can be measured by a range of traits in dairy and beef cattle, either being continuous (e.g. age at puberty, days at first calving), binary (e.g. pregnancy status), or count (e.g. number of inseminations) (Toghiani et al., 2017). The ability to select herd bulls based on the predicted reproductive performance of their female progeny has the potential to significantly improve herd productivity. In extensive production systems of tropical and subtropical northern Australia, uptake of genomic selection has been limited due to practical, financial, and management difficulties associated with collecting individual performance records. In northern commercial cattle herds, following natural multiple-sire joining, heifers are mustered and grouped based on the result of their 2nd joining pregnancy and lactation status (PLS, Table 1). The information that is routinely obtained at the pregnancy testing muster allows cows to be grouped into six categories: 1. DNP = Dry and Not Pregnant; 2.WNP = Wet and Not Pregnant; 3. DEP = Dry and Early Pregnant; 4. DMP = Dry and Mid Pregnant; 5. DLP = Dry and Late Pregnant; 6. WEP = Wet and Early Pregnant. When heifers are genotyped, either individually or pooled as a group based on these categories, their DNA profiles become a valuable resource to explore genomic selection strategies for the improvement of fertility traits in beef cattle (Reverter et al., 2016). The current project was designed to move this novel application of genomics closer to a commercial application by calculating the accuracy of the DNA-based predictions (typically denoted as GEBV for genomic-estimated breeding values) and by determining their cost-benefit to producers and bull-breeding operations. In addition, this project has been a major source of genotypic and phenotypic data for beef cow fertility which can be accessed and utilized by other R&D projects in the NLGC (for example Repronomics, Northern Genomics Project, Female Reproduction Phenobank), as well as contributing valuable data to genetic evaluation systems used by the wider industry, such as BREEDPLAN.

Table 1. Definition of 2nd joining Pregnancy and lactation status of cow herds

Dry and empty (not pregnant)	DNP
Dry and early pregnant	DEP
Dry and mid pregnant	DMP
Dry and late pregnant	DLP
Wet and empty (not pregnant)	WNP
Wet and early pregnant	WEP
Wet and mid pregnant	WMP
Wet and late pregnant	WLP

2. Objectives

The overall project objectives include:

- **Determine the accuracy of pooled-DNA GEBV for commercial cow fertility within participating herds and for the Brahman breed more generally.**

Achieved. Genomic prediction of an individual bull’s contribution to its female progeny was initially performed using conventional GBLUP models with pooled-DNA genotypes of cow herds as a reference population to estimate the GEBVs for individual bulls of different populations. The PLS was treated as a continuous trait by giving scores 1 to 6 from lowest to highest reproductive performance to the 6 categories of PLS. A series of studies were conducted to investigate the impacts of different SNP densities (low vs 50K, 50K vs imputed high-density (HD), and different coding systems to PLS on the genomic ranking of candidate bulls. The results indicate that the high-density SNP panels captured the genetic difference better between reference cows and candidate bulls when pooled genotypes were used. In addition, the results reveal some challenges to the analytical methods. Firstly, single GEBV values from a GBLUP model for individual bulls are difficult to make a biological meaning, given that the PLS is a non-ordinal multiclass phenotype. Secondly, the genomic rankings of candidate bulls are significantly impacted by which score is given to the phenotypes. For example, arguments could be made for the Wet and Non-pregnant phenotype to be scored as either 2 or 5. Therefore, a new method (GA2CAT) was developed for the project (see section 2.4). With it, systematic comparisons were made for genomic rankings of 1,451 MDH and Gipsy Plains bulls using 290 pools of commercial cows as the reference vs those using 1,135 individual cows. There were very low correlations between the genomic predictions using pooled DNAs vs individually genotyped animals.

- **Calculate the cost-benefit of adopting the technology.**

Achieved. The team investigated the suitability of economic models to estimate the economic impact of implementing our genomic predictions. Based on the genetic parameters of the traits studied, we estimated that a 2% increase in the reconception rate of first calf heifers would be achievable in a 10-year timeframe. A preliminary study on the cost: benefit analysis using the BREEDCOW+ herd modelling software showed that for a base herd of 1,500 animals, improving the re-conception rate by 2% on first calf heifers would yield an increase in the herd gross margin (GM) by $1,129.81 after taking into account of the GM interest or $0.75 GM per adult equivalent. The independent assessment of the CSIRO work found that the BreedCow analysis the team had performed is appropriate as an initial estimate of benefit from the genetic changes estimated, and the GMS/AE we derived from BreedCow are within the range expected for the region. A broad suggestion for consideration by CSIRO, MLA, & Bush AgriBusiness is to consider benefit-cost analyses of genetic technology investments at a range of scales e.g. herd, enterprise, region, nationally. Since the scale of the proposed further cost: benefit analysis is beyond the scope of the current project, it is agreed that it should be conducted as an independent project of MLA.

- **Build the “reference population” for Brahman cow fertility GEBV by collecting additional phenotypes and genotypes**

Achieved. The project has built a reference population of both individual and pooled genotypes for “pregnancy and lactation status” on MDH Brahman cows (1,457 samples, see Table 2) and 2,418 individual bulls. The genotypes and phenotypes of the cow population have been used as a reference to predict the genomic ranking of the bulls. In addition, we have also acquired three additional genotypic and phenotypic datasets on Brahman animals from different parts of the industry for validation purposes. These include 1) Repronomics project dataset (MLA project B.NBP.0759), comprised of 2,147 samples with the genotypes from various SNP densities (ranging from 25,651 to 74,153 SNPs) and GEBV for days to calving (DTC) and accuracy for DTC; 2) the MLA Bull Fertility Project dataset (L.GEN.1818), of 6,063 bulls with the genotypes of imputed high density SNPs (522,549) and the GEBV of 10 traits (WT, CS, SC, Sheath score, density, mass, mot, PNS, PD and MP), and 3) The genotype and GEBV information of 71 animals from Roxborough Brahmans.

Table 2 summarises the genomic and phenotypic resources that the project has assembled. A total of 16,861 samples have been genotyped since commencing the project, and their genotypes have been imputed to a high-density panel of 530,000 autosomal SNPs using a population of 861 Beef CRC Brahman cattle that were genotyped with the high-density of 727,270 SNPs as reference. All genotypic and phenotypic data acquired by the project have been deposited in the secure CSIRO big data storage portal.

- **Develop methodology for integrating GEBV for commercial cow fertility into Brahman group BREEDPLAN services**

Achieved. The project developed a novel computational genomic methodology for estimating individual sire's "genomic attribution to a Categorical Trait" (GA2CAT). This method estimates an individual bull's contribution to six categories of their female progeny's pregnancy and lactation status at the second joining, based on the genomic relationships between the bulls and cows with PLS records. In other words, the estimates depend on how closely related each DNA-tested bull to the most fertile cows in a herd with PLS records.

For exploring the value of this prediction in the context of GEBV methodology currently used in the industry, we carried out a series of analyses on different populations using both conventional GBLUP models and the newly developed GA2CAT method to compare the animal rankings. These include: 1) Testing GA2CAT method in bull selections for MDH populations; 2) Validating the utility of GA2CAT method by comparing bull rankings using the GA2CAT estimates versus those using the GEBV of bull reproductive traits from the MLA Bull Fertility Project (L.GEN.1818); 3) Comparing cow fertility rankings using the GA2CAT estimates versus those using BREEDPLAN GEBV from the Repronomics project (B.NBP.0759). 4) Comparing the performance of the method with two commonly used machine learning methods (Random Forests and Supporting Vector Machine). The results indicate that the PLS is a low heritability reproductive trait, and there were low correlations between estimates from GA2CAT and conventional GBLUP models.

3. Methodology

3.1Collection of phenotypic data, DNA samples and genotypic information

McDonald Holdings Pty Ltd (MDH) is one of Australia's largest family owned and operated beef cattle operations with historical ties dating back to 1827 with the first consignment of cattle to Tasmania. To date the family represents seven generations of Australian beef producers. The McDonald family has 14 properties, collectively covering an area of 3.36 million hectares and runs approximately 150,000 head of cattle. The MDH breeding herd is generally located in the Gulf of Carpentaria and Cape York areas of Northern Queensland and are predominantly the Bos indicus Red Brahman breed and Brahman cross breeds. Brahman calves reared on the Northern Queensland properties are backgrounded in the Winton, Cloncurry central highlands and then grain fed finished at the Wallumba feedlot on the Western Darling Downs. MDH has been CSIRO's industry partner for two projects, QLD Smart State 2012 to 2014, and the current MLA project. Both projects involved using genomic technologies to improve bull reproductive performance in the Brahman breed of cattle. It is also acknowledged the long-standing relationship MDH has with its sole Brahman bull supplier Gipsy Plains and their cooperation with CSIRO during both projects. Two-year-old bulls are joined with maiden cows at an age of 18 months to 2 years old and remain together for life. Bulls are run at around 4% of the female herd and the average herd size is approximately 2000 to 3000 head depending on paddock size. Cow herds are normally pregnancy tested once a year to wean calves.

This structure provides a good opportunity for accessing large cohorts of females at their second breeding opportunity under commercial conditions. The long-standing association with the same bull breeder also provided an opportunity to develop genomic predictions of daughter reproductive performance for prospective herd bulls available for selection at the bull breeders.

The collection of pregnancy data and biological samples from cow herds following their second joining was an important component for building the reference population of pooled Brahman cows with commercial fertility data. It allowed us to test what impact the size of reference populations of pooled and individual genotypes has on the accuracy of the GEBV for commercial cow fertility. An additional objective of the project was to test options for integrating GEBV for commercial cow fertility obtained by pooling, with genetic predictions of fertility traits that were available via the BREEDPLAN national genetic evaluation system.

As the project was conducted on a commercial property it encountered the same challenging conditions faced by the industry as a whole. The 2019 floods in North Queensland that meant the bull breeder (Gipsy Plains Brahmans) was unable to provide surplus herd bulls for MDH to make a selection. The travel restrictions imposed both by MDH and CSIRO during 2020 and 2021 due to COVID-19 meant that project team members were unable to travel to Iffley station to attend the pregnancy test muster.

The persistent drought conditions at Iffley Station during 2020 and 2021 saw the cow herd we needed to target at their 2022 pregnancy test muster being transported to Rutland Plains station.

We were able to resolve the issues by:

1. Substituting bull samples by using archived tail hair samples from the consignment of “maiden bulls” delivered to Iffley in 2017 (n = 120), as well as archived tail hair samples from a representative sample of herd bulls in use at Iffley station in 2017 (n = 370);
2. Using 895 tail hair samples from Gipsy Plains Brahmans in 2021;
3. Designing a protocol for the MDH station manager and station staff to carry out the phenotype recording (see Table 1) of veterinarian pregnancy test results and TSU sample collection in both 2020 and 2021. These resulted in the collection of phenotypes and DNA samples from 2,452 (2020 season) and 446 (2021 season) cows at Iffley Station.
4. With the lifting of the travel restrictions in 2022, CSIRO scientists were able to travel to Rutland Plains to attend the pregnancy test muster and the collection of 990 DNA samples from cow herds in 2022.

From 2022, 2-year-old bulls under consideration by MDH had undergone Bull Breeding Soundness Examination (BBSE). Phenotypic data include birth month, coat colour, live weight, scrotal circumference, sperm motility (percent motile), sperm morphology (percent normal sperm), observations of anatomical abnormalities and poll/horn status.

For 2023 herds, although there were about 700 bulls at Gipsy Plains, only 574 were red bulls (the population used to supply herd bulls for Iffley) and the rest were grey Brahmans. Our previous research has shown that the genomic prediction would only work if the reference cow population is red Brahman. After double-checking with the bulls previously genotyped, the samples from 324 bulls were provided.

In addition, we successfully negotiated collaborative agreements for data sharing with the Australian Brahman Breeders Association, MLA and AGBU. Through these, we acquired three additional genotypic and phenotypic datasets on Brahman animals from different parts of the industry for validation purposes. These include:

1) Repronomics project dataset (MLA project B.NBP.0759), comprised of 2,147 samples with the genotypes from various SNP densities (ranging from 25,651 to 74,153 SNPs) and GEBV for days to calving (DTC) and accuracy for DTC;
2) The bull Fertility Project dataset (MLA project L.GEN.1818), of 6,063 bulls with the genotypes of imputed high-density SNPs (522,549) and the GEBV of 10 traits (WT, CS, SC, Sheath score, density, mass, mott, PNS, PD and MP); and
3) The genotype and GEBV information of 71 animals from Roxborough Brahmans.

Table 2 summarises the genomic resources containing the information for both bulls and cows collected/acquired during the project. A total of 16,861 samples were assembled and their genotypes were imputed to a high-density panel of 530,000 autosomal SNPs using the high-density 700K SNPs of 861 Beef CRC Brahman cattle as a reference genome. All genotypic and phenotypic data acquired by the project have been deposited in the secured CSIRO big data storage portal.

Table 2. **Genomic resources of 16,861 cattle samples collected/acquired with imputed high-density genotypes (> 500K SNP)**

Project/population	Description	# Samples	# SNPs	Sub-total
Commercial Cow Fertility P.PSH.1211	Pooled Brahman cow genotypes 2020 (with PLS)	233	54,791	
	Pooled Brahman cow genotypes 2022 (with PLS)	59	54,791	
	Pooled Brahman cow genotypes 2022 (with PLS)	30	54,791 777,962	322
	Individual Brahman cow genotypes 2020 (with PLS)	546	54,791	
	Individual Brahman cow genotypes 2021 (with PLS)	249	54,791	
	Individual Brahman cow genotypes 2022 (with PLS)	290	54.791	
		50	54,791 777,962	1135
	Brahman maiden herd bulls 2019	114	54,791	
	Brahman herd bulls for selection 2020	482	54,791	
	Brahman (Grey) herd bulls for selection 2021	160	54,791	
	Brahman (Red) herd bulls for selection 2021	229	54,791	
	Brahman herd bulls for selection 2022	486	54,791	
	Brahman herd bulls for selection 2022/2023	622	54,791	
	Brahman herd bulls for selection 2022/2023	325	54,791	2418
Beef CRC	Brahman genotypes	861	727,270	
		5040	Imputed 722,208	5901
Roxborough Brahmans	Brahman stud genotypes and EBV	71	Various	71

Smart Futures	Brahman herd bulls deployed 2013-2014	177	74,584	
	Pooled Brahman cow genotypes 2012-2014 (with PLS)	113	74,584	290
Repronomics B.NBP.0759	Brahman cows (with DTC EBV and accuracy)	2,118	25,651 ~ 74,153	2,118
Bull Fertility L.GEN.1818	Tropical bulls with male repro phenotypes (WT, CS, SC, Sheath, dens, mass, mott, PNS, PD, MP)			
	Brahman (CRCBR)	1,051	522,549	
	Tropical Composite (CRCTC)	1,819	522,549	
	Santa Gertrudis (SGT)	929	522,549	
	Droughtmaster (DMT)	760	522,549	
	UltraBlack (BLK)	844	522,549	
	Belmont Red (BEL)	660	522,549	6,063
Total				16,861

3.2Genotype pools of commercial Brahman cows from MDH properties

Genomic prediction of breeding values based on a genomic relationship matrix has revolutionized the ability to identify genetically superior livestock for improving traits that are difficult to measure (van der Werf 2009). However, in commercial herds, it is impractical to individually genotype all animals. DNA pooling of cows with reproductive records can provide a cost-effective way for assessing and predicting the contribution of individual bulls to the fertility of their female offspring (Reverter et al, 2016).

Based on the recommendations for optimal pool size developed by Alexandre et al (2020), a computer script was developed to assign samples to pools, using the following guidelines: Pool sizes range from 5 to 12 and, within phenotype, pool size is as homogeneous as possible. Characteristics of laboratory equipment were considered when developing the logistics for building the pools. To this respect, we considered working on a plate of 12 columns by 8 rows. Each row corresponds to one pool and the columns across are the samples that go to the same pool. Therefore, there is a virtual limit of 12 samples per pool (using a single plate) and if we were to add more samples into a pool we would need to overflow to a second plate increasing the lab work considerably due to more material needed and chances of getting things wrong.

Genotyping of individual tail hair samples was conducted by Neogen Australasia, using the GGP TropBeef 50K chip. This genotyping platform has a large amount of overlap with the SNP platform used to acquire the Smart Futures legacy data available to the project, on which the cow fertility GEBV are based (Reverter et al. 2016).

Table 3 shows details of PLS records of 3,875 MDH cow samples collected from 2020 to 2022 and the information on the genotyped samples. Across all seasons, the majority of females were wet and non-pregnant (WNP), that is, they had weaned their first calves and failed to reconceive. Lactating cows with evidence of pregnancy (any stage) formed the smallest cohort. These observations underscore the importance of lactational anoestrus in commercial Brahman cow herds in North Queensland. Apart from 249 samples in 2021 that were individually genotyped, the DNA pooling and individual genotyping strategies were applied to the cows from both the 2020 and 2022 dry seasons. The reasons for adopting such genotyping strategies in these populations were a) cost-effective when applying a DNA-pooling strategy; b) the impact of DNA pooling and individual genotyping on genomic prediction accuracy of candidate bulls could be systematically evaluated for different populations.

Table 3. Summary of statistics of 3,875 MDH cows sampled from 2020, 2021 and 2022 seasons with second joining Pregnancy and Lactation Status (PLS) records and the number of genotyped cows

		2020			2021		2022		
Pregnancy and lactation status	Code	No. of Cows	No. DNA Pools	No. Individually genotyped	No. of Cows	No. Individually genotyped	No. of Cows	No. DNA Pools	No. Individually genotyped
Dry and empty (not pregnant)	DNP	232	23	92	32	32	61	59 (10 animals/pool)	61
Wet and empty (not pregnant)	WNP	1543	129	256	286	102	759		109
Dry and early pregnant	DEP	219	22	43	34	34	109		109
Dry and mid pregnant	DMP	205	23	23	47	47	45		45
Dry and late pregnant	DLP	192	24	71	15	15	6		6
Wet and early pregnant	WEP	61	12	61	19	19	10		10
Total		**2452**	233	546	**433**	249	**990**	59	340

3.3 Imputation of genotypic data

As shown in Table 2, the bovine genotypic data we acquired (MDH, BeefCRC, Repronomics, Bull Fertility and Roxborough) had been genotyped on a variety of different genotyping platforms. To combine the genomic information from different sources with various SNP densities for analyses, imputation of genotypes from low or medium SNP density to high density was carried out. The genotypes of 861 Beef CRC Brahman cattle acquired using the Neogen HD bovine SNP chip (727,270 SNP) were used as the reference genome. PLINK (Chang et al. 2015) and Eagle v2.4.1 (Loh et al. 2016) were applied for phasing and imputation, respectively. After quality checks with the threshold of R-square value > 0.8 and removing SNP on the X chromosome, the genotypes of MDH, Repronomics, bull fertility, and Roxborough animals were imputed to a high SNP density panel (~530,000).

3.4 Construction of Hybrid GRM

A marker-based genomic relationship matrix (**GRM**), ***G*** was constructed following the method described by Van Raden (2008):

$$\boldsymbol{G} = \frac{Z_a Z_a'}{2\sum_{k=1}^{m} p_k q_k},$$

where matrix ***Z_a*** has dimensions of the number of individuals (*n*) by the number of loci (*m*), with elements that are equal to $2 - 2p_k$ and $-2p_k$ for opposite homozygous (A_1A_1 and A_2A_2 respectively) and $1 - 2p_k$ for heterozygous genotypes, p_k is the frequency of allele A_1 of locus *k*, and $q_k = 1 - p_k$. For the DNA pooled samples, the B-allele frequencies from the genotypes of the pools of cows (≤0.25, >0.25 and <0.75 or ≥0.75, best fitted the three genotypes based on the individual DNA samples and the genotype call algorithm employed by Illumina) were converted into the three possible genotypes (i.e. 0, 1 and 2 for AA, AB, and BB, respectively) and these were merged with the individual genotypes of each bull to generate a single "*hybrid*" GRM relating individually genotyped bulls with DNA pools of cows.

3.5 Principal Component Analysis

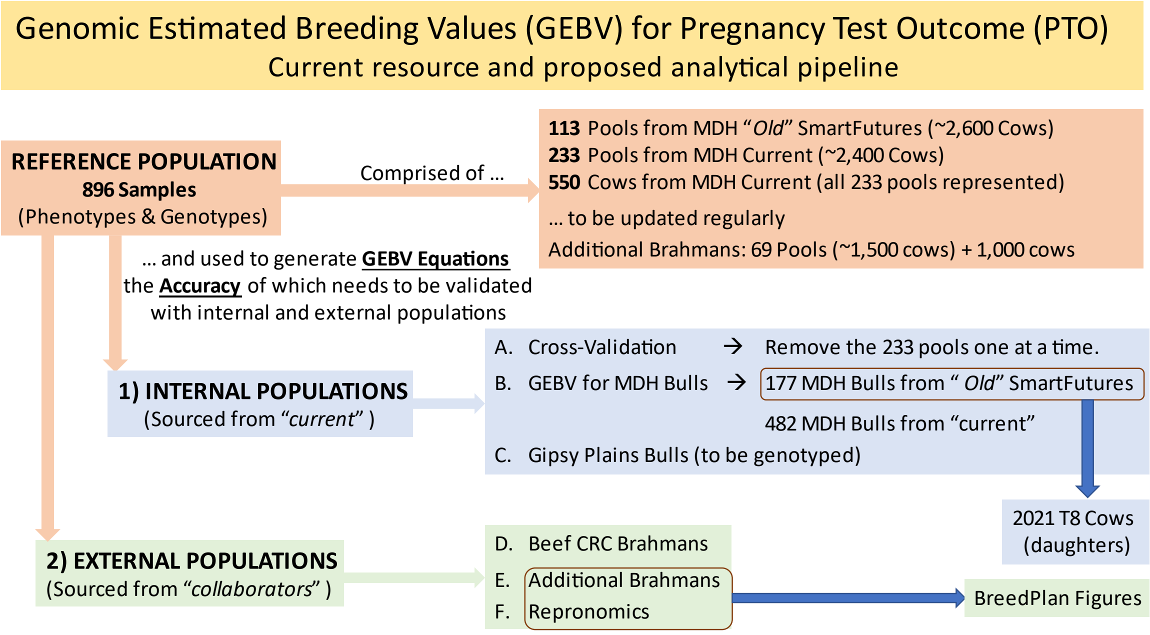
To assist in the visualisation of genetic relationships and population (sub) structure between the individually genotyped bulls of different populations and MDH cows (either DNA pooled or individually genotyped), Principal Components Analysis (PCA) was performed using the software PLINK (Chang et al. 2015).

3.6 Determine the accuracy of pooled-DNA GEBV for commercial cow fertility within participating herds and for the Brahman breed more generally

Genomic prediction of an individual bull's contribution to its female progeny was initially performed for the Brahman bulls from MDH and Gipsy Plains and later with the external populations, using conventional GBLUP models with pooled-DNA genotypes of cow herds as a reference population. The Pregnancy and lactation status (PLS) was treated as a continuous trait by giving scores 1 to 6 to the 6 categories of PLS.

Figure 1 illustrates the genomic resources and analytical pipeline we applied to estimate genomic breeding values (GEBVs) for Pregnancy and lactation status (PLS) of the female progeny of candidate bulls from different populations.

