milestone report

Project code: P.PSH.1154

Prepared by: Sonja Dominik, Toni Reverter and Laercio Porto-Neto
CSIRO Agriculture and Food

Date published: 20/12/2019

PUBLISHED BY
Meat and Livestock Australia Limited
Locked Bag 1961
NORTH SYDNEY NSW 2059

Cost effective DNA pooling strategies to drive genetic gain in the livestock industries

Milestone Number 3 – Final report

This is an MLA Donor Company funded project.

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

Abstract

DNA pooling could provide a cost effective strategy to obtain GEBV for sires based on their commercial progeny performance. The aim of this study was to compare genomic breeding values (GEBV) for sires estimated from individual genotypes to GEBV estimated from pooled DNA samples of the sires' progeny. A phenotype dataset from 2,436 Angus cattle from 174 sires was assembled for yearling weight (YWT), coat score (COAT) and MSA marbling score (MARB). All animals had genotypes for 35,009 SNP. Two pooling strategies were tested, pooling by sorted phenotype and pooling at random. Within each strategy pool of sizes of 2, 5, 10, 15 20 and 25 were explored. We conclude that pools of 10 DNA samples based on phenotype were identified as representing a good compromise between loss of accuracy (~10-15%) and cost savings (~90%) from genotype assays.

Table of contents

1 Milestone description

Final report covering project objectives to be submitted to MLA for review and approval.

2 Project objectives

This project will evaluate the value proposition for the application of DNA pooling in Australian livestock production systems, by exploring the balance between accuracy and cost-effectiveness of different strategies to assess sire performance on commercial farms.

The project outcomes include

- Detailed cost benefit analysis of DNA pooling strategies compared to individual genotyping
- Genomic breeding values for sires and sire rankings for traits of interest based on in silico pooling of genotypes
- Evaluation of pooling strategies for different traits of relevance for production and adaptation
- Validation of the sire ranking from DNA pooling approaches against the ranking from genomic breeding values based on individual genotypes.

3 Methodology

3.1 Data

Data for this study was obtained from the Angus Australia Breed Society. The data has been collected as part of the Angus Sire Benchmarking project, jointly funded by Meat & Livestock Australia Donor Company (MDC) and Angus Australia Breed Society. It includes phenotypes, genotypes and fixed effect information on around 264 sires and 5,000 progeny. Data used for this study is a subset. Phenotype data includes records for three traits of interest to this project.

3.1.1 Phenotypes

Trait 1 – Yearling weight (YWT): a continuous trait with moderate heritability, easy to measure, available on all animals and achieves useful accuracies in genomic predictions.

Trait 2 – Coat score (COAT): categorical trait with high relevance for heat tolerance in beef cattle, as well as a moderate to high heritability.

Trait 3 – MSA marbling score (MARB): difficult to measure of high economic importance and likely to benefit from genomic approaches.

3.1.2 Genotypes

We assembled a dataset of 2,610 Angus cattle from 174 sires averaging 14.02 progeny per sire and ranging from 2 progeny (sires 133300825 and 133417140) to 36 progeny (sire 133537894).

The original data set supplied by Angus Australia contained 3,921 genotyped animals (38,661 SNPs). This set of SNPs was quality control assessed with a threshold minor allele frequency of >0.05. A further quality control step included the removal of problematic data points such as duplicates.

In the quality controlled data, all animals had genotypes for 35,009 SNP that were used to build a genomic relationship matrix (GRM). The resulting GRM was compared against the pedigree-based numerator relationship matrix (NRM) to ensure correctness. The diagonal and off-diagonal elements of the GRM are summarised in Table 1.

Table 1: Summary statistics for (off)diagonal elements of the GRM.

	N	Mean	SD[A]	Min.	Max.
Diagonals	2,610	0.988	0.0324	0.827	1.129
Off-Diagonals	3,404,745	0.001	0.0335	-0.100	1.027

[A]The (essentially) identical variance of relationships within and across animals is an important characteristic of an optimal relationship matrix.

The summary statistics for the GRM shown in Table 1 are as expected with values close to 1 representing an animal matched to itself (diagonal elements) and relationships to other animals represented in the off-diagonals closer to 0. Importantly, a consistent variation in the relationships both within and between animals was observed (*ie*. SD ~ 0.03) indicating a homogeneous genetic (co)variance expected from a single-breed analysis.

3.2 Analysis

3.2.1 Genetic parameter estimation

Genetic parameters and GEBV were estimated from a tri-variate mixed model in Qxpak v.5.05 (Perez-Enciso and Misztal, 2011). A simple model was fitted with adjustment of phenotypes only for contemporary group and sex (concatenated) and age at measurement. All animals (N=2,610 = 2,436 progeny + 174 sires) had genotypes for 35,009 SNP that were used to build a genomic relationship matrix (GRM) and the random additive genetic effect was fitted in the model based on the GRM.

3.2.2 Pooling strategies and process

Only data from sires with suitable sizes of progeny groups (accuracy of gEBV > 50%) and contemporary groups will be retained for analysis. Phenotypes were adjusted for fixed effects and age at measurement and pools of size 2, 5, 10, 15 20 and 25 were explored where pools were created either sorting by phenotype, depending on the trait on a continuous or categorical scale, or at random. When pools were created at random, 10 replicates were examined to provide a measure of sampling variation. In a real experiment, blood of animals from the same pool would be combined to obtain DNA pools. Here, DNA pooling will be simulated by "pooling" genotypes of pools in silico. Once animals were assigned to pools, a genotype for the pools was created based on the genotypes of the animals in each pool based on B-allele frequencies (Bell et al. 2014; Alexandre et al. 2019). As a result, a 'hybrid' GRM was built using the individual genotypes from the sires and the merged genotypes from the pools. Pools of DNA samples were created.