Figure 1. Genetic resources and the analytical pipeline applied to estimate and validate genomic breeding values (GEBVs) for Pregnancy and lactation status (PLS)


Genomic Estimated Breeding Values (GEBV) for Pregnancy Test Outcome (PTO)
Current resource and proposed analytical pipeline
REFERENCE POPULATION
896 Samples
(Phenotypes & Genotypes)
Comprised of …
113 Pools from MDH "Old" SmartFutures (~2,600 Cows)
233 Pools from MDH Current (~2,400 Cows)
550 Cows from MDH Current (all 233 pools represented)
… to be updated regularly
Additional Brahmans: 69 Pools (~1,500 cows) + 1,000 cows
… and used to generate GEBV Equations
the Accuracy of which needs to be validated
with internal and external populations
1) INTERNAL POPULATIONS
(Sourced from "current")
A. Cross-Validation → Remove the 233 pools one at a time.
B. GEBV for MDH Bulls → 177 MDH Bulls from " Old" SmartFutures
482 MDH Bulls from "current"
C. Gipsy Plains Bulls (to be genotyped)
2021 T8 Cows
(daughters)
2) EXTERNAL POPULATIONS
(Sourced from "collaborators")
D. Beef CRC Brahmans
E. Additional Brahmans
F. Repronomics
BreedPlan Figures

Since DNA pooling was based on the phenotype of muti-classes, it is unclear what density SNP panel we should use to rank bulls to achieve accurate prediction of reproductive performance of their progeny, therefore the first study was to investigate the impact of different SNP densities on genomic ranking of candidate bulls.

3.6.1 Impact of different SNP density panels on GEBV predictions of Brahman bulls for commercial cow fertility

3.6.1.1 Animals

Two datasets were used for the study. One is from the SmartF consisting of 290 samples from 2012-2014 herds (177 individual bulls and 113 pools representing 2,648 cows) genotyped with 74,584 SNPs (770K BovineHD BeadChip platform). The other (MDH2020) contains 715 samples from the 2020 herd (482 individual bulls and 233 pools representing 2,452 cows) genotyped with 54,791 SNPs (Neogen Australasia GGP TropBeef 50K chip). DNA pools were formed based on the pregnancy test (i.e. not pregnant or pregnant) and lactation status (dry or wet) of cows at 2nd joining. Details of the phenotype of pregnancy and lactation status (PLS) and pooling techniques can be found in Reverter et al. (2016).

In brief, animals were separated into 6 categories, that is, dry and empty (not pregnant, scored as 1), dry and early pregnant (scored 2), dry and mid pregnant (scored 3), dry and late pregnant (scored 4), wet and empty (not pregnant, scored as 5), and wet and pregnant (scored 6). DNA samples of animals with identical phenotypic scores were pooled together. The individual pool size ranged from 4-45 animals for SmartF (Reverter et al., 2016) and from 5-12 animals for MDH2020, depending on the number of animals available in each category. Details of the two datasets are presented in Table 4.

Table 4. Composition of two genotyped populations: Smart Futures (SmartF) and 2020 MDH (MDH2020)

Population	Sex	Year	DNA samples	Total
SmartF	Cows	2012	41 (pools)	
(74,584 SNPs)		2013	31 (pools)	
		2014	41 (pools)	113
	Bulls	2013	27	
		2014	150	177
MDH2020	Cows	2020	233 (pools)	233
(54,791 SNPs)	Bulls	2020	482	482

Between the two populations, there were 19,089 SNP in common. The imputation from low to high-density SNP genotypes was conducted to both SmartF and MDH2020, using 730,000 SNPs from 5,040 Beef CRC Brahman cattle as the reference. PLINK (Change et al. 2015) and Eagle v2.4.1 (Loh et al. 2016) were applied for phasing and imputation, respectively. After quality checks with the threshold of R-square value >0.8 and removing SNPs on the sex chromosome, this resulted in 615,310 SNPs. To visualize genetic relationships between two populations, we conducted a PCA using genotypes from either the low density (19,089 common SNP) or imputed high-density panel (615,310 SNP, HD).

3.6.1.2 Genomic estimated breeding values (GEBVs) of bulls using DNA pooled genotypes of cows as the reference

Genomic estimated breeding values (GEBVs) of PLS of progeny for individually genotyped bulls were derived within each population. The conventional genomic prediction method was applied to derive GEBVs, that is, a mixed animal model was used by fitting a polygenic random effect with the GRM (genomic relationship matrix). The fixed effects included the size of pool (30 levels) and contemporary group (5 levels) for SmartF, and SNP chip row (3 levels) and column (24 levels) information for different pools in MDH2020, respectively. The GRM was constructed using the method described by Reverter et al (2016) for pooled genotypes. Then the Qxpak5 software program (Pérez-Enciso and Misztal, 2011) was used to fit the GRM in a mixed animal model and obtain genomic estimates of variance components and genomic predictions (GEBVs) for PLS of the testing population. For comparison purposes of different density panels within populations, GEBVs were derived using four GRMs, either with 19,089, 54,791 (for MDH2020 only), 74,584 (for SmartF only), or high density (HD) SNP.

3.6.2 Comparing genomic prediction for individual bull's contribution to their female progeny fertility using individual or pooled genotypes

DNA pooling of cows with reproductive records could provide a cost-effective way for assessing and predicting the contribution of individual bulls to the fertility of their female offspring. However, how do the genomic rankings of bulls using DNA-pooled genotypes compare with those using individual genotypes? To answer this question, we conducted the study.

3.6.2.1 Animals, genotypes and phenotypes

A total of 893 Brahman cattle samples from MDH genotyped with 54,791 SNPs (Neogen Australasia GGP TropBeef50K chip) were used for the study. The population consisted of:

- 114 individually genotyped 2-year-old maiden MDH bulls from 2019;
- 233 DNA pools representing 2,452 cows with 2nd joining PLS records (233pools, Table 5)
- 546 individually genotyped cows (546cows, Table 5), selected from 2,452 cows.

Table 5. Composition of 546 cows and 233 pools with PLS records

Pregnancy and lactation status	Code	Analysis Code	No individual genotyped	No. pools (no. animals /pool)
Dry and empty (not pregnant)	DNP	1	92	23 (10-12)
Wet and empty (not pregnant)	WNP	2	256	129 (7-12)
Dry and early pregnant	DEP	3	43	22 (9-10)
Dry and mid pregnant	DMP	4	23	23 (7-9)
Dry and late pregnant	DLP	5	71	24 (8)
Wet and early pregnant	WEP	6	61	12 (5-6)
Total			546	233

3.6.2.2 Genomic relationship matrices (GRM) and genomic prediction

Two SNP-density panels (50K and imputed HD) were compared. For each SNP panel, three GRMs were constructed. They are:

GRM1: using 114 bulls + 233 pools
GRM2: 114 bulls + 546 individual cows
GRM3: 114 bulls + 233 pools + 546 individual cows

A standard GBLUP model was applied to estimate GEBVs for 114 bulls. In the model, the scores (1-6, Table 5) of PLS were used as the phenotype, a GRM as a random effect and the first three principal components as fixed effects.

3.7Develop a methodology for integrating GEBV for commercial cow fertility into Brahman group BREEDPLAN services

While other categorical score traits in beef cattle are amenable to linear models (for instance studies on body condition score and sheath score), from the study described in section 3.6.1, we encountered a few technical challenges for genomic prediction of a phenotype like PLS with six arbitrary categories. These include 1) the phenotype of PLS is a non-ordinal multi-class categorical trait, the conventional way of analysing an ordered categorical trait, e.g. treating PLS as a continuous score trait (i.e. 1-6) and applying the GBLUP model to generate breeding values, the results are difficult to make biological sense. It produces a single GEBV value for each candidate bull, it is meaningful to use it for ranking candidate bulls for selection, but it is unclear which category of PLS future progeny would fall into; 2) genomic ranking of bulls using GEBVs changes significantly on how the category Wet and Non-pregnant is scored (whether 2 or 5) in 1-6 scoring system; 3) the distribution of PLS normally departures from normal symmetry (e.g. PLS 3 and 4 not necessarily more abundant than 1 or 6) and the uncertainty about importance ranking of categories (e.g. PLS 2 not necessarily much worse than PLS 5) makes the analytical methodology to explore PLS rather cumbersome; 4) Conventional threshold models (treating the multiple class problem as binary classes) do tend to misclassify candidate bulls to the category with the most abundant category. For these reasons, there is a need to develop a new way of genomically ranking commercial bulls for commercial herd application of genomic selection, which is independent of the 1-6 phenotypic scoring system and biologically more meaningful to interpret.

3.7.1 New method - Genomic contributions to a Categorical Trait (GA2CAT)

3.7.1.1 Basic concept

Using the genotypes from individual bulls and a reference population of cows with genotypes and phenotypes for PLS, a genomic relationship matrix (GRM) can be computed using the method of Van Raden et al (2008). The GA2CAT of a testing bull is defined for each PLS category as its average genomic relationship with the cows having that PLS category divided by its average genomic relationship across all cows. Numerically, for the i-th bull, its GA2CAT to the j-th PLS category (j=1, 2, …6, corresponding to DNP, WNP, … WE, respectively) is computed as follows:

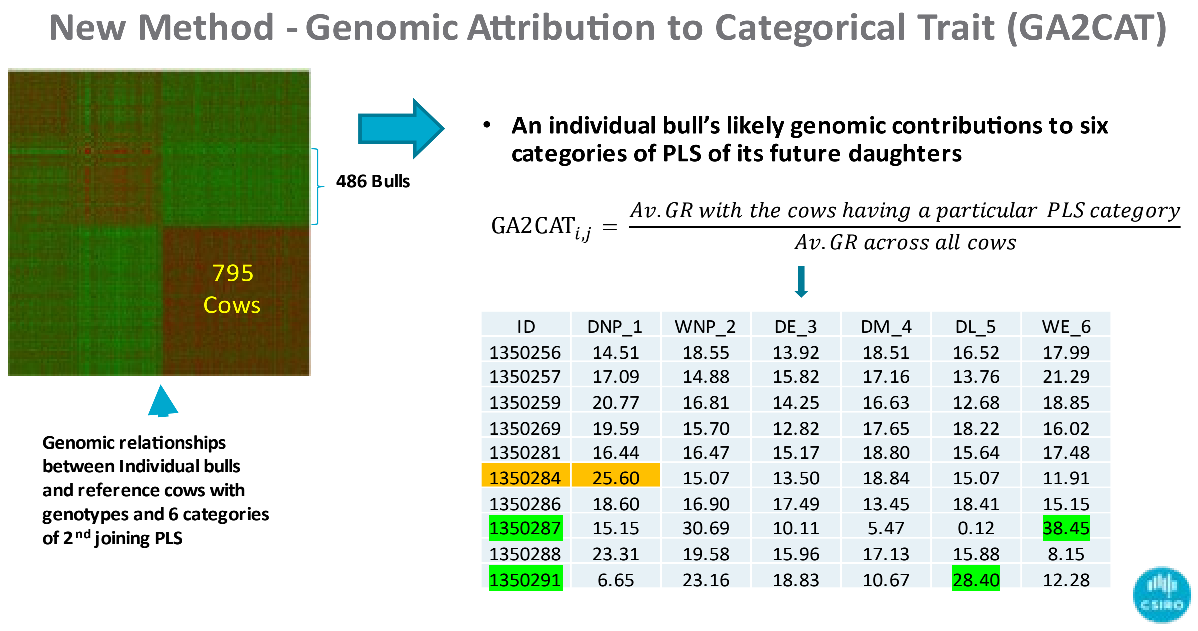
$$\mathrm{GA2CAT}_{i,j} = \frac{\frac{1}{N_j}\sum_{k=1}^{k=N_C}\left[g_{i,k}\times \mathrm{I}(k=j)\right]}{\sum_{j=1}^{j=6}\left(\frac{1}{N_j}\sum_{k=1}^{k=N_C}\left[g_{i,k}\times \mathrm{I}(k=j)\right]\right)},$$

where: N_j = Number of cows in the *j*-th PLS category;
N_C = Total number of cows in the reference population;
$g_{i,k}$ = Genomic relationship between the *i*-th animal and the *k*-th cow;
I(*k=j*) = An indicator function which takes a value of 1 if the k-th cow belongs to the j-th PLS category, and 0 otherwise.

By construction, the GA2CAT of the i-th animal summed across all 6 *j*-th PLS categories is one.

This method estimates an individual bull's contribution to six categories of their female progeny's pregnancy and lactation status at the second joining, based on the genomic relationships between the bulls and cows with PLS records. In other words, the estimates depend on how closely related each DNA-tested bull to the most fertile cows in a herd with PLS records. Figure 2 illustrates the process of generating GA2CAT values for 486 MDH bulls using the genotypes of 795 cows as the reference genome.

Figure 2. Schematic illustration of calculating GA2CAT values for MDH bulls and selection of the bulls with favourable (green highlight) or unfavourable (orange highlight) PLS phenotype



New Method - Genomic Attribution to Categorical Trait (GA2CAT)

- An individual bull's likely genomic contributions to six categories of PLS of its future daughters

$$\mathrm{GA2CAT}_{i,j} = \frac{Av.GR\ with\ the\ cows\ having\ a\ particular\ \ PLS\ category}{Av.GR\ across\ all\ cows}$$

Genomic relationships between Individual bulls and reference cows with genotypes and 6 categories of 2nd joining PLS

ID	DNP_1	WNP_2	DE_3	DM_4	DL_5	WE_6
1350256	14.51	18.55	13.92	18.51	16.52	17.99
1350257	17.09	14.88	15.82	17.16	13.76	21.29
1350259	20.77	16.81	14.25	16.63	12.68	18.85
1350269	19.59	15.70	12.82	17.65	18.22	16.02
1350281	16.44	16.47	15.17	18.80	15.64	17.48
1350284	25.60	15.07	13.50	18.84	15.07	11.91
1350286	18.60	16.90	17.49	13.45	18.41	15.15
1350287	15.15	30.69	10.11	5.47	0.12	38.45
1350288	23.31	19.58	15.96	17.13	15.88	8.15
1350291	6.65	23.16	18.83	10.67	28.40	12.28

To explore the value of GA2CAT in the context of GEBV methodology currently used in the industry, we carried out a series of analyses on different populations using both conventional GBLUP models and the newly developed GA2CAT method to compare the animal rankings. These include: 1) Testing GA2CAT method in bull selections for MDH populations; 2) Validating the utility of GA2CAT method by comparing bull rankings using the GA2CAT estimates versus those using the GEBV of bull reproductive traits from the MLA Bull Fertility Project (L.GEN.1818); 3) Comparing cow fertility rankings using the GA2CAT estimates versus those using BREEDPLAN GEBV from the Repronomics project (B.NBP.0759); 4) Comparing genomic prediction accuracies for cows' reproductive performance using GA2CAT and two machine learning methods.

3.7.2 Genomic prediction of bulls' future progeny performance for MDH populations using GA2CAT

3.7.2.1 MDH animals

Four tropical Brahman cattle populations from MDH were used for the study. The first three populations contained 829 bulls, of which 114 were from the 2020 season (Bulls_114), 229 from 2021 (Bulls_229), and 486 are to be used in the 2022 mating season (Bulls_486). The fourth population consisted of 795 cow samples (Cows_795, which is the combination of two cow populations Cows_546 and Cows_249, Table 2) with PLS records. All samples were individually genotyped with 54,791 SNPs (Neogen Australasia GGP TropBeef 50K chip) and later imputed to a higher density of 529,260 autosomal SNPs. A GRM across all animals was constructed, and the GA2CAT values of individual bulls were then estimated based on their genomic relationships with the 795 cows. For all bull populations, genomic prediction was performed using the new method GA2CAT to estimate individual bull's contributions to their progeny's PLS (six categories). For comparison purposes, genomic estimated breeding values (GEBV) of PLS of progeny for individually genotyped bulls were also derived within each individual population. The conventional genomic prediction method - GBLUP linear model was applied to derive GEBV, that is, a mixed animal model was used by fitting a polygenic random effect with the GRM. Genomic estimated breeding values (GEBV) of PLS of progeny for individually genotyped bulls were derived within each individual population.

3.7.3 Validation of New GA2CAT Method as "genomic prediction machinery"- Comparison of bull rankings using GA2CAT and GEBV for MLA Bull Fertility Project datasets

To further examine the utility of new method (GA2CAT) in predicting individual bull ranking based on their contributions to PLS, we applied the new method to 6,063 bulls from the MLA's Bull Fertility Project (L.GEN.1818). Of these animals, 1,051 were CRC Brahman (CRCBR), 1,819 were CRC Tropical Composite (CRCTC), 929 were Santa Gertrudis (SGT), 760 Droughtmaster (DMT), 844 Ultrablack (BLK), 660 Belmont Red (BEL). All animals were genotyped with 54,791 SNPs and then imputed to a high-density panel. Individual bulls' contribution to six categories of PLS was estimated using the GRM constructed from the 522,549 SNP genotypes (common SNP) of bulls within each breed and 1,028 MDH cows with the phenotypes of PLS. The 1,028 MDH cows consisted of 546 individual cows (2020), 233 DNA pooled (2020), and 246 individual cows. All bulls had GEBV calculated by the research project for 10 traits (WT (liveweight), CS (condition score), SC (scrotal circumference), sheath (sheath score), dens (sperm density), mass (sperm mass activity), mott (sperm motility), PNS (percent normal sperm), PD (proximal cytoplasmic droplets) and MP (midpiece abnormality)). For comparison purposes, a conventional GBLUP model with six categories of PLS as a continuous trait was used to calculate GEBV of PLS values.

3.7.4 Comparison between Brahman commercial cow fertility GA2CAT, GEBV and BREEDBPLAN DTC EBV Values

This study was undertaken to validate the GA2CAT methodology using the Repronomics cow fertility datasets, and to compare Brahman GA2CAT GEBV with Brahman BREEDBPLAN Days to Calving (DTC) EBV.

Using a similar approach to that used for the Bull Fertility datasets, we conducted the following analyses: 1) Constructing a GRM with the high-density genotypes of 2,118 Repronomics cows and 1,028 MDH cows; 2) Estimating GEBV of PLS for 2,118 cows using the conventional GBLUP model; 3) Estimating GA2CAT values for 2,118 cows using the new method; 4) Comparing the correlations between GEBV of PLS, GA2CAT values and GEBV of BREEDPLAN female fertility trait - Days to Calving (DTC). Table 11 summarises the Pearson's correlations between different metrics - six categories of GA2CAT values, GEBV of PLS (PLS_GEBV) and GEBV of DTC (DTC_GEBV).

3.7.5 Cross-validation estimates of empirical accuracy of GA2CAT values in different populations

The accuracy of the genomic predictions of assigning samples to phenotype categories using GA2CAT was assessed empirically in four distinct cross-validation studies.

3.7.5.1 Assigning 546 individually genotyped cows to their DNA pools of origin across 233 pools that were formed based on the pregnancy and lactation status categories

Based on numerical approaches previously described by Reverter et al. (2016) a 'hybrid' GRM was constructed with individual DNA samples from 546 MDH cows and pooled DNA samples from 233 pools. Importantly, these 546 individually genotyped cows were sampled in a such way that all 233 pools were represented. A cow was assigned to the pool in which its genomic relationship was the largest.

3.7.5.2 Assigning 2,107 bulls to their nominal breed across 17 breeds (i.e., Breed as a non-ordinal multi-class phenotype)

For this component of the empirical accuracy study, we resorted to the dataset employed by Reverter et al. (2020) in the context of genomic breed composition, and in which, using resources from the 1000 Bull Genomes Project, a dataset of 2,107 animals from 17 breeds and with genotypes for 1,001,234 SNPs was assembled.

The breeds (number of bulls in brackets) included were as follows: Angus (257), Blonde d'Aquitaine (36), Brahman (146), Brown Swiss (117), Charolais (127), Composite (57), Crossbreed (34), Gelbvieh (53), Hereford (74), Holstein (681), Jersey (88), Kholmogory (20), Limousin (82), Montbeliarde (51), Normande (43), Original Braunvieh (46), and Simmental (195).

Within the context of the current project, we built a GRM across the 2,107 animals and tested GA2CAT by sequentially exploring the genomic relationship of each animal to all the other ones and assigning it to the breed for which its average relationship was the largest.

3.7.5.3 Assigning 795 individual cows to pregnancy and lactation status with the approach of leaving one out at a time

Similar to the approach just described using data from the 1000 Bull Genomes Project, the accuracy of GA2CAT was assessed using 795 MDH cows with pregnancy test phenotypes and with six possible categories (number of cows in brackets) as follows: 1. Dry not pregnant (DNP, N = 124 cows); 2. Wet not pregnant (WNP, N = 358 cows); 3. Dry and early pregnant (DE, N = 77 cows); 4. Dry and mid pregnant (DM, N = 70 cows); 5. Dry and late pregnant (DL, N = 86 cows); and 6. Wet and pregnant (WP, N = 80 cows).

In detail, the accuracy validation approach proceeded as follows:

1. After imputing genotypes for 535,509 SNPs, a GRM was built for the 795 Cows.
2. Within each pregnancy test phenotype, the average genomic relationship between each cow and the cows with that phenotype was recorded (ignoring its self-relationship).
3. Each cow was assigned to the pregnancy test phenotype with the largest average genomic relationship
4. The assigned phenotype was compared against the real phenotype of each cow.

3.7.5.4 Assigning 1,451 MDH bulls to pregnancy and lactation status phenotype category

Assigning 1,451 bulls to pregnancy and lactation status phenotype category based on:

A. Confirming proportions in bulls deviate from random (i.e. proportions in the reference population of 795 cows).
B. Confirming the assignment of the last cohort of 622 bulls does not vary depending on which other bulls are included in the GRM (genomic relationship matrix).

3.7.6 Comparison of genomic prediction accuracies for cows' reproductive performance using GA2CAT and machine learning methods

Classification of the non-ordinal and highly imbalanced phenotype (such as PLS) presents a technical challenge when trying to rank potential sires based on their genomic relationships with phenotyped heifers. To address this issue, we developed a new method GA2CAT to predict an individual sire's contribution to its future daughters' performance. However, the performance of GA2CAT has not been benchmarked against other commonly used methods such as machine learning (ML) for analysing non-ordinal multi-class traits. Therefore, we conducted the study to compare genomic prediction accuracies of GA2CAT and two ML methods (Random Forests (RF) and Supporting Vector Machines (SVM)).

3.7.5.5 Datasets

Two datasets containing 1,135 tropical Brahman cows, 795 cows from 2020-2021 seasons (referred as Cows_795) and 340 of the 2022 season (Cows_340), from MDH were used for the study. All animals with PLS records were individually genotyped for 54,791 SNPs (Neogen Australasia GGP TropBeef 50K chip) which were then imputed to the high density using 700K genotypes of 861 legacy Beef CRC Brahman cattle as the reference genome. Table 6 summarises the composition of the phenotype records in both populations, illustrating unevenly distributed multi-class categories.

For comparison purposes, three different phenotype recoding systems for PLS records were investigated (Table 6). These include: a) treating PLS as a binary trait (2PLS, Non-pregnant "1" vs pregnant "2"); b) as a four-category trait (4PLS, Dry and Non-Pregnant "1", Wet and Non-Pregnant "2", Dry and Pregnant "3", and Wet and Pregnant "4"); and c) as a six- category trait (6PLS, see Table 1 for details).

Table 6. Composition of 2nd Joining Pregnancy and Lactation Status (PLS) records of two Brahman cow populations (795 and 340 cows respectively) and three phenotype recoding systems

		Cow population		Phenotype recoding system		
PLS	Code	Cows_795	Cows_340	2PLS*	4PLS*	6PLS*
Dry and Non-	DNP	124	61	1	1	1
Wet and Non-	WNP	358	109	1	2	2
Dry and Early	DEP	77	109	2	3	3
Dry and Mid	DMP	70	45	2	3	4
Dry and Late	DLP	86	6	2	3	5
Wet and Early	WEP	80	10	2	4	6

Total	795	340

*2PLS: binary categories, 4PLS: four categories; 6PLS: 6 categories

3.7.6.1 Statistical methods

Three analytical methods were used for evaluating classification accuracy, including GA2CAT (Li et al. 2022), RF (Random Forests, Berriman, 2001) and SVM (Support Vector Machine, James *et al.* 2013). In brief, the GA2CAT algorithm applies a standard genomic relationship matrix derived with the method of VanRaden (2008) between the reference and testing populations to predict the likely contributions of an individual animal in the testing population to individual classes of a categorical trait. For PLS, a GA2CAT value of a given animal for a given PLS category is defined as the animal's average genomic relationship with other animals having that PLS category divided by its average genomic relationship across all animals. RF is based on ensemble learning of a large number of decision trees deriving from randomly sampling of various subsets (both SNPs and animals) of a given dataset. It takes the average of decision trees (with replacement) to improve the predicted accuracy of the dataset. The final output (variable importance value) of RF is based on the majority votes of predictions. SVM applies different kernel functions (linear or non-linear) to identify a hyperplane that maximizes the separation of the data points to their potential classes (binary or multi-classes). While a genomic relationship matrix was used for deriving the GA2CAT values, both RF and SVM directly applied SNPs for the analyses.

A 5-fold cross-validation scheme was used for evaluating the classification performance of each method. Each cow population was randomly divided into 5 equal-size groups and each group (68 in Cows_340 or 159 animals in Cows_795) was in turn used as the validation set. Overall accuracy ((true positive +true negative)/(true positive + true Negative + false positive + false negative)) was used for evaluating the prediction performance. The final results were based on the average prediction accuracy of five validation groups. Given the imbalanced multiclass datasets used here, we also applied the Matthews correlation coefficient (MCC, Chicco and Jurman 2020) as a measure of the quality for multiclass classification. MCC values normally range from -1 to 1, with 1 representing a perfect prediction, 0 an average random prediction, and -1 a perfect misprediction.

A range of hyper-parameter values was examined for each ML method to determine the critical parameters that minimize prediction errors. These include: for RF, the size of forest trees (Ntree =100, 500), and the number of SNP markers at each sampling event (Mtry = 100, 500, 1000 and 5000); for SVM, insensitivity zone (gamma = 0.001, 1, 5, 10) and the penalty parameter (C= 0.001, 1, 10). All other parameters for each method took default values. The RF and SVM classifiers in the scikit-learn Python package (https://scikit-learn.org/stable/) were used for classification predictions.

3.8Calculate the cost-benefit of adopting the technology

The team investigated the suitability of economic models to estimate the economic impact of implementing our genomic predictions. The technology intervention being evaluated is most relevant for a beef breeding business in Northern Australia, based on Brahman cows. This type of enterprise is typically found in northern Australia, particularly northern and northwest Queensland. Three modelling programs were investigated and all of them could do the job but require experienced modelers to run the programs due to complexity, including CSIRO-developed CLEM (https://www.apsim.info/clem/Content/Home.htm), DynMod (https://livtools.cirad.fr/dynmod, France) and Breedcow+ ((https://breedcowdynama.com.au, QDPI). Of the three, we chose the Breedcow+ because it is already tailored to beef enterprises in Northern Australia. This software allows users to compare the profitability of different herd management strategies and includes regional herd templates for beef breeding regions in Queensland. These templates take achievable growth rates, mortality rates and other regional parameters of Queensland beef businesses into account. The Breedcow+ software reports the profitability of the modelled herds as gross margin per adult equivalent and assumes a steady-state herd structure. A key concept underpinning Breedcow+ analyses is that of adult equivalents. The calculation of the total adult equivalents for each modelled herd structure indicates the relative grazing pressure exerted by the herd structure and, if herds have similar total adult equivalents, a meaningful comparison of relative profitability can be made. Reproductive parameters can be changed in Breedcow+ and the outcomes on herd profitability modelled. We decided to adopt this online tool to model a north Queensland beef enterprise before our intervention and 10 years after introducing GA2CAT.

4. Results

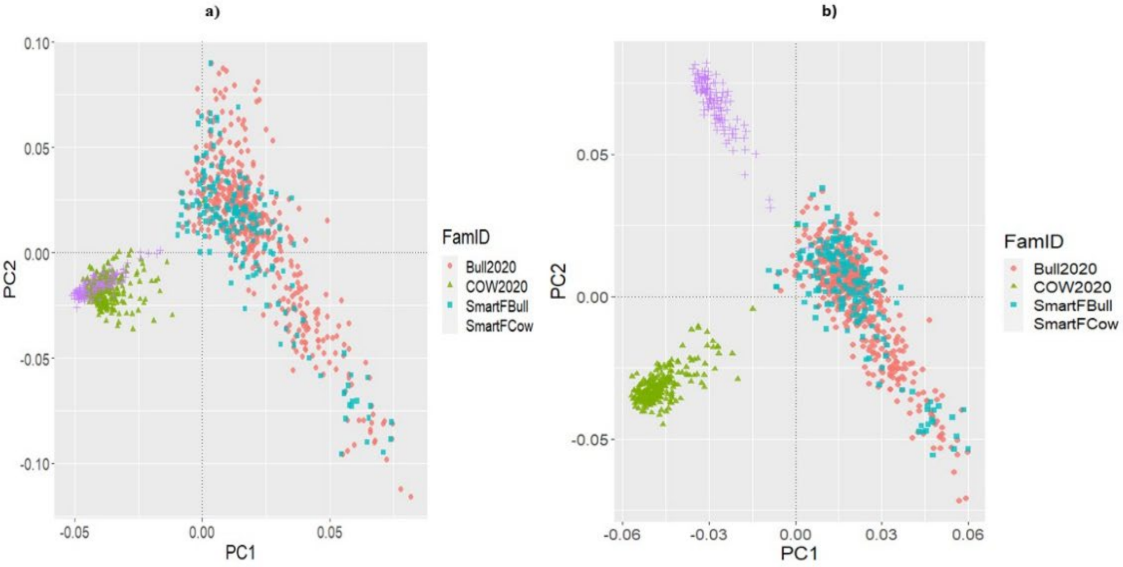
4.1 Determine the accuracy of pooled-DNA GEBV for commercial cow fertility within participating herds and for the Brahman breed more generally

4.1.1 The impact of SNP panels of different densities on rankings of bulls using DNA pooling and conventional GBLUP models

4.1.1.1 Relationships between animals of two populations

The results from the PCA on all 1,005 animals (290 from SmartF and 715 from MDH2020) are shown in Fig 3. When a low-density SNP panel data (19,089, Fig. 3a) was used, 346 DNA pooled cow samples from both populations were clustered together with very small variation among them, suggesting high similarity in the number of alleles between pooled samples. For the 659 individually genotyped bulls (red and blue dots), there was a much wider range of variation than for cows. However, when the high-density SNP panel was applied (HD, Fig. 3b), there was a clear separation of cow samples of within and across two populations. But bulls remained mixed up as low-density results show, with a much narrower range of variation. This indicates that the bulls in the two populations had some degree of relatedness among themselves, but not among the cows. Therefore, the separation of pooled cows would not have been detected if the HD was not used.

Figure 3. Principal Component Analysis of 1,005 genotyped samples, in which 482 were individually genotyped bulls (Bull2020), 233 were pools of cow DNA samples (COW2020), 177 were individually genotyped SmartF bulls (SmartFBull) and 113 were pools of SmartF cows (SmartFCow). a) 19,089 common SNP; b) High density SNP



4.1.1.2 Genomic predictions of PLS in bulls – using different SNP density panels

Assuming the results from HD are true, Table 7 presents the estimates of genetic and phenotypic parameters from the GBLUP models using three different SNP densities. Across both populations, HD panel produced higher genetic variance and heritability estimates.

Table 7. Heritability and variance estimates from using 19,089, 54,791 and HD (615,310) SNP panels within MDH2020 and SmartF populations, respectively.

Populations	GRM	SNP Density	σ_a^2	σ_e^2	h^2
MDH	482 Bulls + 233 Pools	19,089	0.3481	0.7531	0.3161
		54,791	0.4726	0.7105	0.3990
		HD	0.6818	0.4157	0.6200
SmartF	117 bulls +113 pools	19,089	0.746	1.335	0.3585
		74584	0.7467	1.3028	0.3643
		HD	0.7389	1.0867	0.4048

Table 8 shows the Pearson's correlations among the PLS GEBVs from three SNP panels (19,089, 54,791 and HD) in the MDH2020 and SmartF respectively. Within MDH202, the correlations between GEBVs of PLS of 482 bulls were 0.74 between 19,089 and HD, and 0.82 between 54,791 and HD. The correlations were much lower (0.39 and 0.45 respectively) if only the top 25% bulls were considered (see Table 8 correlation for top 25%). Similar trends were observed in SmartF when the correlations of GEBVs for 177 bulls were compared (Table 8), despite slightly higher correlations between 19,089 and 74,584 with HD when the top 25% bulls were selected (0.54-0.59, Table 8). These suggest that if low-density panels were used to genotype pooled DNA cows for estimating the EBVs of PLS of bulls, at least 40-50% of the best bulls would not be selected.

Table 8. Pearson's correlations among GEBVs estimated from using 19,089, 54,791 and HD SNP panels within MDH2020 and SmartF populations, respectively.

Population		MDH2020			SmartF		
	SNP	19089	54791	HD	19089	74584	HD
All bulls	19089	1	0.90	0.74	1	0.76	0.72
	54791 / 74584		1	0.82		1	0.81
	HD			1			1
Top 25%	19089	1	0.81	0.39	1	0.52	0.54
	54791 / 74584		1	0.45		1	0.59
	HD			1			1

When further investigating the bull GEBVs of PLS estimated using HD, Table 9 illustrates the profiles of the GEBVs of 482 MDH2020 bulls and 177 SmartF bulls in different quartiles. The average GEBV difference between top and bottom 25% of 482 MDH2020 bulls was 0.292**,** which is much larger than the difference obtained using low density panels (0.120 from 19,089 or 0.158 from 54,791, see the Supplementary Tables 1 and 2 for details). To put this in perspective, the difference between animals being dry and empty (PLS score 1) and wet and pregnant (score 6), represents 21-27 months of a cow's productive life. The GEBV difference of 0.292 from HD would translate into earlier conception by 1.31 months for the female progeny of the top 25% sires. A similar trend was observed with the 177 SmartF bulls, in which the average GEBV difference between top and bottom 25% of the bulls was 0.286**,** which is also much larger than the difference obtained using lower density panels (0.0561 from 19,089 or 0.136 from 74,584 SNP, Supplementary Tables 1 and 2).

Table 9. Average genomic breeding values (GEBVs) of progeny pregnancy and lactation status (PLS) of the MDH2020 bulls in four quartiles using HD SNP panel

	MDH2020			
Quartile	# Bulls	Av. GEBV	Min	Max
1 -Top 25%	120	0.136	0.0833	0.323
2	120	0.055	0.0275	0.0831
3	121	-0.004	-0.0341	0.0261
4 – Bot. 25%	121	-0.156	-0.2771	-0.0345
All	482	0.023	-0.277	0.323

	SmartF			
Quartile	# Bulls	Av. GEBV	Min	Max
1 -Top 25%	44	0.0746	0.0208	0.2521

2	44	-0.0217	-0.0618	0.0175
3	44	-0.0967	-0.1339	-0.0618
4 – Bot. 25%	45	-0.2112	-0.3402	-0.1365
All	177	-0.0646	-0.3402	0.2521

The study presents the results for the comparison of different panels of SNP density in ranking commercial bulls in two populations. The phenotype score (1-6) of the 2nd joining pregnancy and lactation status was treated as a continuous trait. It highlights the need for extreme caution to be taken when applying SNP panels of low or medium densities to study genetic relationships, and rank and select top bulls for commercial beef production based on DNA pooling technology.

4.1.2 Comparing genomic prediction of bulls' contribution to their female progeny fertility using individual cow genotypes versus DNA-pooled cow genotypes

4.1.2.1 Comparison of genetic and phenotypic parameters from GBLUP models of two different SNP densities and two different genotype sources

Table 10 shows that the genomic prediction of GEBVs for 114 bulls using 50K genotypes and PLS records of 233 pools produced much larger genetic variance and a very small error variance, hence a very high heritability estimate (0.99), in comparison to using HD (0.68) . This was a stark contrast to a consistent low h^2 estimate (HD: 0.054; 50K: 0.059) when using the genotypes of 546 individual cows as the reference genome.

Table 10. Comparison of estimates of variances and genetic parameters

GRM Source	SNP density	No. Samples	σ_a^2	σ_e^2	h^2
114 + 233 pools	50K	347	1.1	0.015	0.99
114 + 546 Cows	50K	660	0.60	9.59	0.059
114 + combined	50K	893	0.74	0.12	0.87
114 + 233 pools	529,260	347	0.30	0.14	0.68
114 + 546 cows	529,260	660	0.55	9.62	0.054
114 + combined	529,260	893	0.42	0.078	0.84

4.1.2.2 Correlations between GEBVs of PLS when using different SNP densities and GRM sources

Table 11 summarises the genetic correlations between GEBVs of different methods. It can be seen that:

- When using the same GRM source, e.g., 114 bulls plus 233 pools, regardless the size of SNP-panel (50K or HD), the correlation was high (0.791, 50K_233pools vs HD_233pools, Table 11). There was an almost perfect correlation between 50K and HD when individual genotypes of 546 cows were used as the reference genome (0.969, 50K_546cows and HD_546cows).

- There was very little genetic correlation between the GEBVs estimated using DNA pools and those of using the individual genotypes of 546 cows (0,062, 50K_233pools and 50K_546cows; 0.070, HD_233pools and HD_546cows). The results suggest that there were large allele frequency variations between pools and 546 individuals.

The results indicate that for a low heritability trait such as PLS, due to a very low correlation between GEBVs of using DNA-pooling and individual genotyping methods, genomic prediction using DNA pooling should be combined with individual genotypes that represent different phenotype pools.

Table 11. Pair-wise Pearson's correlations between GEBVs of PLS for 114 bulls using two different SNP densities (HD vs 50K) and prediction methods (233 pools vs 546 cows vs combined)

	HD_233pools	**HD_546Cows**	**HDcombined**	**50K_233pools**	**50K_546cows**	**50K_combined**
HD_233pools	1	**0.070**	0.906	0.791	0.014	0.715
HD_546Cows		1	-0.08	**0.038**	0.969	0.119
HD_combined			1	0.669	-0.054	**0.739**
50K_233pools				1	**0.062**	0.878
50K_546cows					1	0.167
50K_combined						1

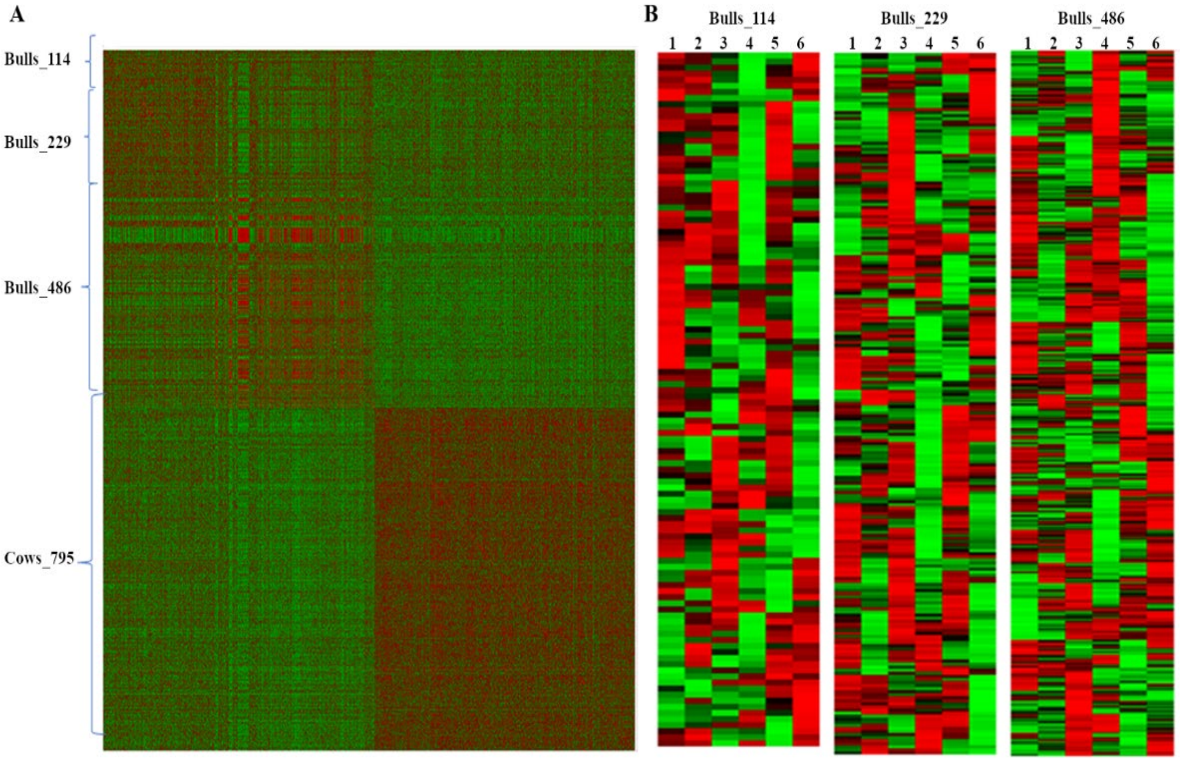
4.2 Develop methodology for integrating GEBV for commercial cow fertility into Brahman group BREEDPLAN services

4.2.1 Genomic prediction using the newly developed GA2CAT method on MDH bulls

4.2.1.1 Genomic relationships between bull and cow populations

The intensity differences in Fig. 4A reveals that: 1) The animals in the Cows_795 set were more closely related among themselves than with the 829 bulls from three bull populations; 2) The two bull populations (Bulls_114 and Bulls_229) had a closer relationship than either of them with the third bull population (Bulls_486); 3) The relationship between bulls and the cow population (Cows_795) were stronger for the Bulls_114 and Bulls_229 than for the Bulls_486.

Figure 4. (A). Pairwise genomic relationships between animals of different populations. Each off-diagonal dot represents a genomic relationship matrix element between two animals. The red intensity represents closeness between animals. (B). Heatmap of individual bulls' GA2CAT values across six categories within each bull population.



4.2.1.2 GA2CAT estimates for three bull populations

The closer genomic relationships between the first two bull populations (Bulls_114 or Bulls_229) and the 795 cows, allowed for large variations in average GA2CAT values (Table 12, see columns 2 and 3) in these two bull populations (ranging from 0.139 to 0.182, 0.139 to 0.194, respectively), in comparison to the narrow range of the bull population Bulls_486 (0.159 – 0.172). Accordingly, the standard deviation (STD) of the GA2CAT in the first two bull populations were also consistently larger than those in the third bull population. When clustering bulls based on their GA2CAT values (Figure 4B), it is easy to identify the bulls with distinguished value differences across the six categories. This allows, for example, selection of the desirable bulls with the highest GA2CAT values for 6_WEP (red colour) and the lowest values for 1_DNP (green colour).

Table 12. Average values of bull's genomic attribution to pregnancy and lactation status (GA2CAT) across the six PLS categories for three bull populations.

Population	**Bulls_114**	**Bulls_229**	**Bulls_486**
PLS Category*	Mean (STD)	Mean (STD)	Mean (STD)
1_DNP	0.182 (0.0572)	0.170 (0.0521)	0.172 (0.0463)
2_WNP	0.169 (0.0451)	0.161 (0.0436)	0.159 (0.0299)
3_DEP	0.176 (0.0567)	0.194 (0.0676)	0.178 (0.0499)
4_DMP	0.139 (0.0636)	0.139 (0.0608)	0.159 (0.0514)
5_DLP	0.178 (0.0594)	0.171 (0.0626)	0.170 (0.0461)
6_WEP	0.156 (0.0709)	0.165 (0.0675)	0.163 (0.0513)

*1_DNP = Dry and Not Pregnant; 2_WNP = Wet and Not Pregnant; 3._DEP = Dry and Early Pregnant; 4_DMP = Dry and Mid Pregnant; 5_DLP = Dry and Late Pregnant; 6_WEP = Wet and Early Pregnant.

Based on the individual GA2CAT values, the client and their bull breeder were able to prioritise bulls for retention in the bull breeding population and deployment as commercial herd bulls. An example of the estimated GA2CAT values for the top 10 bulls (ranked on the 6_WEP) from the population Bulls_229 is given in Table 13.

Table 13. GA2CAT estimates for the top 10 bulls among herd bulls available for selection in 2021

Rank	**AnimalID**	**1_DNP***	**2_WNP**	**3_DEP**	**4_DMP**	**5_DLP**	**6_WEP**
1	13040318	0.211	0.030	0.164	0.033	0.019	0.542
2	13040140	0.134	0.126	0.091	0.069	0.054	0.527
3	1324737	0.184	0.076	0.058	0.019	0.157	0.506
4	13040151	0.279	0.069	0.106	0.037	0.033	0.476
5	13040416	0.370	0.053	0.008	0.009	0.112	0.449
6	13040395	0.198	0.043	0.100	0.116	0.130	0.414
7	13040417	0.219	0.050	0.170	0.054	0.095	0.412
8	1324629	0.178	0.115	0.153	0.029	0.113	0.411
9	1324498	0.188	0.164	0.045	0.129	0.073	0.401
10	1324696	0.153	0.157	0.065	0.068	0.166	0.392

*1_DNP = Dry and Not Pregnant; 2_WNP = Wet and Not Pregnant; 3._DEP = Dry and Early Pregnant; 4_DMP = Dry and Mid Pregnant; 5_DLP = Dry and Late Pregnant; 6_WEP = Wet and Early Pregnant.

4.2.1.3 Correlations Between Bull's contributions to six categories of PLS and GEBV of PLS from using conventional GBLUP model

One question that remains unanswered is how the bull rankings based on the estimated GA2CAT values compare with those based on the conventional GBLUP model that treats the six categories of

PLS as a continuous trait. For this purpose, we conducted the GBLUP analyses to derive the GEBV for PLS in the three bull populations separately. From Table 14, the GA2CAT values for WNP (PLS category 2) had a consistent negative correlation with the GEBV across all three bull populations (see yellow highlights), while the GA2CAT for WEP (PLS category 6) had a strong positive correlation (>0.47, yellow highlights) with the GEBV in both Bulls_229 and Bulls_486.

Table 14. Pearson correlations between GA2CAT values for six PLS categories and GEBV in three bull populations

	GEBV within Population		
PLS Category	Bulls_114	Bulls_229	Bulls_486
1_DNP	0.04	-0.09	0.22
2_WNP	-0.45	-0.38	-0.69
3_DEP	-0.11	-0.09	0.05
4_DMP	-0.26	-0.44	-0.49
5_DLP	0.37	0.21	0.10
6_WEP	0.27	0.48	0.47

*1_DNP = Dry and Not Pregnant; 2_WNP = Wet and Not Pregnant; 3._DEP = Dry and Early Pregnant; 4_DMP = Dry and Mid Pregnant; 5_DLP = Dry and Late Pregnant; 6_WEP = Wet and Early Pregnant.

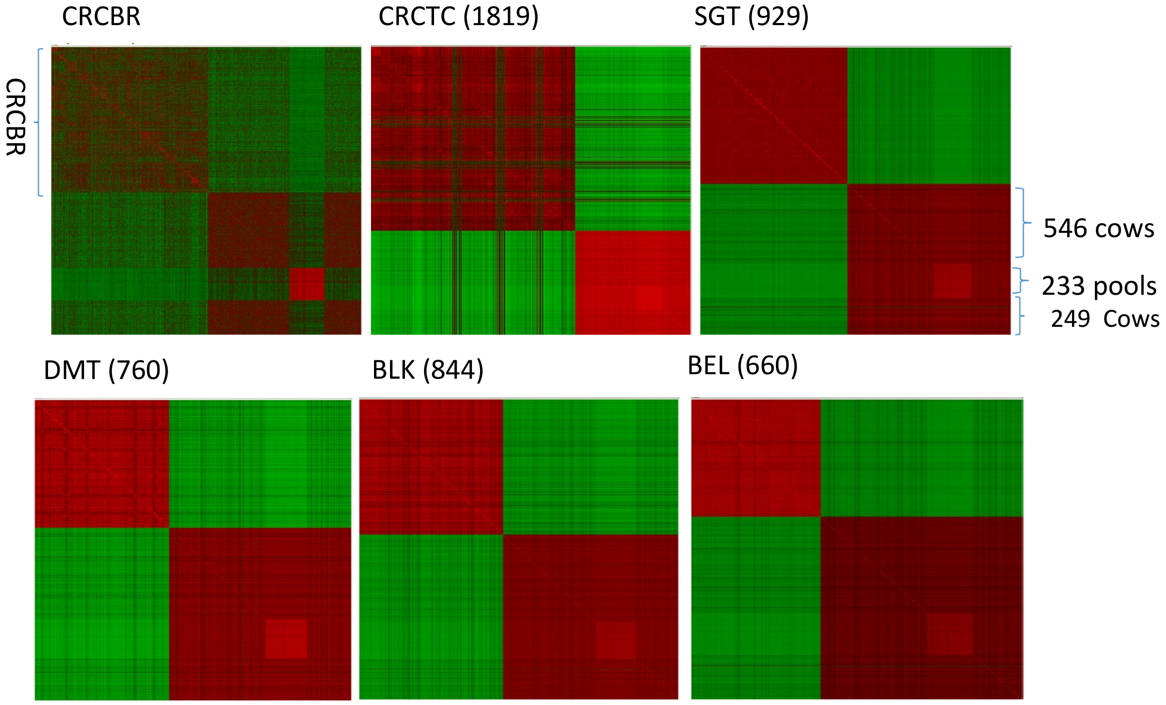
The newly developed method GA2CAT provides an effective way of assessing individual sire's genomic contributions to six categories of PLS by using the standard GRM values between sires and cows with PLS records. The merits of the method include: 1) It is biologically meaningful and easy for the commercial beef industry to adopt to select the bulls with desirable progeny reproductive performance, 2) the GA2CAT method will not be influenced by the phenotype scoring system; 3) The population variation of GA2CAT values will truly reflect the degree of genomic closeness between testing and reference populations, 4) The method could easily be extended to any phenotype with a categorical or ordinal nature.

4.2.2 Validation of New GA2CAT Method as "genomic prediction machinery"- Comparison of bull rankings using GA2CAT and GEBV for MLA Bull Fertility Project datasets

4.2.2.1 GRM relationships between bulls of six breeds and 1,028 MDH Brahman cows with PLS records

As expected, of 6 the breeds (CRCBR, CRCTC, SGT, DMT, BLK and BEL), only the CRCBR bulls had genomic relationships with the MDH Brahman cows (see Fig. 5 for CRCBR). To a lesser degree, CRCTC had some relatedness with the MDH cows. The bulls from the remaining four breeds had little relationship with the MDH cows.

Figure 5. Pairwise genomic relationships between bulls of each breed and MDH Brahman cow samples. Each off-diagonal dot represents a genomic relationship matrix element between two animals. The red intensity represents closeness between animals. CRCBR -CRC Brahman, CRCTC -CRC Tropical Composite, SGT- Santa Gertrudis, DMT- Droughtmaster, BLK -Ultrablack, BEL -Belmont Red.



4.2.2.2 GA2CAT estimates for six bull populations

Table 15 summarises the basic statistics of bulls' GA2CAT in six breeds for the PLS category wet and early pregnant (WEP_6) and the difference between WEP_6 and wet and not pregnant (WNP_2). Since there were little genomic relationships between the bulls in four breeds (SGT, DMT, BLK and BEL) and MDH Brahman cows that had PLS records, therefore, there was very little variation in the estimated bulls' GA2CAT values in these breeds for WE_6 and WEP6_minus_WNP2. However, the opposite was true for the CRCBR and CRCTC in which some genomic relationships were existing between the bulls in these populations with the cows in MDH population that were used for the genomic prediction. For example, the variation for WEP_6 (Table 15, top part) ranged from 16.58 (minimum) – 17.20 (maximum) in SGT, 16.71 – 17.07 in BEL, in comparison to the significant variations 2.90 -50.50 in CRCBR and 6.05 -35.60 in CRCTC. A similar trend was observed for the WEP6_minus_WNP2.

Table 15. Basic statistics of bulls' GA2CAT values of six breeds for WEP_6 and WEP6_minus_WNP2. CRCBR- CRC Brahman, CRCTC - CRC Tropical Composite, SGT- Santa Gertrudis, DMT-Droughtmaster, BLK -Ultrablack, BEL -Belmont Red.

WEP_6	CRCBR*	CRCTC	SGT	DMT	BLK	BEL
Min.	2.90	6.05	16.58	15.36	16.51	16.71
Mean	17.74	16.85	16.87	17.01	16.83	16.87
Max.	50.50	35.60	17.20	18.02	17.35	17.07
Std	3.81	0.66	0.09	0.28	0.07	0.05
WEP6_minus_WNP2						
Min.	-15.44	-15.28	0.15	-0.26	-0.11	0.29
Mean	1.93	0.40	0.47	0.83	0.39	0.48
Max.	49.82	19.44	0.82	2.19	0.84	0.73
Std	5.17	0.82	0.10	0.32	0.08	0.06

The range of GA2CAT values clearly reflected the close or weak genomic relationships between the two populations from which the GRM was constructed.

4.2.2.3 Correlations between GA2CAT values and GEBV of 10 Traits within the individual breed

From Table 16, there were very little correlations between the individual bull's contributions to six categories of PLS and the GEBV of Brahman bull fertility traits. This suggests that the bull rankings based on the new method would be different from those based on the GEBV of sperm quality traits.

Supplementary Tables 3 – 7 show that there were no correlations between these traits in the Tropical composite breeds studied. These could be due to: 1) The genetic distances between the animals in these populations and 1,028 MDH cows that were used as the reference population were far apart, therefore there was not enough power for genomic prediction; 2) Sperm related fertility traits had low heritability values and were largely impacted by animal physiology, environmental and management factors.

Table 16. Pearson's correlations between GA2CAT values for 6 categories of PLS and GEBV of 10 growth and fertility traits from bulls of CRCBR.

	wt	cs	sc	sheath	dens	mass	mott	pns	pd	mp	DNP_1	WNP_2	DE_3	DM_4	DL_5	WE_6
DNP_1	0	0.02	0.023	0.023	-0.03	-0.098	-0.014	-0.044	0.017	0.082	1	-0.081	-0.06	-0.19	-0.276	-0.299
WNP_2	0.005	0.036	-0.033	0.052	0.137	0.065	-0.042	-0.023	-0.017	-0.044	-0.081	1	0.082	0.039	-0.214	-0.447
DEP_3	0.17	-0.05	0.008	-0.087	0.088	0.038	-0.03	0.065	-0.076	-0.072	-0.06	0.082	1	-0.011	-0.38	-0.438
DMP_4	0.057	0.014	-0.083	0.044	0.044	-0.044	-0.064	-0.044	-0.005	0.006	-0.19	0.039	-0.011	1	-0.286	-0.395
DLP_5	-0.144	0.022	-0.027	0.006	-0.069	-0.017	0.045	-0.032	0.089	-0.004	-0.276	-0.214	-0.38	-0.286	1	0.092
WEP_6	-0.063	-0.025	0.076	-0.021	-0.102	0.05	0.07	0.053	-0.008	0.015	-0.299	-0.447	-0.438	-0.395	0.092	1

Abbreviations: wt: live weight; cs: condition score, sc: scrotal circumference; sheath: sheath score; dens: sperm density; mass: sperm mass mobility; mott: sperm motility; pns: percent normal sperm; proximal cytoplasmic droplets; mp : midpiece abnormalities. *DNP_1 = Dry and Not Pregnant; WNP_2 = Wet and Not Pregnant; DEP_3 = Dry and Early Pregnant; DMP_4 = Dry and Mid Pregnant; DLP_5 = Dry and Late Pregnant; WEP_6 = Wet and Early Pregnant.*

4.2.3 Comparison between Brahman commercial cow fertility GEBV and BREEDBPLAN DTC EBV

With the imputed HD genotype datasets from the 2,118 cows of the Repronomic fertility project and 1,028 MDH cow samples (795 individuals and 233 DNA pools), the comparison was undertaken to examine the genomic rankings using the GA2CAT estimates for PLS, the GEBV for PLS estimated with a conventional GBLUP model, and the Brahman BREEDBPLAN Days to Calving (DTC) EBV. Table 17 summarises Pearson's correlations between different metrics - six categories of GA2CAT values, GEBV of PLS (PLS_GEBV) and GEBV of DTC (DTC_GEBV).

Table 17. Pearson's correlations between six categories of GA2CAT values, GEBV of PLS (PLS_GEBV) and GEBV of DTC (DTC_GEBV) in 2,118 Repronomics animals

	WEP_6	PLS_GEBV	DTC_GEBV
DNP_1	-0.32	-0.03	0.11
WNP_2	-0.33	-0.27	0.07
DEP_3	-0.42	0.03	0.19
DMP_4	-0.22	-0.08	-0.02
DLP_5	-0.21	0.12	-0.03
WEP_6	1	0.1	-0.22
PLS_GEBV	0.1	1	0.26
DTC_GEBV	-0.22	0.26	1

DNP_1 = Dry and Not Pregnant; WNP_2 = Wet and Not Pregnant; DEP_3 = Dry and Early Pregnant; DMP_4 = Dry and Mid Pregnant; DLP_5 = Dry and Late Pregnant; WEP_6 = Wet and Early Pregnant

It can be seen that: 1) The GA2CAT values for the preferred fertility category Wet and Early Pregnant (WEP_6) had a negative correlation (-0.22) with the GEBV of DTC. This indicates that using the WEP_6 as a selection metric for ranking bulls will drive selection in the same direction as favorable DTC; 2) The GEBV of PLS had a positive correlation (0.26) with the GEBV of DTC.

4.2.4 Cross-validation estimates of empirical accuracy of GA2CAT values in different populations

4.2.4.1 Using GA2CAT to assign 546 MDH individual cows to their DNA pools of origin across 233 pools

Of 546 cows, 428 (or 78.4%) were correctly assigned to their original pool (Table 18). In addition, cows correctly assigned to their pools had a higher genomic relationship to their assigned pool than cows wrongly assigned (6.25% vs. 2.70%, P-value < 0.001).

Taking a nominal 2.5% and 5% for the minimal genomic relationship thresholds before making an assignment, we found that 85.9% of cows (or 354 out of 412) and 94.5% of cows (or 185 out of 195) were correctly assigned, for 2.5% and 5% minimum genomic relationship, respectively.

Table 18. Summary statistics for % genomic relationship (GR) for the 546 cows by correct or incorrect assignment to their pool of origin.

Assignment	N Cows	Mean GR	STD	Min.	Max.
Correct	428	6.24841	6.10772	0.103878	72.9935
Incorrect	118	2.70406	1.45772	0.380383	7.92351

However, pool size had the largest effect on the accuracy of assignment (Table 19) with large pools being more difficult to characterize. For instance, the 44 cows from pools of sizes 5, 6, or 7 were all correctly assigned. Meanwhile, only 64.8% of cows (or 166 out of 256) from pools of size 12 were correctly assigned.

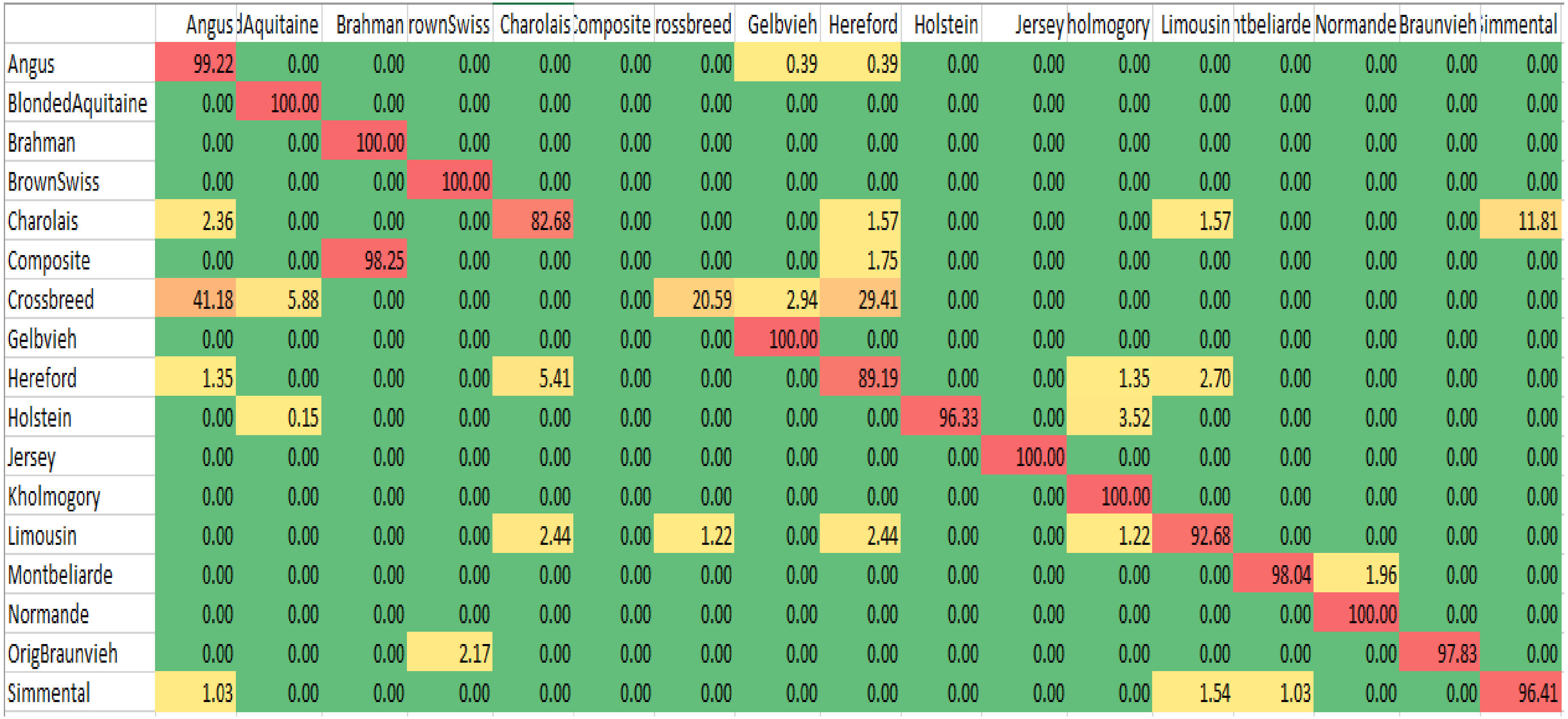
Table 19. Summary statistics for % genomic relationship (GR) for the 546 cows by correct or incorrect assignment to their pool of origin and by pool size.

Pool Size	Assignment	N Cows	Mean GR	STD	Min.	Max.
5	Correct	55	13.9076	6.41746	3.95133	30.3664
6	Correct	6	9.80834	1.77274	6.97329	12.2554
7	Correct	3	10.3752	1.86059	8.70149	12.3786
8	Correct	69	7.8335	8.90428	2.05687	72.9935
8	Incorrect	2	3.92746	0.772493	3.38123	4.4737
9	Correct	21	4.67785	1.93258	1.32736	8.68746
9	Incorrect	3	2.92528	2.45242	0.82855	5.62203
10	Correct	106	4.61489	2.23379	1.64901	13.5906
10	Incorrect	23	3.06334	1.46708	1.13303	6.51702
12	Correct	166	4.11513	4.31059	0.103878	49.0141
12	incorrect	90	2.57769	1.42646	0.380383	7.92351

4.2.4.2 Using GA2CAT to assign 2,107 bulls to their nominal breed across 17 breeds

We found that 92.6% (or 1,951 out of 2,107) of bulls were correctly assigned to their nominal breed. However, there was some variation in the accuracy of assignment across breeds (Fig. 6). For instance, while all 146 Brahman were assigned correctly, only 82.68% (or 105 out of 127) of Charolais were assigned to Charolais with 11.81% (or 15) being “incorrectly” assigned to Simmental. Worse performances were observed for loosely defined breeds such as “Composite” and “Crossbreds” but the difficulty in correctly assigning these breeds was also observed and discussed in the original work of Reverter et al. (2020).

Figure 6. Heatmap of the average % assignment for the 17 breeds. Within a row, averages across columns add to 100%.



	Angus	dAquitaine	Brahman	rownSwiss	Charolais	omposite	rossbreed	Gelbvieh	Hereford	Holstein	Jersey	holmogory	Limousin	ntbeliarde	Normande	Braunvieh	immental
Angus	99.22	0.00	0.00	0.00	0.00	0.00	0.00	0.39	0.39	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BlondedAquitaine	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Brahman	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BrownSwiss	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Charolais	2.36	0.00	0.00	0.00	82.68	0.00	0.00	0.00	1.57	0.00	0.00	0.00	1.57	0.00	0.00	0.00	11.81
Composite	0.00	0.00	98.25	0.00	0.00	0.00	0.00	0.00	1.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Crossbreed	41.18	5.88	0.00	0.00	0.00	0.00	20.59	2.94	29.41	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gelbvieh	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Hereford	1.35	0.00	0.00	0.00	5.41	0.00	0.00	0.00	89.19	0.00	0.00	1.35	2.70	0.00	0.00	0.00	0.00
Holstein	0.00	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	96.33	0.00	3.52	0.00	0.00	0.00	0.00	0.00
Jersey	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
Kholmogory	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00
Limousin	0.00	0.00	0.00	0.00	2.44	0.00	1.22	0.00	2.44	0.00	0.00	1.22	92.68	0.00	0.00	0.00	0.00
Montbeliarde	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	98.04	1.96	0.00	0.00
Normande	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00
OrigBraunvieh	0.00	0.00	0.00	2.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	97.83	0.00
Simmental	1.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.54	1.03	0.00	0.00	96.41

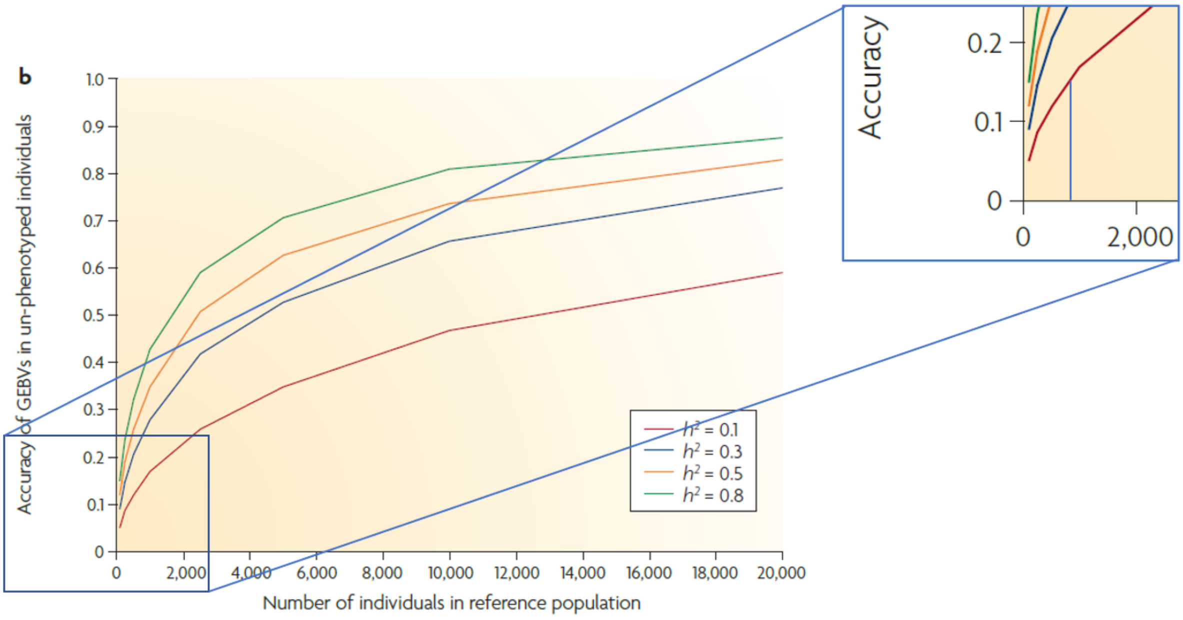
4.2.4.3 Using GA2CAT to assign 795 individual cows to pregnancy and lactation status with the approach of leaving one out at a time

We found that 12.7% of cows (or 101 out of 795) were assigned to their correct phenotype category. This seemingly low level of accuracy is in line with what could be expected from theory and based on a reference population of ~800 animals and a trait with a h^2 of ~10% (Fig. 7 and Goddard and Hayes, 2009).

Unlike the situation where we were assigning individual cows to their DNA pools of origin, the largest average genomic relationship used to assign cows to phenotype categories was not significantly affected by correct/incorrect assignment or phenotype category (P-value > 0.10). Across all 795 cows, this largest average genomic relationship averaged 28.3% (range 21.5% to 54.9%).

Importantly, the observed distribution of assignments being 121, 75, 118, 158, 151 and 172 cows into categories 1 to 6, respectively, was significantly different from the one expected by chance alone, ie. the distribution of the original phenotypes (Chi-squared with 5 df = 511.173, P-value < 0.0001). This translated into a % of correct assignment significantly different (P-value < 0.01) across the six phenotype categories: 16.9%, 8.6%, 10.4%, 18.6%, 20.9% and 12.5% for categories 1 to 6, respectively. Therefore, while only 8.6% (or 31 out of 358 cows) were correctly assigned to category 2 (Wet not pregnant), 20.9% (or 18 out of 86 cows) were correctly assigned to category 5 (Dry and late pregnant).

Figure 7. The expected accuracy of GEBVs of un-phenotyped individuals with an increasing number of individuals in the reference populations and various levels of heritability (h^2). The insert highlights the region corresponding to a reference population of ~800 animals where a trait with a h^2 of ~10% should allow for an accuracy of around 15%. Adapted from Figure 3 in Goddard and Hayes (2009).



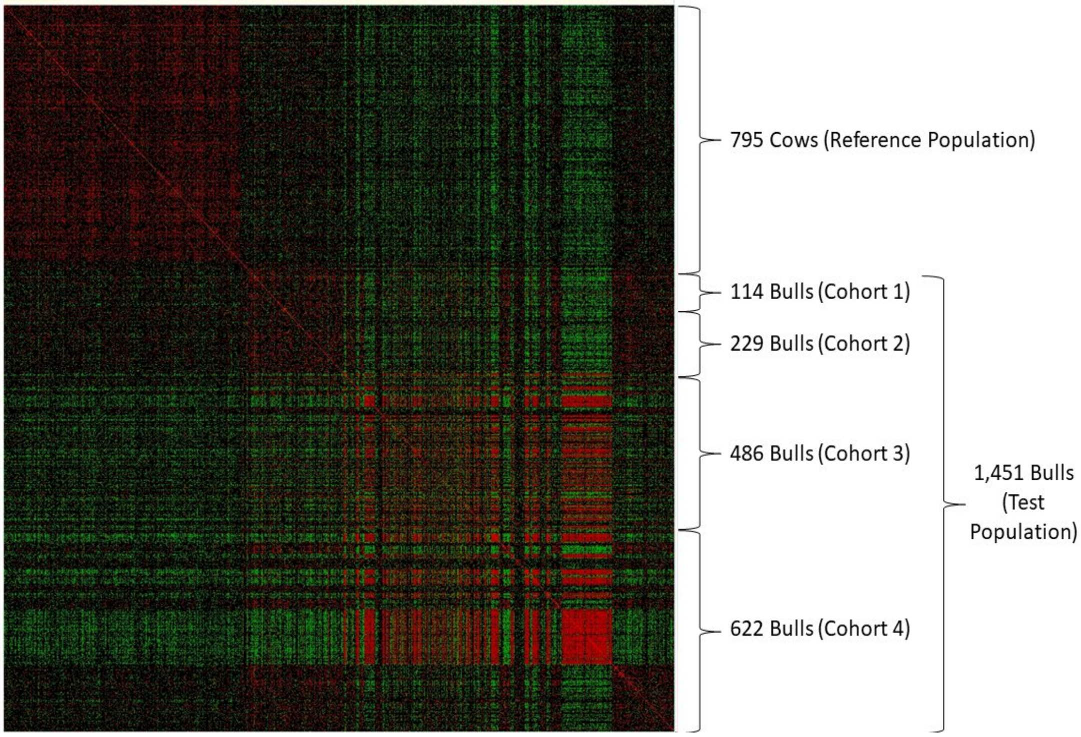
To further explore the accuracy of GA2CAT, we collapsed the six categories into four by grouping the dry and early, medium and late pregnant into a single “Dry and pregnant” category. Hence, the new four categories are as follows: 1. Dry not pregnant (DNP, N = 124 cows); 2. Wet not pregnant (WNP, N = 358 cows); 3. Dry and pregnant (DP, N = 233 cows); and 4. Wet and pregnant (WP, N = 80 cows).

Using this new coding system, 23.1% (or 184 out of 795) of cows were correctly assigned which compares favourably with the 12.7% observed using the 6 categories system. Again, the observed distribution of assignments (210, 153, 177 and 255 for categories 1 to 4, respectively) was significantly different from the one expected by chance alone, ie. the distribution of the original phenotypes (Chi-squared with 3 df = 573.305, P-value < 0.0001). Similarly, the % of correct assignments was significantly different (P-value < 0.01) across the four categories and equal to 30.6%, 18.45, 25.3% and 26.2% for categories 1 to 4, respectively.

4.2.4.4 Using GA2CAT to assign 1,451 MDH bulls to pregnancy and lactation status phenotype category

After imputing genotypes for 535,509 SNPs, a GRM was built for the 795 Cows (comprising the reference population) and 1,451 Bulls (validation or test population) from 4 cohorts (Fig 7). While all cows are Red Brahman, bulls are a mixture of Red and Grey Brahman and these distinctions are reflected in the GRM.

Figure 8. The genomic relationship matrix (GRM) across 2,246 animals including 795 cows (reference population) and 1,451 bulls (test population) genotyped for 535,509 SNPs.



The GRM was used to generate GA2CAT predictions for the 1,451 bulls and the quality of the predictions was assessed based on two criteria:

a. Confirming proportions in bulls deviate from random (ie. proportions in the reference population of 795 cows).
b. Confirming the assignment of the last cohort of 622 bulls does not vary depending on which other bulls are included in the GRM.

The proportion of bulls assigned to each category is given in Table 20. Compared to what was observed in the cows, category 2 (WNP, wet not pregnant) was poorly assigned across all bull populations. This was attributed to the many wet and non-pregnant cows being more genetically diverse.

Similar reasoning applies to categories 3 and 4 (dry and early or mid-pregnant) being poorly represented among cows (9.69% and 8.81%) and more highly represented among bulls. The exception is the 114 Bulls (Cohort 1) and category 4 with only 7.02% of bulls assigned to it. This finding further justifies the grouping of categories 3, 4 and 5 into a single "Dry and pregnant" category.

Table 20. **Percentage of animals across the six pregnancy test phenotype categories and population**.

PLS category	795 Cows	All Bulls	114 Bulls	RB 229 Bulls	486 Bulls	622 Bulls
DNP_1*	15.60	17.02	21.93	17.03	15.84	17.04
WNP_2	45.03	6.75	12.28	9.17	4.53	6.59
DEP_3	9.69	21.64	20.18	29.69	20.37	19.94
DMP_4	8.81	21.50	7.02	10.92	25.72	24.76
DLP_5	10.82	16.88	19.30	18.34	15.23	17.20
WEP_6	10.06	16.20	19.30	14.85	18.31	14.47

*DNP_1 = Dry and Not Pregnant; WNP_2 = Wet and Not Pregnant; DEP_3 = Dry and Early Pregnant; DMP_4 = Dry and Mid Pregnant; DLP_5 = Dry and Late Pregnant; WEP_6 = Wet and Early Pregnant

Across all 8,706 genomic relationships (where 8,706 = 1,451 bulls by 6 categories), the average genomic relationship was 16.66% (ie. the expected 1/6) with a SD of 4.85% and a range of 0.01% to 53.85%. Also, the 5th and 95th percentiles were 8.35% and 24.68%, respectively. Table 21 presents the means and standard deviation of % genomic relationship of bulls across the 6 categories. Within a column, the means add to 100% and under random allocation, each category would capture a 1/6 or 16.66% which also the average observed within a column.

The lower variation in % genomic relationship for category 2 (wet non-pregnant) across all bulls and bull populations reflects the lower prediction accuracy that can be expected for this category.

Table 21. Means and standard deviation of % genomic relationship of bulls across the 6 categories.

PLS category	All Bulls	114 Bulls	RB 229 Bulls	486 Bulls	622 Bulls
Means					
DNP_1*	17.183	18.105	17.364	17.080	17.028
WNP_2	16.402	16.848	16.444	15.952	16.656

DEP_3	17.801	17.185	19.492	17.603	17.445
DMP_4	15.315	14.225	14.058	16.008	15.436
DLP_5	16.804	17.443	16.482	16.732	16.861
WEP_6	16.497	16.193	16.161	16.626	16.575
STDev					
DNP_1	4.560	4.917	4.813	4.458	4.463
WNP_2	3.495	4.290	4.194	3.021	3.367
DEP_3	5.000	4.942	6.070	4.674	4.695
DMP_4	5.329	5.848	5.475	5.409	5.005
DLP_5	4.652	5.455	5.319	4.273	4.511
WEP_6	5.457	6.718	5.736	5.313	5.206

*DNP_1 = Dry and Not Pregnant; WNP_2 = Wet and Not Pregnant; DEP_3 = Dry and Early Pregnant; DMP_4 = Dry and Mid Pregnant; DLP_5 = Dry and Late Pregnant; WEP_6 = Wet and Early Pregnant

The last cohort of 622 bulls (Cohort 4) was the subject of further scrutiny. Their genomic assignment to the 4 categories was assessed with 3 GRM:

1. GRM1: Dimension = 2,246 animals including a reference of 795 cows plus 1,451 bulls (ie. the GRM discussed just now and depicted in Fig. 8).
2. GRM2: Dimension = 1,417 animals including a reference of 795 cows plus 622 bulls from Cohort 4.
3. GRM3: Dimension = 796 animals including a reference of 795 cows plus 1 bull in turn from the 622 bulls from Cohort 4 (ie. a total of 622 GRMs).

The aim was to evaluate the impact of different ways of constructing GRM on the rankings of the 622 bulls. Table 22 shows the number of bulls assigned to each category and their average % genomic relationship according to each GRM model. While the number of bulls assigned to each category was similar across the three GRM models (with fewer bulls assigned to categories 2 and 6 than to the other categories), the average genomic relationship was similar for GRM1 and GRM2 and at ~22% but higher for GRM3 at ~28%.

Table 22. Number of bulls (N) assigned to each category and their average % genomic relationship (%GR) according to each GRM model: GRM1, GRM2 and GRM3

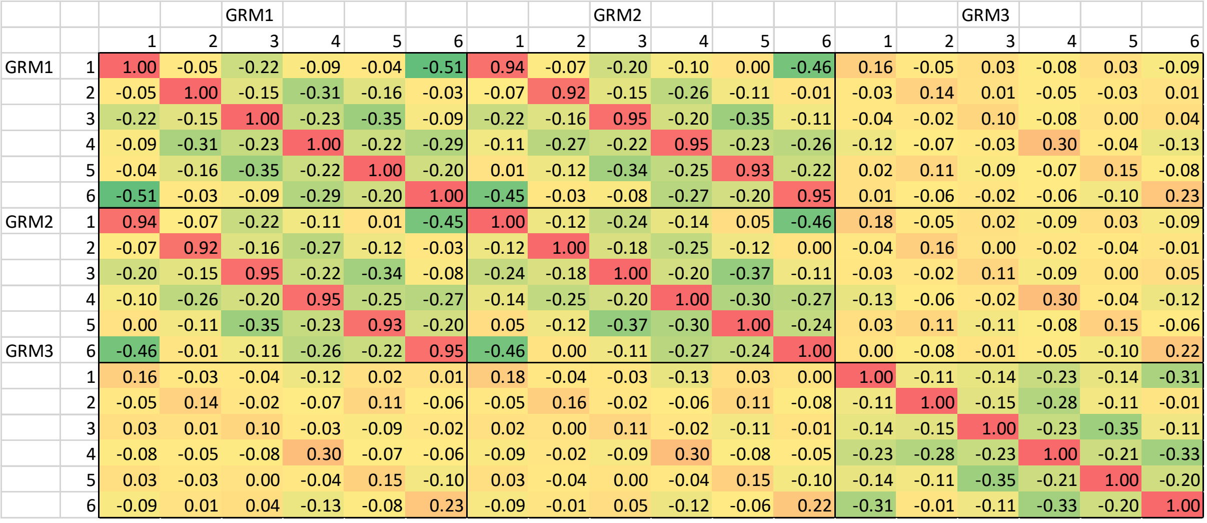
PLS Category	GRM1		GRM2		GRM3	
	N	%GR	N	%GR	N	%GR
DNP_1*	106	22.751	108	23.228	105	28.129
WNP_2	41	21.549	34	20.938	28	28.164
DEP_3	124	23.001	156	23.833	192	27.964
DMP_4	154	19.893	134	19.808	126	29.003
DLP_5	107	22.004	111	23.055	116	28.681
WEP_6	90	23.907	79	24.180	55	28.042

*DNP_1 = Dry and Not Pregnant; WNP_2 = Wet and Not Pregnant; DEP_3 = Dry and Early Pregnant; DMP_4 = Dry and Mid Pregnant; DLP_5 = Dry and Late Pregnant; WEP_6 = Wet and Early Pregnant

The very strong similarity between GRM1 and GRM2 compared to GRM3 further highlighted by the correlation across all 18 genomic relationships (6 categories by 3 GRM models) which is pictured in Fig. 9. Within a phenotype category, the correlation between GRM1 and GRM2 was always > 0.90 while this correlation dropped to 0.10 to 0.30 with GRM3 (but still positive).

A total of 208 bulls (or 33.4%) were assigned to the same category by the three GRM models. This included 34, 8, 54, 55, 40 and 17 in categories 1 to 6, respectively.

Figure 9. Heatmap of the correlation matrix across the 18 genomic relationships observed for the 6 pregnancy test phenotype categories across the three GRM models.



				GRM1						GRM2						GRM3			
		1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
GRM1	1	1.00	-0.05	-0.22	-0.09	-0.04	-0.51	0.94	-0.07	-0.20	-0.10	0.00	-0.46	0.16	-0.05	0.03	-0.08	0.03	-0.09
	2	-0.05	1.00	-0.15	-0.31	-0.16	-0.03	-0.07	0.92	-0.15	-0.26	-0.11	-0.01	-0.03	0.14	0.01	-0.05	-0.03	0.01
	3	-0.22	-0.15	1.00	-0.23	-0.35	-0.09	-0.22	-0.16	0.95	-0.20	-0.35	-0.11	-0.04	-0.02	0.10	-0.08	0.00	0.04
	4	-0.09	-0.31	-0.23	1.00	-0.22	-0.29	-0.11	-0.27	-0.22	0.95	-0.23	-0.26	-0.12	-0.07	-0.03	0.30	-0.04	-0.13
	5	-0.04	-0.16	-0.35	-0.22	1.00	-0.20	0.01	-0.12	-0.34	-0.25	0.93	-0.22	0.02	0.11	-0.09	-0.07	0.15	-0.08
	6	-0.51	-0.03	-0.09	-0.29	-0.20	1.00	-0.45	-0.03	-0.08	-0.27	-0.20	0.95	0.01	-0.06	-0.02	-0.06	-0.10	0.23
GRM2	1	0.94	-0.07	-0.22	-0.11	0.01	-0.45	1.00	-0.12	-0.24	-0.14	0.05	-0.46	0.18	-0.05	0.02	-0.09	0.03	-0.09
	2	-0.07	0.92	-0.16	-0.27	-0.12	-0.03	-0.12	1.00	-0.18	-0.25	-0.12	0.00	-0.04	0.16	0.00	-0.02	-0.04	-0.01
	3	-0.20	-0.15	0.95	-0.22	-0.34	-0.08	-0.24	-0.18	1.00	-0.20	-0.37	-0.11	-0.03	-0.02	0.11	-0.09	0.00	0.05
	4	-0.10	-0.26	-0.20	0.95	-0.25	-0.27	-0.14	-0.25	-0.20	1.00	-0.30	-0.27	-0.13	-0.06	-0.02	0.30	-0.04	-0.12
	5	0.00	-0.11	-0.35	-0.23	0.93	-0.20	0.05	-0.12	-0.37	-0.30	1.00	-0.24	0.03	0.11	-0.11	-0.08	0.15	-0.06
GRM3	6	-0.46	-0.01	-0.11	-0.26	-0.22	0.95	-0.46	0.00	-0.11	-0.27	-0.24	1.00	0.00	-0.08	-0.01	-0.05	-0.10	0.22
	1	0.16	-0.03	-0.04	-0.12	0.02	0.01	0.18	-0.04	-0.03	-0.13	0.03	0.00	1.00	-0.11	-0.14	-0.23	-0.14	-0.31
	2	-0.05	0.14	-0.02	-0.07	0.11	-0.06	-0.05	0.16	-0.02	-0.06	0.11	-0.08	-0.11	1.00	-0.15	-0.28	-0.11	-0.01
	3	0.03	0.01	0.10	-0.03	-0.09	-0.02	0.02	0.00	0.11	-0.02	-0.11	-0.01	-0.14	-0.15	1.00	-0.23	-0.35	-0.11
	4	-0.08	-0.05	-0.08	0.30	-0.07	-0.06	-0.09	-0.02	-0.09	0.30	-0.08	-0.05	-0.23	-0.28	-0.23	1.00	-0.21	-0.33
	5	0.03	-0.03	0.00	-0.04	0.15	-0.10	0.03	-0.04	0.00	-0.04	0.15	-0.10	-0.14	-0.11	-0.35	-0.21	1.00	-0.20
	6	-0.09	0.01	0.04	-0.13	-0.08	0.23	-0.09	-0.01	0.05	-0.12	-0.06	0.22	-0.31	-0.01	-0.11	-0.33	-0.20	1.00

*1 = Dry and Not Pregnant; 2 = Wet and Not Pregnant; 3 = Dry and Early Pregnant; 4 = Dry and Mid Pregnant; 5 = Dry and Late Pregnant; 6 = Wet and Early Pregnant

The testing and validation of the GA2CAT algorithm have been performed using the existing MDH animals (cows and bull populations) and external populations that have the traits of non-ordinal multi-class categories. The results showed that the method worked well when the categorical traits are distinguishable between classes.

4.2.5 Comparison of genomic prediction accuracies for cows' reproductive performance using GA2CAT and machine learning methods

4.2.5.1 Comparison of classification performance of GA2CAT, Random Forest (RF) and Supporting Vector Machines (SVM)

The overall average prediction accuracies (standard deviations in the brackets) of the three methods from a five-fold cross-validation scheme are summarised in the top part of Table 23. When changing the coding of PLS from two to four to six categories, the overall classification accuracy decreased significantly in both populations for all methods in the small population Cows_340, but much less extent in the big population Cows_795.

Table 23. Classification performance of GA2CAT, RF and SVM under different PLS coding systems in two cow populations, using a five-fold cross-validation scheme. A) The overall average classification accuracies (standard deviations in brackets); b) Matthews correlation coefficients (MCC)

A. Overall Accuracy	Cow population					
	Cows_340			Cows_795		
Method	GA2CAT	RF	SVM	GA2CAT	RF	SVM
2PLS	0.46 (0.097)	0.51 (0.073)	0.47 (0.032)	0.53 (0.027)	0.61 (0.029)	0.61 (0.033)
4PLS	0.18 (0.034)	0.43 (0.063)	0.47 (0.018)	0.24 (0.027)	0.44 (0.061)	0.45 (0.052)
6PLS	0.091 (0.024)	0.25 (0.054)	0.29 (0.034)	0.12 (0.025)	0.46 (0.052)	0.45 (0.052)

B. MCC	Cow population					
	Cows_340			Cows_795		
Method	GA2CAT	RF	SVM	GA2CAT	RF	SVM
2PLS	-0.071 (0.19)	0.020 (0.143)	0.000 (0.000)	0.059 (0.049)	0.077 (0.040)	0.000 (0.000)
4PLS	-0.039 (0.029)	-0.037 (0.036)	0.000 (0.000)	0.013 (0.026)	-0.027 (0.043)	0.000 (0.000)
6PLS	-0.017 (0.036)	-0.062 (0.073)	0.000 (0.000)	-0.023 (0.036)	0.053 (0.043)	0.000 (0.000)

RF: Random Forest; SVM: Support Vector Machine.

The poor performance of the three methods under 6PLS could be due to the phenotype of PLS being a non-ordinal multi-class categorical trait. The separation of animals for three Dry and Pregnant classes, i.e. early, mid, and late pregnancy, was not clean-cut as those in the binary situation (2PLS, non-pregnant vs pregnant). For the GA2CAT, the genomic relationships between animals in these three classes in the training populations were very similar, therefore the predicted contributions of the animals in the validation populations to six categories of PLS (i.e. GA2CAT values) were very similar. As result, it made the correct assignment of the animals in the testing populations to different categories extremely difficult. The results indicate the necessity of recoding PLS records before applying different analytical methods to achieve reliable results.

Across two cow populations, for the same coding system, e.g. 6 categories (6PLS), the two ML methods (RF and SVM) seemed to outperform the GA2CAT (see the average accuracies in Table 23). The margin was huge in the population Cows_795 (0.46 (RF), 0.45 (SVM) vs 0.12 (GA2CAT). The difference between RF and SVM was little in comparison to either of them with the GA2CAT. However, when investigating further on the classes correctly classified, we found that both RF and SVM assigned all of the individuals in the validation datasets to the category of Wet and Non-Pregnant. This was the class with the largest number of phenotypic observations in Cows_795. This confirms the downside of ML methods that bias toward the majority class by over-sampling the abundant classes and under-sampling minor classes (Chicco and Jurman 2020).

When evaluating the performance of three methods by the MCC values (the bottom part of Table 23), all three methods had the MCC values either zero (SVM) or close to zero. These suggest that: a) the phenotype PLS is a low heritability trait, as all three methods followed a random prediction behaviour (MCC values ~ 0.00). In addition, the accuracy values for the GA2CAT fitted the random sampling expected prediction accuracies of 0.5 (PLS2), 0.34 (PLS4) and 0.25 (PLS6); b) there was no significant classification performance difference among the GA2CAT, RF and SVM.

The results from a five-fold cross-validation scheme indicate that different coding systems of PLS categories greatly impacted the classification outcome of the GA2CAT. For highly imbalanced non-ordinal multiclass datasets, using the average overall accuracy value for evaluating the classification performance of the GA2CAT and ML methods can be misleading and MCC values should be applied. A GA2CAT value is the weighted average of genomic relationships between reference and validation populations for a particular category, it reflects better the heritability nature of a phenotypic trait.

4.3 Economic models for commercial cow fertility and cost:benefit estimates

4.3.1 Estimating responses to selection

A desktop modelling study was undertaken to compare the contribution to both beef herd productivity and profitability associated with selection of bulls for cow reproduction performance based on our GA2CAT genomic predictions approach.

Even though our phenotype is measured in terms of the Pregnancy Test, the true trait of interest is Heifer's Second Joining Pregnancy rate (HPR) for genetic progress analyses, which in turn translates into a weaning rate (WR) for economic analyses. Importantly, only 50% of the genetic progress (which is only achieved via sires) is transmitted to cows, ie. a 4% HPR translates into 2%.

Estimated response to selection was based on annual genetic progress (*d*G) and determined according to the following formula (Bourdon 1997):

$$dG = \frac{ACC \times i \times \sigma_G}{L}$$

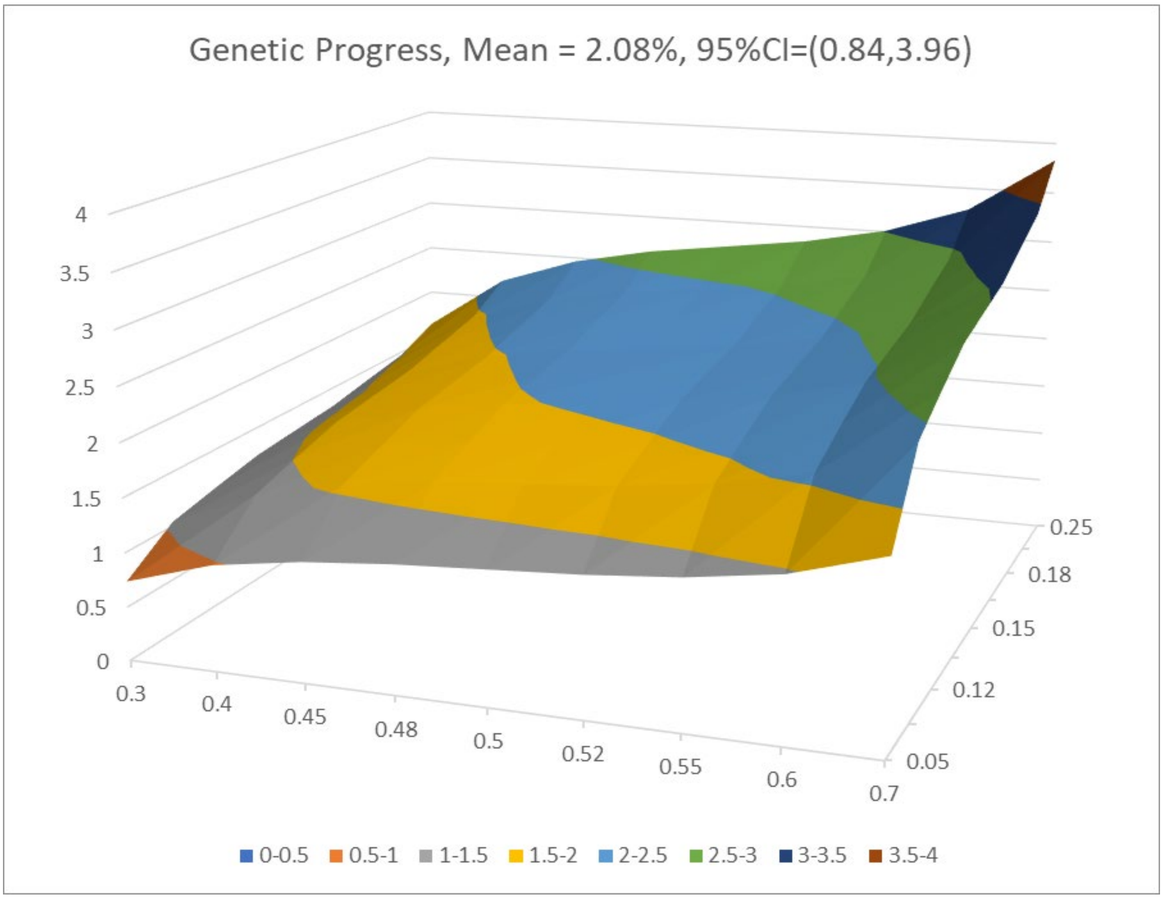
ACC is the accuracy of selection, *i* is the selection intensity, σ_G is the genetic standard deviation and L is the generation interval.

For the modelling study, the following parameters and ranges were assumed:

a) An HPR phenotypic variance of 0.25 (rate, not %) confirmed via both GA2CAT real phenotypes and GAPLS predicted proportions.
b) A "plausible" range of h^2 values for HPR are 0.05 to 0.25.
c) A "plausible" range of ACC values is 0.20 to 0.50
d) A "plausible" range of L is 4, 5, 6 years
e) The intensity of selection (*i*) can be estimated based on the proportion (*p*) of selected animals as a percentage of those available for selection. A "realistic" *p* for sires is 30% (ie. buying 300 sires out of 1,000 available for purchase). But worth exploring *p* 20% (better) or 40% (worse).
f) A "realistic" *p* for cows is a function of replacement rates for cows and drafting out empty cows. A good assumption of 20% at first joining will not get pregnant.

The data collected by the team on the heritability value of the trait, range of observed variance, and generation interval, have led to an estimate of a 2 % annual genetic improvement in pregnancy rate at the second joining opportunity (Fig. 10).

Figure 10. Modelling response to selection: the mean of 2%, right at the intercept between yellow and blue shades, can be observed across all ranges of accuracies (width, 0.3 to 0.7) and heritabilities (depth, 0.05 to 0.25).



4.3.2 Modelling the impact of introducing better herd bulls over time

A cow fertility intervention that first targets the selection of sires of commercial cows, will take time to achieve increased pregnancy rates. Based on the 2 % p.a. improvement of this sire trait that we have estimated, it will take 10 years from introducing the first cohort of improved bulls to achieve a 6% improved pregnancy rate in the cow herd.

Our observations at Iffley station in 2020 and 2021 have shown that the pregnancy rate for Brahman cows at their second joining was around 45 %. The 6 % improvement after 10 years would therefore result in a second joining pregnancy rate of 48 % (Table 24).

The CRC for Beef Genetic Technologies has shown that the ability to reconceive at the second joining is a different trait from the ability to conceive at the first joining. These traits are somewhat correlated (Johnston et al. 2013). Lifetime fertility is also positively correlated, but a separate, and less heritable trait (Johnston et al. 2013).

The correlated changes in those traits that could be estimated over 10 years are therefore very small. In the simple calculations set out below, we have assumed that no correlated increases in the heifer pregnancy rate and lifetime fertility rate will occur.

Table 24. Changes in reproductive performance over a 10-year timespan after introducing herd bulls with improved genetic value for daughter pregnancy rate at the second joining opportunity

Year of first joining	Bull genetic value	Heifer genetic value	Year brand of mating 1	Heifer preg rate (mating 1)	Year brand of mating 2	First calf preg rate (mating 2)	subsequent preg rate
2021	1.02	1.00	2020	0.78	2019	0.45	0.70
2022	1.04	1.00	2021	0.78	2020	0.45	0.70
2023	1.06	1.00	2022	0.78	2021	0.45	0.70
2024	1.08	1.01	2023	0.78	2022	0.45	0.70
2025	1.10	1.02	2024	0.78	2023	0.45	0.70
2026	1.13	1.03	2025	0.78	2024	0.45	0.70
2027	1.15	1.04	2026	0.78	2025	0.46	0.70
2028	1.17	1.05	2027	0.78	2026	0.46	0.70
2029	1.20	1.07	2028	0.78	2027	0.47	0.70
2030	1.22	1.08	2029	0.78	2028	0.47	0.70
2031	1.24	1.09	2030	0.78	2029	0.48	0.70

4.3.3 Using BreedCow+ to compare gross margin for herd

A preliminary study on the impact of a small increase in pregnancy rate at the 2nd mating opportunity in a Brahman herd in northern Australia was conducted using the BreedCow+ program. "North Queensland Gulf" was selected from the DCAP regional base files in Breedcow+ to model a 1,500-cow herd. Bulls were assumed to be sold as weaners and heifer weaners would be retained each year and first join would occur at 2-year-olds. Livestock prices were updated to the 6 June 2022 NLRS indicators for Northern Queensland.

An online NQGulf Breedcowplus file in the excel version was compiled and the re-conception rate of the first calf heifers was set as 2%. The parameter assumptions for calving and death rates, sale prices for heifers and bulls, and herd structures for steers and females are given in Appendix 7.3.

The preliminary results for gross margins of a herd of 1500 (total adult equivalents) are shown in Table 25. From the second and third columns in Table 25, it can be seen that comparing the herd without any improvement (NQ Gulf base), improving the re-conception rate by 2% (i.e. plus 2% on first calf heifers, the second column in Table 25) would yield an increase in the herd gross margin (GM) by $1,129.81 after taking into account of the GM interest (i.e. $1,129.81 =$182457.58 -$181327.77). If the base herd was further optimized for maximum herd GM, then the mature cow cull age would fall to 8-9 years, and the benefit of increasing the re-conception rate of first-calf heifers by 2% would see GM being $1,083.76 after taking into account of the GM interest (i.e. $1083.76 =$185660.04 -$184576.28, Table 25, last column). If considering the gross margin per adult equivalent (GM/PAE), a 2% increase in the re-conception rate would see a net profit of $0.75 per head and $0.72 respectively for the two above-mentioned scenarios. The results suggest that there is certainly herd gross margin benefit when improving the re-conception rate of first calf heifers by 2% if the change can be fully implemented.

Table 25. Comparison of gross margins using the herds with and without a 2% improvement in reconception rate

	NQ Gulf base	**plus 2% on first calf heifers**	**optimum cull base**	**optimum base +2% hfrs**
Total adult equivalents	1500	1500	1500	1500
Total cattle carried	1739	1740	1744	1743
Weaner heifers retained	233	233	236	235
Total breeders mated	805	803	803	801
Total breeders mated & kept	751	749	740	739
Total calves weaned	465	466	473	475
Weaners/total cows mated	57.76%	58.11%	58.86%	59.34%
Wnrs/cows mated and kept	61.91%	62.25%	63.85%	64.32%
Overall breeder deaths	2.50%	2.50%	2.50%	2.50%
Female sales/total sales %	47.89%	47.90%	47.96%	47.99%
Total cows and heifers sold	202	203	206	207
Maximum cow culling age	11	11	8	8
Heifer joining age	2	2	2	2
Weaner heifer sale & spay	0.00%	0.00%	0.00%	1.00%
One yr old heifer sales %	0.00%	0.00%	0.00%	0.00%
Two yr old heifer sales %	60.55%	61.06%	39.70%	39.70%
One yr old heifer spay and unmated %	0.00%	0.00%	0.00%	0.00%
Two yr old heifer spay and unmated %	0.00%	0.00%	0.00%	0.00%
Total steers & bullocks sold	220	221	224	225
Max bullock turnoff age	3	3	3	3
Average female price	$668.30	$668.53	$657.94	$653.17
Average steer/bullock price	$841.75	$841.75	$841.75	$841.75
Capital value of herd	990292.7	989830.2	991580	990940.2
Imputed interest on herd val	49514.64	49491.51	49579	49547.01
Net cattle sales	320427.4	321481.4	323844.7	324822.9
Direct costs excluding bulls	63107.41	63137.47	63287.86	63282.24
Bull replacement	26477.61	26394.87	26401.51	26333.56
Gross margin (GM) for herd	230842.41	231949.09	234155.28	235207.05
GM after imputed interest	181327.77	182457.58	184576.28	185660.04
		1129.81		**1083.76**
GM per adult equivalent (GM/PAE)	153.89	154.63	156.1	156.8
GM/PAE after interest	120.89	121.64	123.05	123.77
		0.75		**0.72**

4.3.4 Estimating the cost of implementing GA2CAT in a beef enterprise

We based our estimates of implementing the GA2CAT intervention on our experience of working with cow herds at Iffley station over the last 3 years. On a property capable of supporting about 10,000 head, in any given year there will be around 2,000 cows that present for pregnancy testing at their second mating opportunity. We have assumed that the pregnancy test muster and vet costs are not an additional cost to the enterprise, but the collection of phenotypes and TSU for DNA analysis would be. The estimated costs over 10 years for a bull buyer or a bull seller are given in Table 26.

Table 26. The cost estimates for the implementation of GA2CAT over 10 years for a bull buyer or a bull seller

Activity or material	Estimated cost	Cost over 10 years
Bull buyer		
Pregnancy testing and sample collection include extra farm staff (n=2,000)	$11,600	$69,600 (assume six collections)
Building DNA pools (lab work) and genotyping	$9,100	$54,600 (assume six collections)
Recalibrating genomic predictions with new data	$2,000	$12,000 (assume six rounds)
GA2CAT analyses	$1,200	$12,000 (assume every year)
Sub Total	--	**$148,200**
Bull seller		
Candidate sires genotyping (n=500)**	$20,000	$200,000 (assume every year)
Sub Total	--	**$200,000**
Grand total		**$348,200**

** The costs for genotyping candidate sires might not incur directly on the property buying the bulls. The genotypes might be already collected for other reasons, e.g. registration with a breed society, or parent verification and internal management.

The results show that the implementation of GA2CAT over 10 years would incur the cost of $148,200 for a bull buyer, or $200,000 for a bull seller. The reference population for the genomic estimation is tailored for individual property (i.e., the reference population is specific for the property in which cow phenotypes are collected). Therefore, the cost incurred for creating and expanding this reference population should be made by individual bull buyers willing to choose better bulls for their herds. On the other hand, the genotyping cost of candidate sires is a cost for bull sellers. It is worth mentioning that the genotypes of candidate sires will potentially be already available due to property management e.g. registration with breed societies, or parent verification as part of internal management. Reusing the genotypes for more than one outcome will add value to the technology and might encourage producers to access and benefit from the technology that is already widely used in other sectors of the livestock industry.

The GA2CAT algorithm will become more accurate when more phenotypic and genotypic data are available in the reference population. We have therefore assumed that one year of data will be accumulated before herd bulls are selected and that genotyping of pools will be continued on a bi-annual basis to achieve the reference population required to generate genomic prediction accuracies of pregnancy test outcomes for bulls with useful accuracy.

It is worth mentioning that the results presented above are preliminary and were based on publicly available tools for beef industry economic analysis. A more comprehensive study should be considered across multiple MLA projects to investigate the requirement of reference populations and the cost-benefit analysis of genetic technology investments at a range of scales e.g. herd, enterprise, region and national levels.

4.4 Implementation plan for routine estimation of gEBV for candidate bulls established based on commercial cow performance data

During the development, testing and validation of GA2CAT with different populations, we have developed a UNIX Shell script and a FORTRAN95 source code that can be used for rapidly generating GA2CAT values for candidate bulls of any population.

Both the UNIX and the FORTRAN95 codes have been tested independently and they have produced identical GA2CAT values when the same datasets are used. The executable programs can be made available once potential collaborators or clients are identified and formal license agreements between MLA, CSIRO and third party are put into place.

Since the GA2CAT programs are based on genomic relationships between reference cows with both genotype and performance data and candidate bulls with genotype information, users will need to employ their own choice of a program (either public domain available or commercial licensed) to generate genomic relationship matrix before applying the GA2CAT program. Should producers or collaborators decide to use the methodology, an agreement will need to be developed between the MLA, CSIRO and third parties.

5. Conclusion

All milestone criteria have been met.

- There were very low correlations between the genomic predictions using pooled DNAs vs individually genotyped animals. This suggests that the genomic ranking of candidate bulls using DNA pools of cows of the same phenotype of PLS had different outcomes when using individually genotyped cows. The reproductive trait – pregnancy and lactation status is a lowly heritable trait. As a precautionary measure, we recommend the use of individual genotypes to make genomic predictions of bulls' reproductive performance';
- The genomic prediction of 1,451 MDH and Gipsy Plains bulls from 2020, 2021, and 2022 seasons were performed to estimate their contributions to their female progeny's reproductive performance using GA2CAT. The rankings of the animals were provided to the farms on time to assist their sire selection decision;
- The testing and validation of the GA2CAT algorithm have been performed using the existing MDH animals (cows and bull populations) and external populations that have the traits of non-ordinal multi-class categories. The results showed that the method worked well when the categorical traits are distinguishable between classes with a strong genomic background and the method can be easily extended to deal with other phenotypes with non-ordinal multi-class features.
- Two programs (a UNIX Shell script and a FORTRAN95 source code) have been developed for the efficient computation of the GA2CAT algorithm. They have been independently tested

with identical results. The programs can be used for rapidly generating GA2CAT values for candidate bulls of any population.

- A preliminary study on the cost: benefit analysis using the BREEDCOW+ herd modeling software showed that for a base herd of 1500 animals, improving the re-conception rate by 2% on first calf heifers would yield an increase in the herd gross margin (GM) by $1,129.81 after taking into account of the GM interest or $0.75 GM per adult equivalent.

5.1 Key findings

- The GA2CAT results are easy to interpret, promoting adoption by commercial producers wanting to improve the reproductive performance of their herds;
- The GA2CAT method will not be influenced by the phenotype scoring system allowing producers to tailor selection strategies to their production system;
- The population variation of GA2CAT values reflects the degree of relatedness between testing and reference populations: as would be expected, GA2CAT predictions are valid for the most closely related cattle populations;
- The negative correlation between GA2CAT values for Wet and Early Pregnancy and GEBV of DTC (-0.2) further indicates GA2CAT can be used as a good tool to select bulls.

5.2 Benefits to industry

- Optimised methodology for dealing with GEBV for a categorical fertility phenotype (GA2CAT) has been developed
- Genomic ranking of herd bulls available for selection in a timely fashion has been achieved and the results communicated to the client
- Validation population of thousands of tropical beef cattle genotypes with associated phenotypes has been assembled

5.3 Future research and recommendations

- As a low heritability and imbalanced multi-class trait, the genomic selection for 2nd joining pregnancy and lactation status presents serious challenges. The new metric “GA2CAT” is recommended to apply in conjunction with individual bull’s BBSE (Bull Breeding Soundness Examination) results when selecting candidate bulls for mating.
- Due to 2019 flood, the project could only collect the candidate bulls from Gipsy Plains during the 2021-2022 and 2022-2023 seasons. As result, we have not been able to conduct a systematic validation of the accuracies of genomic prediction by our new method GA2CAT for all individual bulls that sired the MDH cows (with 2nd joining pregnancy and lactation status records) from 2019-2022 . A future study is needed to evaluate the accuracy of the new prediction tool;
- Further research is required to identify a new phenotype that is more heritable than the phenotype of pregnancy and lactation status.

6. References

Breiman, L. (2001). Random Forests. Mach. Learn. **45**: 5.

Caraux G, Pinloche S. 2005. PermutMatrix: a graphical environment to arrange gene expression profiles in optimal linear order." Bioinformatics. 21(7):1280-1281.

Chang C.C., Chow C.C., Tellier L.C.A.M., Vattikuti S., Purcell S.M. and Lee J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience, **4**:7. DOI 10.1186/s13742-015-0047-8.

Chicco, D., and Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC genomics, *21*, 1-13.

James G., Witten D., Hastie T., and Tobshirani (2013). 'An Introduction to Statistical Learning'. Springer, Heidelberg, Germany

Li Y., Lehnert S.A., Porto-Neto L., McCulloch R., McWilliam S., Alexandre P, McDonald J., Smith C., and Reverter A. (2022). Proceedings of 12th World Congress on Genetics Applied to Livestock Production. Rotterdam. The Netherlands. 3-8 July 2022.

Li Y., Porto-Neto L., McCulloch R., McWilliam S., Alexandre P., *et al.* (2021). Ranking Brahman bulls for female reproductive performance in northern Australian commercial environments using DNA pooling. Proc. Assoc. Advmt. Anim. Breed, Genet, 24: 204-207.

Loh P.-R., Palamara P.F. and Price A.L. (2016). Fast and accurate long-range phasing in a UK Biobank cohort. Nat. Genet. 48: 811-816.

Reverter A., Porto-Neto L.R., Fortes M.R., McCulloch R., Lyons R.E., *et al.* (2016). Genomic analyses of tropical beef cattle fertility based on genotyping pools of Brahman cows with unknown pedigree. J Anim Sci. 94: 4096-4108. doi:10.2527/jas2016-0675.

VanRaden, P.M, 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. (91):4414-4423. doi: 10.3168/jds.2007-0980.

7. Appendix

7.1 Abbreviations

BBSE – Bull Breeding Soundness Examination

CSIRO - The Commonwealth Scientific and Industrial Research Organisation

DTC - Days to calving

EBV - Estimated breeding values

GA2CAT - Genomic contributions to a Categorical Trait

GBLUP - Genomic best linear unbiased prediction

GEBV - Genomic estimated breeding value

GRM - Genomic relationship matrix

MDH – McDonald Holdings Pty Ltd

MLA – Meat & Livestock Australia

PLS - Pregnancy and lactation status

SNP – Single nucleotide polymorphisms

7.2 Supplementary Tables

Supplementary Table 1. Average genomic breeding values (GEBV) of progeny pregnancy testing outcome (PLS) of bulls in four quartiles of two populations using low density (19,089) SNP panel

	MDH2020			
Quartile	# Bulls	Av. GEBV	Min	Max
1 -Top 25%	120	0.0619	0.0372	0.14
2	120	0.0199	0.0055	0.0371
3	121	-0.0098	-0.0267	0.0321
4 – Bot. 25%	121	-0.0581	-0.1986	-0.0271
All	482	0.0033	-0.1986	0.1252
	SmartF			
Quartile	# Bulls	Av. GEBV	Min	Max
1 -Top 25%	44	0.0189	-0.0395	0.0750
2	44	0.0013	-0.0369	0.0481
3	44	-0.0109	-0.0499	0.0356
4 – Bot. 25%	45	-0.0372	-0.1266	0.0150
All	177	-0.007	-0.127	0.0750

Supplementary Table 2. Average genomic breeding values (GEBV) of progeny pregnancy testing outcome (PLS) of bulls in four quartiles of two populations using mid-density SNP panels

	MDH2020 (with 54,791 SNPs)			
Quartile	# Bulls	Av. GEBV	Min	Max
1 -Top 25%	120	0.0836	0.0448	0.221
2	120	0.0272	0.0096	0.0447
3	121	-0.0102	-0.0315	0.0409
4 – Bot. 25%	121	-0.0741	-0.322	-0.0315
All	482	0.0064	-0.322	0.1504
	SmartF (with 74,584 SNPs)			
Quartile	# Bulls	Av. GEBV	Min	Max
1 -Top 25%	44	0.0519	-0.0322	0.0695
2	44	-0.0009	-0.0472	0.1621
3	44	-0.0374	-0.1371	0.0293
4 – Bot. 25%	45	-0.0838	-0.1802	-0.0043
All	177	-0.0192	-0.1802	0.1621

Supplementary Table 3. Pearson's correlations between GA2CAT values for 6 categories of PLS and GEBV of 10 growth and fertility traits from bulls of CRCTC

	wt	cs	sc	sheath	dens	mass	mott	pns	pd	mp	DNP_1	WNP_2	DE_3	DM_4	DL_5	WE_6
DNP_1	-0.01	-0.021	-0.021	0.049	-0.036	-0.04	-0.032	-0.031	0.038	0.027	1	0.154	-0.301	-0.525	-0.075	0.035
WNP_2	0.058	-0.011	0.037	0.014	0.008	0.025	0.037	0.054	-0.027	0.007	0.154	1	-0.104	-0.252	0.175	-0.392
DE_3	0.061	-0.03	0.051	-0.029	0.047	0.045	0.018	-0.006	-0.009	0.034	-0.301	-0.104	1	0.004	-0.33	-0.183
DM_4	-0.002	0.023	-0.035	-0.03	0.002	-0.01	-0.026	-0.005	0.017	-0.037	-0.525	-0.252	0.004	1	-0.189	-0.333
DL_5	-0.022	-0.018	0.001	0.047	-0.04	-0.038	-0.012	0.002	0.009	-0.014	-0.075	0.175	-0.33	-0.189	1	-0.464
WE_6	-0.043	0.034	-0.006	-0.03	0.019	0.026	0.027	0.003	-0.032	0.003	0.035	-0.392	-0.183	-0.333	-0.464	1

Abbreviations: wt: live weight; cs: condition score, sc: scrotal circumference; sheath: sheath score; dens: sperm density; mass: sperm mass mobility; mott: sperm motility; pns: percent normal sperm; proximal cytoplasmic droplets; mp : midpiece abnormalities. *DNP_1 = Dry and Not Pregnant; WNP_2 = Wet and Not Pregnant; DEP_3 = Dry and Early Pregnant; DMP_4 = Dry and Mid Pregnant; DLP_5 = Dry and Late Pregnant; WEP_6 = Wet and Early Pregnant.*

Supplementary Table 4. Pearson's correlations between GA2CAT values for 6 categories of PLS and GEBV of 10 growth and fertility traits from bulls of SGT

	wt	cs	sc	sheath	dens	mass	mott	pns	pd	mp	DNP_1	WNP_2	DE_3	DM_4	DL_5	WE_6
DNP_1	0.005	-0.075	-0.063	-0.046	-0.007	-0.004	-0.048	-0.022	0.021	0.036	1	-0.171	-0.157	-0.11	-0.069	-0.396
WNP_2	0.009	0.055	0.026	0.032	0.117	0.162	0.084	0.022	-0.03	-0.05	-0.171	1	-0.045	-0.163	-0.233	-0.013
DE_3	-0.022	-0.052	0.059	-0.006	0.078	0.073	0.019	0.1	-0.04	-0.012	-0.157	-0.045	1	-0.226	-0.182	-0.321
DM_4	0.041	0.082	-0.1	-0.097	0.11	0.04	0.111	0.097	-0.147	-0.026	-0.11	-0.163	-0.226	1	-0.299	-0.161
DL_5	-0.007	-0.057	0.037	0.123	-0.099	-0.047	-0.071	-0.163	0.162	0.091	-0.069	-0.233	-0.182	-0.299	1	-0.347
WE_6	-0.017	0.059	0.032	-0.012	-0.114	-0.127	-0.044	-0.005	0	-0.054	-0.396	-0.013	-0.321	-0.161	-0.347	1

Supplementary Table 5. Pearson's correlations between GA2CAT values for 6 categories of PLS and GEBV of 10 growth and fertility traits from bulls of DMT

	wt	cs	sc	sheath	dens	mass	mott	pns	pd	mp	DNP_1	WNP_2	DE_3	DM_4	DL_5	WE_6
DNP_1	-0.032	-0.009	-0.109	-0.009	-0.032	-0.079	-0.091	-0.09	0.149	0.155	1	-0.257	-0.226	0.017	-0.213	-0.332
WNP_2	-0.041	-0.029	-0.061	0.057	0.024	0.022	0.133	0.115	-0.13	-0.182	-0.257	1	-0.043	-0.286	-0.052	-0.021
DE_3	0.157	0.03	0.023	0.014	0.072	0.042	0.007	-0.022	0.049	-0.032	-0.226	-0.043	1	-0.267	-0.149	-0.246
DM_4	-0.066	-0.033	-0.025	-0.079	0.047	0.023	0.003	-0.015	-0.017	0.04	0.017	-0.286	-0.267	1	-0.258	-0.243
DL_5	0.028	0.081	0.096	-0.001	-0.023	-0.003	0.021	0.014	-0.025	0.007	-0.213	-0.052	-0.149	-0.258	1	-0.347
WE_6	-0.048	-0.045	0.042	0.032	-0.063	0.002	-0.025	0.029	-0.055	-0.041	-0.332	-0.021	-0.246	-0.243	-0.347	1

Abbreviations: wt: live weight; cs: condition score, sc: scrotal circumference; sheath: sheath score; dens: sperm density; mass: sperm mass mobility; mott: sperm motility; pns: percent normal sperm; proximal cytoplasmic droplets; mp : midpiece abnormalities. *DNP_1 = Dry and Not Pregnant; WNP_2 = Wet and Not Pregnant; DEP_3 = Dry and Early Pregnant; DMP_4 = Dry and Mid Pregnant; DLP_5 = Dry and Late Pregnant; WEP_6 = Wet and Early Pregnant*

Supplementary Table 6. **Pearson's correlations between GAPLS values for 6 categories of PLS and GEBV of 10 growth and fertility traits from bulls of BLK**

	wt	**cs**	**sc**	**sheath**	**dens**	**mass**	**mott**	**pns**	**pd**	**mp**	**DNP_1**	**WNP_2**	**DE_3**	**DM_4**	**DL_5**	**WE_6**
DNP_1	0.022	0.036	-0.153	-0.016	-0.055	-0.074	-0.064	-0.017	-0.04	0.04	1	-0.224	-0.262	-0.188	-0.058	-0.191
WNP_2	0.032	0.051	0.099	-0.038	0.175	0.165	0.153	0.053	-0.02	-0.006	-0.224	1	0.007	-0.126	-0.316	-0.023
DE_3	-0.05	-0.043	-0.041	-0.076	-0.031	-0.032	-0.062	-0.125	0.159	0.049	-0.262	0.007	1	-0.182	-0.266	-0.332
DM_4	0.124	-0.037	-0.114	-0.001	-0.037	-0.11	-0.094	-0.08	0.078	0.105	-0.188	-0.126	-0.182	1	-0.151	-0.344
DL_5	-0.013	0.02	0.122	0.106	-0.026	-0.052	0.031	0.038	-0.048	0.011	-0.058	-0.316	-0.266	-0.151	1	-0.273
WE_6	-0.086	-0.003	0.091	0.013	0.025	0.132	0.073	0.134	-0.129	-0.173	-0.191	-0.023	-0.332	-0.344	-0.273	1

Abbreviations: wt: live weight; cs: condition score, sc: scrotal circumference; sheath: sheath score; dens: sperm density; mass: sperm mass mobility; mott: sperm motility; pns: percent normal sperm; proximal cytoplasmic droplets; mp : midpiece abnormalities. *DNP_1 = Dry and Not Pregnant; WNP_2 = Wet and Not Pregnant; DEP_3 = Dry and Early Pregnant; DMP_4 = Dry and Mid Pregnant; DLP_5 = Dry and Late Pregnant; WEP_6 = Wet and Early Pregnant.*

Supplementary Table 7. Pearson's correlations between GAPLS values for 6 categories of PLS and GEBV of 10 growth and fertility traits from bulls of BEL

	wt	cs	sc	sheath	dens	mass	mott	pns	pd	mp	DNP_1	WNP_2	DE_3	DM_4	DL_5	WE_6
DNP_1	-0.03	0.013	-0.109	-0.101	0.113	0	-0.193	-0.049	-0.051	0.19	1	-0.073	-0.054	-0.116	-0.228	-0.359
WNP_2	0.058	0.046	0.092	0.029	-0.07	-0.054	0.013	-0.113	0.181	-0.038	-0.073	1	-0.05	-0.371	-0.19	-0.015
DEP_3	-0.012	0.072	-0.023	0.023	0.02	-0.02	-0.059	0.017	-0.102	0.129	-0.054	-0.05	1	-0.258	-0.141	-0.395
DM_4	0.028	-0.007	-0.088	-0.024	-0.032	-0.04	-0.022	0.044	-0.035	-0.045	-0.116	-0.371	-0.258	1	-0.158	-0.225
DLP_5	0.111	0.04	0.073	-0.034	-0.061	0.007	0.069	-0.049	-0.003	-0.031	-0.228	-0.19	-0.141	-0.158	1	-0.278
WEP_6	-0.114	-0.12	0.061	0.086	0.018	0.075	0.143	0.089	0.047	-0.157	-0.359	-0.015	-0.395	-0.225	-0.278	1

Abbreviations: wt: live weight; cs: condition score, sc: scrotal circumference; sheath: sheath score; dens: sperm density; mass: sperm mass mobility; mott: sperm motility; pns: percent normal sperm; proximal cytoplasmic droplets; mp : midpiece abnormalities. *DNP_1 = Dry and Not Pregnant; WNP_2 = Wet and Not Pregnant; DEP_3 = Dry and Early Pregnant; DMP_4 = Dry and Mid Pregnant; DLP_5 = Dry and Late Pregnant; WEP_6 = Wet and Early Pregnant.*

7.3 Basic parameter assumptions for the economic modelling of herd gross margins (GM) or per adult equalvelent (PAE) using BreedCow+ program

Section A. Calving and death rate assumptions

Cattle age start year	Weaners	1	2	3	4	5	6	7	8	9	10	11	12	13
Cattle age end year	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Expected conception rate for age group (%)	na	0.0%	78.0%	48.0%	70.0%	70.0%	70.0%	70.0%	65.0%	65.0%	60.0%	60.0%	60.0%	60.0%
Expected calf loss from conception to weaning (%)	na	0.0%	16.4%	9.5%	11.8%	11.8%	11.8%	11.8%	13.7%	13.7%	13.7%	13.7%	13.7%	13.7%
Proportion of empties (PTE) sold (%)	na	0%	100%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%
Proportion of pregnants sold (%)	na	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Calves weaned/cows retained	na	0.0%	83.6%	45.8%	63.6%	63.6%	63.6%	63.6%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Female death rate	2.5%	2.5%	2.5%	2.5%	2.5%	2.5%	2.5%	2.5%	2.5%	2.5%	2.5%	2.5%	2.5%	2.5%
Spayed or unmated female death rate	2.5%	2.5%	2.5%	2.5%	2.5%	2.5%	2.5%	2.5%	2.5%	2.5%	2.5%	2.5%	2.5%	2.5%
Male death rate	2.5%	2.5%	2.5%	2.5%	2.5%									

Section B. Sale Prices

Age at sale	Weaners	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Heifers/cows	$239	$483	$693	$635	$635	$635	$635	$635	$635	$635	$635	$635	$635	$635	$635
Spays or unmated females	na	$483	$693	$635	$635	$635	$635	$635	$635	$635	$635	$635	$635	$635	$635
Steers/bullocks ..	$311	$571	$843	$835	$0	$0									

Section C. Steer and bullock herd structure

Maximum male turnoff age (integer)	3	*(Enter 0 for weaners)*				
Steer or bullock age in months	*5 to 11*	*12 to 23*	*24 to 35*	*36 to 47*	*48 to 59*	*60 plus*
Steer or bullock age group	0	1	2	3	4	5
Number available at start year .	238	232	226	37	0	0
Number reserved as bulls	0	0	0	na	na	na

Optional sales %	0.00%	0.00%	83.00%	0.00%	0.00%	na
Transfers to bull herd	na	na	0	na	na	na
Sales at each age .	0	0	187	37	0	0
Total steers and bullocks sold	225	Average price =		$841.75		

Section D: Bull Requirements

Bull/cow ratio	4.00%	Bulls required		32	
Bulls purchased/yr. % of bulls required	20.00%	6	*@ price*	$5,000	
Home bred bulls kept/yr. % of required	0.00%	0	*@ value*	$843	
Bulls sold/yr...	*Calculated*	5	*@ price*	$1,186	
Bull deaths	5.00%	2			
	Average value per head of bulls on hand				$3,365
	Net bull replacement cost (total)				$26,334
	Net bull replacement cost per calf weaned				$55.42

Section E: Female Herd Structure

Weaner heifers to be retained	235	*(Enter*	*235.21*	*to give required AEs)*		
Age at first joining (max 3 yrs)	2		If 1, % joined =	100.00%		
Cow culling age (integer, max 13)	8					
Required herd size (AE)	1500					
Surplus weaner heifers sold or spayed ...	2		Weaner % spayed =	0.00%	*Number =*	0

7.4 2021 AAABG Paper. Proc. Assoc. Advmt. Anim. Breed. Genet: 24, 204-207.

RANKING BRAHMAN BULLS FOR FEMALE REPRODUCTIVE PERFORMANCE IN NORTHERN AUSTRALIAN COMMERCIAL ENVIRONMENTS USING DNA POOLING

Y. Li[1], L. Porto-Neto[1], R. McCulloch[1], S. McWilliam[1], P. Alexandre[1], J. McDonald[2], A. Reverter[1] and S. Lehnert[1]

[1] CSIRO Agriculture and Food, St Lucia, QLD, 4067 Australia
[2] MDH Pty Ltd, Cloncurry QLD 4824, Australia

SUMMARY

Female fertility is one of the important reproductive traits that directly impact the profitability of commercial beef breeding herds. DNA pooling of cows with reproductive records can provide a cost-effective way for assessing and predicting the contribution of individual bulls to the fertility of their female offspring. However, panels of different SNP density exist and their impact on genomic prediction is unknown when DNA pooling is applied. In this study, using the genotype and phenotype (pregnancy test and lactation status) from two Brahman cattle populations in north Queensland, one containing 715 samples genotyped with 54,791 SNPs, the other consisting of 290 samples genotyped with 74,584 SNPs, we investigated genetic relationships between the two populations as well as rankings of individual bulls based on genomic prediction for pregnancy test outcome of their progeny. Our results show different outcomes obtained from using different density SNP panels in separating cow pooling samples, and estimating genomic breeding values for pregnancy test outcome of individual bull's progeny. The research highlights that extreme caution needs to be taken for choosing SNP panels of different densities to rank and select bulls for commercial beef production based on DNA pooling technology.

INTRODUCTION

Genomic prediction of breeding values based on a genomic relationship matrix has revolutionized the ability to identify genetically superior livestock for improving traits that are difficult to measure (van der Werf 2009). However, in commercial herds, it is impractical to individually genotype all animals. DNA pooling of cows with reproductive records can provide a cost-effective way for assessing and predicting the contribution of individual bulls to the fertility of their female offspring (Reverter et al, 2016). A question that remains to be answered is what density SNP panel should be used to genotype DNA pooled cows to rank bulls to achieve accurate prediction of reproductive performance of their progeny? In this study, using two Brahman cattle populations in north Queensland, we aimed to investigate the impact of SNP panels of different density on the ranking of bulls.

MATERIALS AND METHODS

Animals. Datasets from two Brahman cattle populations in north Queensland were used for the study. One (SmartF) consists of 290 samples from 2012-2014 herds (177 individual bulls and 113 pools representing 2,648 cows) genotyped with 74,584 SNPs (770K

BovineHD BeadChip platform). The other (MDH2020) contains 715 samples from the 2020 herd (482 individual bulls and 233 pools representing 2,452 cows) genotyped with 54,791 SNPs (Neogen Australasia GGP TropBeef 50K chip). DNA pools were formed based on the pregnancy test (i.e. not pregnant or pregnant) and lactation status (dry or wet) of cows at 2nd joining. Details of the phenotype of pregnancy test outcome (PTO) and pooling techniques can be found in Reverter et al. (2016). In brief, animals were separated into 6 categories, that is, dry and empty (not pregnant, scored as 1), dry and early pregnant (scored 2), dry and mid pregnant (scored 3), dry and late pregnant (scored 4), wet and empty (not pregnant, scored as 5), and wet and pregnant (scored 6). DNA samples of animals with identical phenotypic scores were pooled together. The individual pool size ranged from 4-45 animals for SmartF (Reverter et al., 2016) and from 5-12 animals for MDH2020, depending on the number of animals available in each category. Details of the two datasets are presented in Table 1.

Table 1. Composition of two genotyped populations

Population	Sex	Year	DNA samples	Total
SmartF	Cows	2012	41 (pools)	
(74,584 SNPs)		2013	31 (pools)	
		2014	41 (pools)	113
	Bulls	2013	27	
		2014	150	177
MDH2020	Cows	2020	233 (pools)	233
(54,791 SNPs)	Bulls	2020	482	482

Imputation of genotypic data. Between the two populations, there were 19,089 SNP in common. The imputation from low to high-density SNP genotypes was conducted to both SmartF and MDH2020, using 730,000 SNPs from 5,040 Beef CRC Brahman cattle as the reference. PLINK (Change et al. 2015) and Eagle v2.4.1 (Loh et al. 2016) were applied for phasing and imputation, respectively. After quality checks with the threshold of R-square value >0.8 and removing SNPs on the sex chromosome, this resulted in 615,310 SNPs.

Principal Component Analysis (PCA). To visualize genetic relationships between two populations, we conducted a PCA using genotypes from either the low density (19,089 common SNP) or imputed high-density panel (615,310 SNP, HD).

Genomic prediction. Genomic estimated breeding values (GEBVs) of PTO of progeny for individually genotyped bulls were derived within each population. The conventional genomic prediction method was applied to derive GEBVs, that is, a mixed animal model was used by fitting a polygenic random effect with the GRM (genomic relationship matrix). The fixed effects included the size of pool (30 levels) and contemporary group (5 levels) for SmartF, and SNP chip row (3 levels) and column (24 levels) information for different pools in MDH2020, respectively. The GRM was constructed using the method described by Reverter et al (2016). In brief, the B-allele frequencies from the genotypes of the pools of cows (≤0.25, >0.25 and <0.75 or ≥0.75, best fitted the three genotypes based on the individual DNA samples and the genotype call algorithm employed by Illumina) were converted into the three possible genotypes (i.e. 0, 1 and 2 for AA, AB, and BB, respectively) and these were merged with the individual genotypes of each bull to generate a single GRM relating bulls with pools of cows. Then the Qxpak5 software program (Pérez-Enciso and Misztal, 2011) was used to fit the GRM in a mixed animal model and obtain genomic estimates of variance components and genomic predictions (GEBVs) for PTO of the testing population. For comparison purposes of different density panels within populations, GEBVs were derived using four GRMs, either with 19,089, 54,791 (for MDH2020 only), 74,584 (for SmartF only), or high density (HD) SNP.

RESULTS AND DISCUSSION

Relationships between animals of two populations. The results from the PCA on all 1,005 animals (290 from SmartF and 715 from MDH2020) are shown in Figure 1. When a low-density SNP panel data (19,089, Figure 1a) was used, 346 DNA pooled cow samples from both populations were clustered together with very small variation among them, suggesting high similarity in the number of alleles between pooled samples. For the 659 individually genotyped bulls (red and blue dots), there was a much wider range of variation than for cows. However, when the high-density SNP panel was applied (HD, Figure 1b), there was a clear separation of cow samples of within and across two populations. But bulls remained mixed up as low-density results show, with a much narrower range of variation. This indicates that the bulls in the two populations had some degree of relatedness among themselves, but not among the cows. Therefore, the separation of pooled cows would not have been detected if the HD was not used.

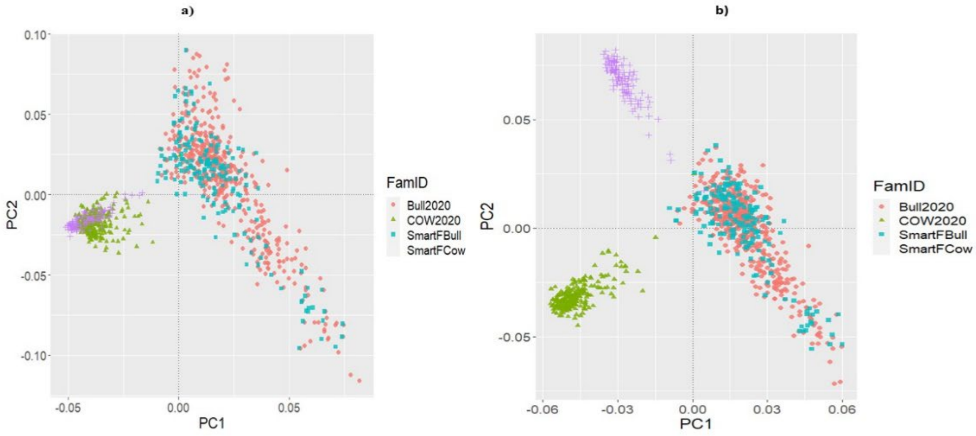


Figure 1. Principal Component Analysis of 1,005 genotyped samples, in which 482 were individually genotyped bulls (Bull2020), 233 were pools of cow DNA samples (COW2020), 177 were individually genotyped SmartF bulls (SmartFBull) and 113 were pools of SmartF cows (SmartFCow). a) 19,089 common SNP; b) High density SNP

Genomic predictions of bull's PTO with different panels of SNP density. Assuming the results from HD are true, Table 2 shows the Pearson's correlations among the PTO GEBVs from three SNP panels (19,089, 54,791 and HD) in the MDH2020 and SmartF respectively. Within MDH202, the correlations between GEBVs of PTO of 482 bulls were 0.74 between 19,089 and HD, and 0.82 between 54,791 and HD. The correlations were much lower (0.39 and 0.45 respectively) if only the top 25% bulls were considered (see Table 2 correlation for top 25%). Similar trends were observed in SmartF when the correlations of GEBVs for 177 bulls were compared (Table 2), despite slightly higher correlations between 19089 and 74584 with HD when the top 25% bulls were selected (0.54-0.59, Table 2). These suggest that if low-density panels were used to genotype pooled DNA cows for estimating the EBVs of PTO of bulls, at least 40-50% of the best bulls would not be selected.

When further investigating the bull GEBVs of PTO estimated using HD, Table 3 illustrates the profiles of the GEBVs of 482 MDH2020 bulls in different quartiles. The average GEBV difference between top and bottom 25% of bulls was 0.292**,** which is much larger than the difference obtained using low-density panels (0.120 from 19,089 or 0.158 from 54,791, results are not shown here). For animals being dry and empty (score 1) to become wet and pregnant (score 6), there could take conservatively up to 21-27 months to achieve. The GEBV difference of 0.292 from HD would translate into earlier conception by 1.31 months for the female progeny of the top 25% sires.

The study presents preliminary results for the comparison of different panels of SNP density in ranking commercial bulls in two populations. The phenotype score (1-6) of the 2nd joining pregnancy test outcome was treated as a continuous trait in which wet and non-pregnant was scored as “5” instead of “2”. Further research is underway to explore the impact of different score systems on ranking differences.

CONCLUSION

This research highlights the need for extreme caution to be taken when applying SNP panels of low or medium densities to study genetic relationships, and rank and select top bulls for commercial beef production based on DNA pooling technology.

Table 2. Pearson’s correlations among GEBVs estimated from using 19,089, 54,791 and HD SNP panels within MDH2020 and SmartF populations, respectively.

Population		MDH2020			SmartF		
	SNP	19089	54791	HD	19089	74584	HD
All bulls	19089	1	0.90	0.74	1	0.76	0.72
	54791 / 74584		1	0.82		1	0.81
	HD			1			1
Top 25%	19089	1	0.81	0.39	1	0.52	0.54
	54791 / 74584		1	0.45		1	0.59
	HD			1			1

Table 3. Average genomic breeding values (GEBVs) of progeny pregnancy testing outcome (PTO) of the MDH2020 bulls in four quartiles using HD SNP panel

Quartile	# Bulls	Av. GEBV	Min	Max
1 -Top 25%	120	0.136	0.0833	0.323
2	120	0.055	0.0275	0.0831
3	121	-0.004	-0.0341	0.0261
4 – Bot. 25%	121	-0.156	-0.2771	-0.0345
All	482	0.023	-0.277	0.323

ACKNOWLEDGEMENTS

The authors wish to acknowledge Meat and Livestock Australia (MLA) for their financial support to the project P.PSH.1211 “Validation of pooled DNA gEBV for Brahman commercial cow fertility”. We also like to thank Clint Smith, manager of Iffley Station of MDH Pty Ltd, for his great assistance in collecting samples and phenotypic information for the study.

REFERENCES

Chang C.C., Chow C.C., Tellier L.C.A.M., Vattikuti S., Purcell S.M. and Lee J.J. (2015) *GigaScience*, **4**.

Loh P.-R., Palamara P.F. and Price A.L. (2016). *Nat. Genet.* **48**: 811.

Reverter A., Porto-Neto L.R., Fortes M.R., McCulloch R., Lyons R.E., Moore S., Nicol D., Henshall J., and Lehnert S.A. (2016). *J Anim Sci*. **94**: 4096.

van der Werf J. H. J. (2009). *Proc. Assoc. Advmt. Anim. Breed. Genet.* **18**:38.

7.5 **2022 WCGALP Paper.**

GAPLS – A novel method to genomically rank bulls for daughter reproductive performance

Y. Li[1*], S. Lehnert[1], L. R. Porto-Neto[1], R. McCulloch[1], S. McWilliam[1], P. Alexandre[1], J. McDonald[2], C. Smith[2], and A. Reverter[1]

[1]CSIRO Agriculture & Food, 306 Carmody Rd., St. Lucia, Brisbane, QLD 4067, Australia; [2]MDH Pty Ltd, Cloncurry, QLD 4824, Australia; *yutao.li@csiro.au

Abstract

Female reproductive traits directly impact the profitability of commercial beef herds. The ability to select herd bulls, based on the predicted reproductive performance of their female progeny, has the potential to significantly improve herd productivity. Genomic selection in beef cattle run in the extensive production systems of tropical and subtropical northern Australia is hindered by technical and economic barriers, including the difficulty of obtaining accurate records on reproductive performance of individual females. In this study we present a novel method, Genomic Attribution to Pregnancy and Lactation Status (GAPLS), to predict an individual bull's contribution to six categories of their progeny's 2nd joining pregnancy and lactation status. Using a simulated example and four commercial Brahman datasets, we demonstrate the merit of the method.

Introduction

In extensive production systems of tropical and subtropical northern Australia, uptake of genomic selection has been limited due to practical, financial and management complications associated with collecting individual performance records. In northern commercial cattle herds, following natural syndicate joining, heifers are mustered and grouped based on the result of their 2nd joining pregnancy and lactation status (PLS). Heifers are commonly grouped into six categories in these situations: 1. DNP = Dry and Not Pregnant; 2.WNP = Wet and Not Pregnant; 3. DEP = Dry and Early Pregnant; 4. DMP = Dry and Mid Pregnant; 5. DLP = Dry and Late Pregnant; 6. WEP = Wet and Early Pregnant. When heifers are genotyped, either individually or pooled as a group based on these categories, their DNA profiles become a valuable resource to explore genomic selection strategies for the improvement of fertility traits in beef cattle (Reverter et al., 2016). However, the categorical nature of PLS coupled with departures from normal symmetry (e.g. PLS 3 and 4 not necessarily more abundant than 1 or 6) and the uncertainty about importance ranking of categories (e.g. PLS 2 not necessarily much worse than PLS 5), makes the analytical methodology to explore PLS rather cumbersome.

Here, we present Genomic Attribution to PLS (GAPLS), a novel and intuitive non-parametric proportion-based metric of genomic relatedness of a given individual animal (presumably a sire candidate for selection) across the six PLS categories.

Materials & Methods

The basic concept of Genomic Attribution to PLS (GAPLS). Using the genotypes from individual bulls and a reference population of cows with genotypes and phenotypes for PLS, a genomic relationship matrix (GRM) can be computed as described by VanRaden et al. (2008). The GAPLS of a tested bull is defined for each PLS category as its average genomic relationship with the cows having that PLS category divided by its average genomic relationship across all cows. Numerically, for the i^{th} bull, its GAPLS to the j^{th} PLS category (j=1, 2, …6, corresponding to DNP, WNP, … WE, respectively) is computed as follows:

$$\text{GAPLS}_{i,j} = \frac{\frac{1}{N_j}\sum_{k=1}^{k=N_C}[g_{i,k}\times \text{I}(k=j)]}{\sum_{j=1}^{j=6}\left(\frac{1}{N_j}\sum_{k=1}^{k=N_C}[g_{i,k}\times \text{I}(k=j)]\right)},$$

where: N_j = Number of cows in the j^{th} PLS category, N_C = Total number of cows in the reference population, $g_{i,k}$ = Genomic relationship between the i^{th} bull and the k^{th} cow; I(k=j) = An indicator that takes a value of 1 if the k^{th} cow belongs to the j$^{\text{th}}$ PLS category, and 0 otherwise. By construction, the GAPLS of the i^{th} bull summed across all 6 j^{th} PLS categories is one.

Simulated Example. For simple illustration purposes, consider a situation with 15 cows in the reference population and 2 candidate sires in the testing population (Table 1). The number of cows belonging to the PLS categories 1 to 6 is 2, 1, 3, 4, 1, and 4, respectively (second column of Table 1). Similarly, the genomic relationships between each cow and the hypothetical sires (Sire 1 and Sire 2) are given in the last two columns of Table 1.

Table 1. Hypothetical scenario with 15 cows with pregnancy and lactation status (PLS) and their genomic relationships ($g_{i,k}$) with two testing sires.

Cows		Sire's $g_{i,k}$	
K	PLS	Sire 1 (i = 1)	Sire 2 (i = 2)
1	6	0.5	0.3
2	3	0.2	0.0
3	4	0.2	0.0

4	1	0.1	0.1
5	6	0.1	0.4
6	3	0.1	0.1
7	6	0.0	0.3
8	1	0.0	0.0
9	4	0.0	0.1
10	5	0.0	0.2
11	2	0.5	0.2
12	4	0.2	0.1
13	6	0.2	0.4
14	3	0.1	0.0
15	4	0.1	0.2

In this scenario, the GAPLS of Sire 1 to PLS category 1 is:

$$\text{GAPLS}_{1,1} = \frac{\frac{1}{2}(0.1+0.0)}{\frac{1}{2}(0.1+0.0)+\frac{1}{1}(0.5)+\frac{1}{3}(0.2+0.1+0.1)+\frac{1}{4}(0.2+0.0+0.2+0.1)+\frac{1}{1}(0.0)+\frac{1}{4}(0.5+0.1+0.0+0.2)} = 0.0496.$$

Similarly, the GAPLS of Sire 2 to PLS category 6 is:

$$\text{GAPLS}_{2,6} = \frac{\frac{1}{4}(0.3+0.4+0.3+0.4)}{\frac{1}{2}(0.1+0.0)+\frac{1}{1}(0.2)+\frac{1}{3}(0.0+0.1+0.0)+\frac{1}{4}(0.0+0.1+0.2+0.2)+\frac{1}{1}(0.2)+\frac{1}{4}(0.3+0.4+0.3+0.4)} = 0.375.$$

Table 2 presents the average GAPLS values for both sires across the 6 PLS categories based on the original data and after 1,000 permutation tests of PLS. Since the first (1= DNP) and last two PLS categories (5=DLP and 6=WEP) indicate poor and good fertility, respectively, therefore Sire 2 would be preferred over Sire 1 because of its highest attribution value for category 6 (0.375). If using completely random data, i.e., a sire has no close relationship with any group of cows of six categories of PLS, the GAPLS of the sire to each PLS category is expected to equal 1/6 = 0.1667. Large deviations of GAPLS will be desirable when selecting sires because the PLS categories are unlikely to be equally represented in real scenarios. As shown in Table 2 (standard deviations in brackets), a robust measure of sampling variation can be quickly computed by permuting the PLS labels while maintaining the genomic relationships fixed.

Table 2. Average genomic attribution to pregnancy test outcome (GAPLS) values across the six PLS categories for the two sires after 1,000 permutations of six PLS values.

PLS Category	Sire 1 (STD[1])	Sire 2 (STD)
1_DNP	0.0496 (0.115)	0.0536 (0.096)
2_WNP	0.4959 (0.137)	0.2143 (0.125)

3_DEP	0.1322 (0.100)	0.0357 (0.082)
4_DMP	0.1240 (0.098)	0.1071 (0.081)
5_DLP	0.0000 (0.144)	0.2143 (0.126)
6_WEP	0.1983 (0.091)	0.3750 (0.079)

[1]STD -Standard deviation

Real data application. Datasets from four tropical Brahman cattle populations in north Queensland were used for the study. The first three populations contained 829 bulls, of which 114 were from the 2020 season (Bulls_114), 229 from 2021 (Bulls_229), and 486 are to be used in the 2022 mating season (Bulls_486). The fourth population consisted of 795 cows with PLS records. All samples were individually genotyped with 54,791 SNPs (Neogen Australasia GGP TropBeef 50K chip) and later imputed to a higher density of 529,260 autosomal SNPs. A GRM across all animals was constructed, and the GAPLS values of individual bulls were then estimated based on their genomic relationships with the 795 cows.

Results

Genomic relationships between bull and cow populations. The intensity difference in Figure 1A reveals that: 1) The animals in the Cows_795 set were more closely related among themselves than with the 829 bulls from the three bull populations; 2) The two bull populations (Bulls_114 and Bulls_229) had a closer relationship with each other than either had with the third bull population (Bulls_486); 3) The relationship between bulls and the cow population (Cows_795) was stronger for the Bulls_114 and Bulls_229 than for the Bulls_486.

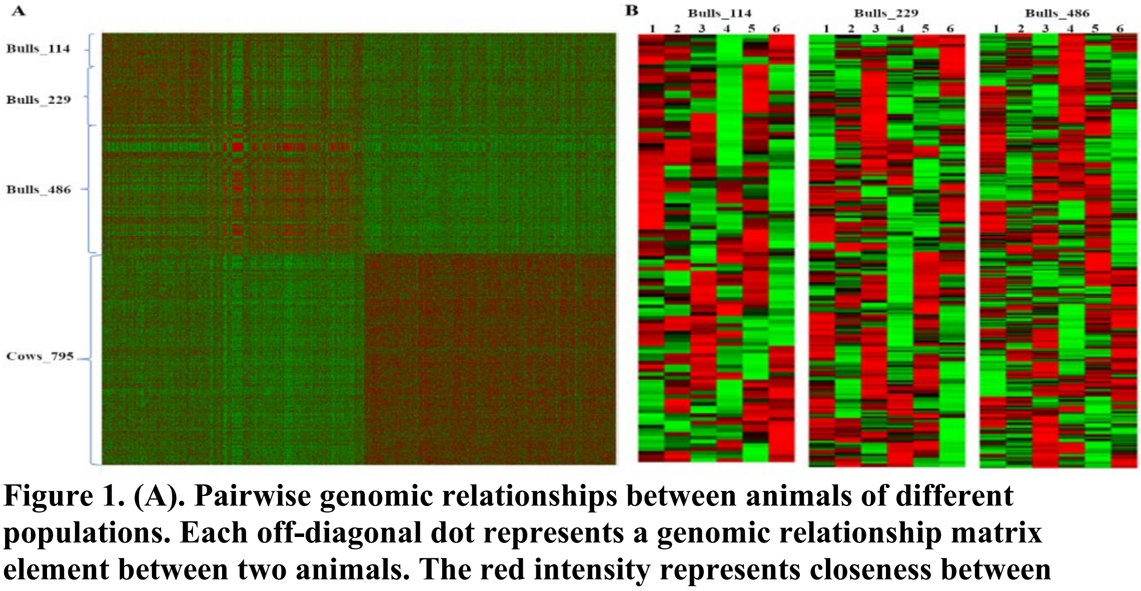


Figure 1. (A). Pairwise genomic relationships between animals of different populations. Each off-diagonal dot represents a genomic relationship matrix element between two animals. The red intensity represents closeness between

animals. (B). Heatmap of individual bull's GAPLS values across six categories within each bull population.

Table 3. Average values of bull's genomic attribution to pregnancy and lactation status (GAPLS) across the six PLS categories for three bull populations.

Population	Bulls_114	Bulls_229	Bulls_486
PLS Category	Mean (STD)	Mean (STD)	Mean (STD)
1_DNP	0.182 (0.0572)	0.170 (0.0521)	0.172 (0.0463)
2_WNP	0.169 (0.0451)	0.161 (0.0436)	0.159 (0.0299)
3_DEP	0.176 (0.0567)	0.194 (0.0676)	0.178 (0.0499)
4_DMP	0.139 (0.0636)	0.139 (0.0608)	0.159 (0.0514)
5_DLP	0.178 (0.0594)	0.171 (0.0626)	0.170 (0.0461)
6_WEP	0.156 (0.0709)	0.165 (0.0675)	0.163 (0.0513)

GAPLS estimates for three bull populations. The closer genomic relationships between the first two bull populations (Bulls_114 or Bulls_229) and the 795 cows, allowed for large variations in average GAPLS values (Table 3, columns 2 and 3) in these two bull populations (ranging from 0.139 to 0.182 and 0.139 to 0.194, respectively). In comparison, the bull population Bulls_486 had a much narrower range (0.159 – 0.172). Accordingly, the standard deviations (STDs) of the GAPLS in the first two bull populations were also consistently larger than those in the third bull population. When clustering bulls based on their GAPLS values (Figure 1B), it is easy to identify the bulls with distinguished value differences (color gradient) across the six categories. For example, selecting the desirable bulls with the highest GAPLS values for 6_WEP (red color) and the lowest values for 1_DNP (green color).

Discussion

Genomic prediction for a phenotype like PLS, with six arbitrary categories, is challenging because traditional GBLUP models are inadequate. Firstly, it is hard to interpret the biological significance of a single GEBV that is obtained after treating the six PLS categories as a continuous trait. Secondly, how to define the scoring system for the six categories is open for discussion. For instance, if categories are coded from worst to best for fertility, WNP could be coded as 2 or as 5 and this has a significant impact on GBLUP results (Li et al. 2021). The newly proposed GAPLS provides an effective way of assessing individual sire's genomic contributions to six categories of PLS by using the standard GRM values between sires and the cows with PLS records. The merits of the method include: 1) results are easy to interpret, promoting adoption by commercial producers wanting to improve the reproductive performance of their herds, 2) the GAPLS method will not be influenced by the phenotype scoring system allowing producers to tailor selection strategies to their production system; 3) The population variation of GAPLS values will truly reflect the degree of relatedness between testing and reference populations and, 4) The method could easily be extended to any phenotype of a categorical or ordinal nature.

Acknowledgments

The authors wish to acknowledge Meat and Livestock Australia (MLA) for their financial support to the project P.PSH.1211 "Validation of pooled DNA gEBV for Brahman commercial cow fertility", and Dr. Brad Hine for his valuable suggestions to improve the paper.

References

Li Y., Porto-Neto L., McCulloch R., McWilliam S., Alexandre P., *et al.* (2021). Proc. Assoc. Advmt. Anim. Breed. Genet. 24:204-207. 52Li24204.pdf (aaabg.org)
Reverter A., Porto-Neto L.R., Fortes M.R., McCulloch R., Lyons R.E., *et al.* (2016). J Anim Sci. 94(10):4096-4108. http://dx.doi.org/10.2527/jas.2016-0675
VanRaden, P.M, 2008. J. Dairy Sci. 91(11):4414-4423. https://doi.org/10.3168/jds.2007-0980

7.6 2023 AAABG paper

COMPARING GENOMIC PREDICTION ACCURACIES FOR COMMERCIAL COWS' REPRODUCTIVE PERFORMANCE USING GA2CAT AND TWO MACHINE LEARNING METHODS

Y. Li[1], S. Hu[1], L. Porto-Neto[1], R. McCulloch[1], S. McWilliam[1], J. McDonald[2], C. Smith[2], P. Alexandre[1], S. Lehnert[1], and A. Reverter[1]

[1]CSIRO Agriculture & Food, St Lucia, QLD, 4067 Australia
[2]MDH Pty Ltd, Cloncurry, QLD, Australia

SUMMARY

Heifers' second joining pregnancy and lactation status (PLS) is an important fertility trait for commercial cattle herds in North Queensland. Genomic prediction of a candidate bull's contribution to its female progeny's PLS presents a technical challenge because the trait has a non-ordinal multi-class nature. We have developed a new algorithm, Genomic Attributions to a Categorical Trait (GA2CAT) to tackle the problem. However, the merit of the method has not been evaluated against those of machine learning methods. In this study, using two commercial cow populations (795 and 340 cows respectively) with high-density SNP genotypes and imbalanced PLS phenotypes, we compared the classification performance of the new method GA2CAT with two machine learning approaches (Random Forests (RF) and Support Vector Machines (SVM)). The results from a five-fold cross-validation scheme indicate that the classification accuracy of GA2CAT was greatly impacted by the coding system of PLS categories. For highly imbalanced non-ordinal multiclass datasets, using the average overall accuracy value for evaluating the classification performance of the GA2CAT and ML methods was misleading and Matthews correlation coefficient values should be applied.

INTRODUCTION

Female reproductive traits directly impact the profitability of commercial beef herds. Among many reproductive traits, fertility-related ones are the most important. In dairy and beef cattle, they are measured by a range of continuous (e.g. age of puberty, days at first calving), binary (e.g. pregnancy status) or count traits (e.g. number of inseminations) (Toghiani *et al*. 2017). However, in Australian northern commercial cattle herds, following natural syndicate joining, heifers are usually mustered and grouped based on the result of their 2nd joining pregnancy and lactation status (PLS). Females can be assigned to six PLS categories: 1. DNP = Dry and Not Pregnant; 2.WNP = Wet and Not Pregnant; 3. DEP = Dry and Early Pregnant; 4. DMP = Dry and Mid Pregnant; 5. DLP = Dry and Late Pregnant; 6. WEP = Wet and Early Pregnant (Reverter *et al.*2016). This non-ordinal multi-class phenotype presents a technical challenge when trying to rank potential sires based on their genomic relationships with phenotyped heifers. To address this issue, we have developed a new method called Genomic Attributions to a Categorical Trait (GA2CAT) to predict an individual sire's contribution to its future daughters' performance (Li *et al.* 2022). However, the performance of GA2CAT has not been benchmarked against other methods commonly used for analysing non-ordinal multi-class traits, such as the machine learning (ML) based

Random Forests (RF) and Support Vector Machines (SVM). Therefore, we conducted the study to compare genomic prediction accuracies of GA2CAT and two ML methods.

MATERIALS AND METHODS

Datasets. Two datasets containing 1,135 tropical Brahman cows, 795 of 2020 season (referred as Cows_795) and 340 of 2021 season (Cows_340), from a north Queensland commercial property were used for the study. All animals with PLS records were individually genotyped for 54,791 SNPs (Neogen Australasia GGP TropBeef 50K chip) which were then imputed to the high density using 700K genotypes of 861 legacy BeefCRC Brahman cattle as the reference genome. Table 1 summarises the composition of the phenotype records in both populations, illustrating unevenly distributed multi-class categories.

Phenotypic data recoding. For comparison purposes, three different phenotype recoding systems for PLS records were investigated (Table 1). These include: a) treating PLS as a binary trait (2PLS, Non-pregnant "1" vs pregnant "2"); b) as a four-category trait (4PLS, Dry and Non-Pregnant "1", Wet and Non-Pregnant "2", Dry and Pregnant "3", and Wet and Pregnant "4"); and c) as a six- category trait (6PLS, see Table 1 for details).

Table 1. Composition of 2nd Joining Pregnancy and Lactation Status (PLS) records of two Brahman cow populations (795 and 340 cows respectively) and three phenotype recoding systems

		Cow population		Phenotype recoding system		
PLS	Code	Cows_795	Cows_340	2PLS*	4PLS*	6PLS*
Dry and Non-Pregnant	DNP	124	61	1	1	1
Wet and Non-Pregnant	WNP	358	109	1	2	2
Dry and Early Pregnant	DEP	77	109	2	3	3
Dry and Mid Pregnant	DMP	70	45	2	3	4
Dry and Late Pregnant	DLP	86	6	2	3	5
Wet and Early Pregnant	WEP	80	10	2	4	6
Total		795	340			

*2PLS: binary categories, 4PLS: four categories; 6PLS: 6 categories

Statistical methods. Three analytical methods were used for evaluating classification accuracy, including GA2CAT (Li et al. 2022), RF (Berriman, 2001) and SVM (James *et al.* 2013). In brief, the GA2CAT algorithm applies a standard genomic relationship matrix derived with the method of VanRaden (2008) between the reference and testing populations to predict the likely contributions of an individual animal in the testing population to individual classes of a categorical trait. For PLS, a GA2CAT value of a given animal for a given PLS category is defined as the animal's average genomic relationship with other animals having that PLS category divided by its average genomic relationship across all animals. RF is based on ensemble learning of a large number of decision trees deriving from randomly sampling of various subsets (both SNPs and animals) of a given dataset. It takes the average of decision trees (with replacement) to improve the predicted accuracy of the dataset. The final output (variable importance value) of RF is based on the majority votes of predictions. SVM applies different kernel functions (linear or non-linear) to identify a hyperplane that maximizes the separation of the data points to their potential classes (binary or multi-classes). While a genomic relationship matrix was used for deriving the GA2CAT values, both RF and SVM directly applied SNPs for the analyses.

A 5-fold cross-validation scheme was used for evaluating the classification performance of each method. Each cow population was randomly divided into 5 equal-size groups and each group (68 in Cows_340 or 159 animals in Cows_795) was in turn used as the validation set. Overall accuracy ((true positive +true negative)/(true positive + true Negative + false positive + false negative)) was used for evaluating the prediction performance. The final results were based on the average prediction accuracy of five validation groups. Given the imbalanced multiclass datasets used here, we also applied the Matthews correlation coefficient (MCC, Chicco and Jurman 2020) as a measure of the quality for multiclass classification. MCC values normally range from -1 to 1, with 1 representing a perfect prediction, 0 an average random prediction, and -1 a perfect misprediction.

Hyperparameter tuning for RF and SVM. A range of hyper-parameter values was examined for each ML method to determine the critical parameters that minimize prediction errors. These include: for RF, the size of forest trees (Ntree =100, 500), and the number of SNP markers at each sampling event (Mtry = 100, 500, 1000 and 5000); for SVM, insensitivity zone (gamma = 0.001, 1, 5, 10) and the penalty parameter (C= 0.001, 1, 10). All other parameters for each method took default values. The RF and SVM classifiers in the scikit-learn Python package (https://scikit-learn.org/stable/) were used for classification predictions.

RESULTS AND DISCUSSION

Comparison of classification performance of GA2CAT, RF and SVM. The overall average prediction accuracies (standard deviations in the brackets) of the three methods from a five-fold cross-validation scheme are summarised in the top part of Table 2. When changing the coding of PLS from two to four to six categories, the overall classification accuracy decreased significantly in both populations for all methods in the small population Cows_340, but much less extent in the big population Cows_795.

Table 2. Classification performance of GA2CAT, RF and SVM under different PLS coding systems in two cow populations, using a five-fold cross-validation scheme. A) The overall average classification accuracies (standard deviations in brackets); b) Matthews correlation coefficients (MCC)

A. Overall Accuracy	Cow population					
	Cows_340			Cows_795		
Method	GA2CAT	RF	SVM	GA2CAT	RF	SVM
2PLS	0.46 (0.097)	0.51 (0.073)	0.47 (0.032)	0.53 (0.027)	0.61 (0.029)	0.61 (0.033)
4PLS	0.18 (0.034)	0.43 (0.063)	0.47 (0.018)	0.24 (0.027)	0.44 (0.061)	0.45 (0.052)
6PLS	0.091 (0.024)	0.25 (0.054)	0.29 (0.034)	0.12 (0.025)	0.46 (0.052)	0.45 (0.052)
B. MCC	Cow population					
	Cows_340			Cows_795		
Method	GA2CAT	RF	SVM	GA2CAT	RF	SVM

2PLS	-0.071 (0.19)	0.020 (0.143)	0.000 (0.000)	0.059 (0.049)	0.077 (0.040)	0.000 (0.000)
4PLS	-0.039 (0.029)	-0.037 (0.036)	0.000 (0.000)	0.013 (0.026)	-0.027 (0.043)	0.000 (0.000)
6PLS	-0.017 (0.036)	-0.062 (0.073)	0.000 (0.000)	-0.023 (0.036)	0.053 (0.043)	0.000 (0.000)

RF: Random Forest; SVM: Support Vector Machine.

The poor performance of the three methods under 6PLS could be due to the phenotype of PLS being a non-ordinal multi-class categorical trait. The separation of animals for three Dry and Pregnant classes, i.e. early, mid, and late pregnancy, was not clean-cut as those in the binary situation (2PLS, non-pregnant vs pregnant). For the GA2CAT, the genomic relationships between animals in these three classes in the training populations were very similar, therefore the predicted contributions of the animals in the validation populations to six categories of PLS (i.e. GA2CAT values) were very similar. As result, it made the correct assignment of the animals in the testing populations to different categories extremely difficult. The results indicate the necessity of recoding PLS records before applying different analytical methods to achieve reliable results.

Across two cow populations, for the same coding system, e.g. 6 categories (6PLS), the two ML methods (RF and SVM) seemed to outperform the GA2CAT (see the average accuracies in Table 2). The margin was huge in the population Cows_795 (0.46 (RF), 0.45 (SVM) vs 0.12 (GA2CAT). The difference between RF and SVM was little in comparison to either of them with the GA2CAT. However, when investigating further on the classes correctly classified, we found that both RF and SVM assigned all of the individuals in the validation datasets to the category of Wet and Non-Pregnant. This was the class with the largest number of phenotypic observations in Cows_795. This confirms the downside of ML methods that bias toward the majority class by over-sampling the abundant classes and under-sampling minor classes (Chicco and Jurman 2020).

When evaluating the performance of three methods by the MCC values (the bottom part of Table 2), all three methods had the MCC values either zero (SVM) or close to zero. These suggest that: a) the phenotype PLS is a low heritability trait, as all three methods followed a random prediction behavior (MCC values ~ 0.00). In addition, the accuracy values for the GA2CAT fitted the random sampling expected prediction accuracies of 0.5 (PLS2), 0.34 (PLS4) and 0.25 (PLS6); b) there was no significant classification performance difference among the GA2CAT, RF and SVM.

CONCLUSION

The results from a five-fold cross-validation scheme indicate that different coding systems of PLS categories greatly impacted the classification outcome of the GA2CAT. For highly imbalanced non-ordinal multiclass datasets, using the average overall accuracy value for evaluating the classification performance of the GA2CAT and ML methods was misleading and MCC values should be applied. A GA2CAT value is the weighted average of genomic relationships between reference and validation populations for a particular category, it reflects better the heritability nature of a phenotypic trait.

ACKNOWLEDGEMENT

The authors wish to acknowledge Meat and Livestock Australia (MLA) for their financial support to the project P.PSH.1211 “Validation of pooled DNA gEBV for Brahman commercial cow fertility”.

REFERENCES

Breiman, L. (2001). *Mach. Learn*. **45**: 5.
Chicco, D., and Jurman, G. (2020). *BMC genomics*, *21*, 1-13.
James G., Witten D., Hastie T., and Tobshirani (2013). ‘An Introduction to Statistical Learning’. Springer, Heidelberg, Germany.
Li Y., Lehnert S.A., Porto-Neto L., McCulloch R., McWilliam S., Alexandre P, McDonald J., Smith C., and Reverter A. (2022). Proceedings of 12th World Congress on Genetics Applied to Livestock Production. Rotterdam. The Netherlands. 3-8 July 2022.
Reverter A., Porto-Neto L.R., Fortes M.R., McCulloch R., Lyons R.E., Moore S., Nicol D., Henshall J. and Lehnert S.A (2016). *J Anim Sci*. **94**(10)**:**4096.
VanRaden, P.M (2008). *J. Dairy Sci*. **91**(11)**:**4414.