3.2.3 Benefit Cost analysis

With a fixed number of individuals, here 2,026, different pooling strategies can be applied, creating different number of pools by manipulating the number of progeny per pool. Different pooling strategies were tested for each of the traits. The more pools, the smaller the number of contributing DNA samples and the closer the resemblance to individual genotyping. However, the downside is that the more pools are created, the higher the cost. Genomic breeding values for sires and their accuracies estimated based on individual progeny genotypes (individual gEBV), individual genotypes are the most costly option. To establish the number of pools and therefore genotypes required, the number of progeny was divided by the pool size. We assumed current price for 50K of ~ $30 which was multiplied by the number of genotypes required

4 Success in meeting the milestone

4.1 Preliminary analysis prior to pooling

We assembled a dataset of 2,436 Angus cattle from 174 sires averaging 14.02 progeny per sire and ranging from 2 progeny (sires 133300825 and 133417140) to 36 progeny (sire 133537894).

Three phenotypes were explored including yearling weight (YWT; N = 1,589 records), coat score (COAT; N = 2,026 records) and MSA marbling score (MARB; N = 1,304 records) with the following summary statistics:

Table 2: Summary statistics for phenotypes used in this analysis.

Phenotype		N	Mean	SD	Min.	Max.
YWT	Age, d	1,589	430.55	28.33	357	496
	Weight, kg	1,589	398.03	77.42	183	692
COAT	Age, d	2,026	591.72	79.15	436	1,038
	Score	2,026	2.38	0.80	1	5
MARB	Age, d	1,304	779.67	96.05	501	990
	Score	1,304	512.76	119.97	160	1,030

A tri-variate GREML analysis was undertaken for all 3 traits. Estimates of heritabilities, correlations (genetic and residual) and GEBV and accuracies are provided in the following Tables 3-5. The GEBVs for sires will be used as the benchmark when assessing the performance of the various pooling strategies.

Table 3: Heritabilites (diagonals, bold), genetic (upper diagonal) and residual (lower) correlations.

	YWT	COAT	MARB
YWT	**0.5675**	-0.0328	0.0471
COAT	-0.0599	**0.4198**	-0.0190
MARB	-0.0151	-0.0533	**0.4902**

Moderate to relatively high estimates of heritabilities were obtained at 56.7%, 42.0% and 49.0% for YWT, COAT and MARB, respectively; while estimates of genetic correlation were close to zero across all pair-wise traits. Of particular relevance are the COAT GEBVs as this trait has been flagged by Angus Australia as “a trait of importance, particularly for the adaptability of Angus genetics in hotter, more tropical environments” with Research Breeding Values (RBV) recently published in their website (https://www.angusaustralia.com.au/content/uploads/2019/12/Full-Report_Coat-Type_November-2019.pdf). Their report indicates the use of ~5,000 measurements (compared to our ~2,000) and a heritability of 0.25 (compared to our 0.42). On closer examination, the correlation between RBV published by Angus Australia and the GEBV computed here was 0.83 adding confidence to the results of this study.

Table 4: Summary statistics for GEBVs by phenotype and for sires and progeny.

	YWT		COAT		MARB	
	SIRES	**PROGENY**	**SIRES**	**PROGENY**	**SIRES**	**PROGENY**
N	174	2,436	174	2,436	174	2,436
Mean	9.91	1.86	-0.061	-0.007	18.91	2.57
SD	18.56	17.74	0.251	0.228	64.98	51.56
Min.	-48.71	-76.18	-0.789	-0.831	-137.70	-174.92
Max.	59.43	58.87	0.708	0.954	224.85	217.08

Table 5: Summary statistics for GEBV Accuracies by phenotype and for sires and progeny.

	YWT		COAT		MARB	
	SIRES	**PROGENY**	**SIRES**	**PROGENY**	**SIRES**	**PROGENY**
N	174	2,436	174	2,436	174	2,436
Mean	0.768	0.541	0.780	0.531	0.735	0.466
SD	0.076	0.121	0.040	0.056	0.089	0.126
Min.	0.547	0.257	0.547	0.282	0.481	0.270
Max.	0.884	0.706	0.867	0.637	0.836	0.661

The accuracies for GEBV are as expected and higher for sires than progeny. The GEBV and accuracies will provide the benchmark for comparison with the resulting GEBV and accuracies from the pooling strategies. Across traits, GEBV accuracies ranged from 48% to 89% for sires and from 27% to 71% for progeny. For the sires and as expected, there was a strong correlation (r = 0.856) between number of progeny and GEBV accuracies.

4.2 Performance of pooling strategies

Two different pooling strategies were investigated 1) pooling by phenotype (ByPheno) and random pooling (ByRandom). Each strategy was conducted by pooling 2, 5, 10, 15, 20 or 25 genotypes. Fig. 1 demonstrates that for all three traits, categorical and continuous, pooling by phenotype performs best, with an impressive correlation of r=0.8 between the GEBV and the pooled GEBV. Random pools allow the computation of sire GEBVs that are moderately correlated (i.e., r > 0.5 at pool sizes ≤ 10) with those obtained without pooling. Overall it can be observed that with increasing number of DNA samples in the pool the correlation between GEBV and pooled GEBV decreases.

Fig. 2 shows the strong (pooling by phenotype) and moderate (pooling at random) correlation in both cases, but the vastly reduced range of GEBVs when the pools are made at random, as can be seen when the scale of the y-axes are compared. This reduction is even more accentuated for larger random pools.

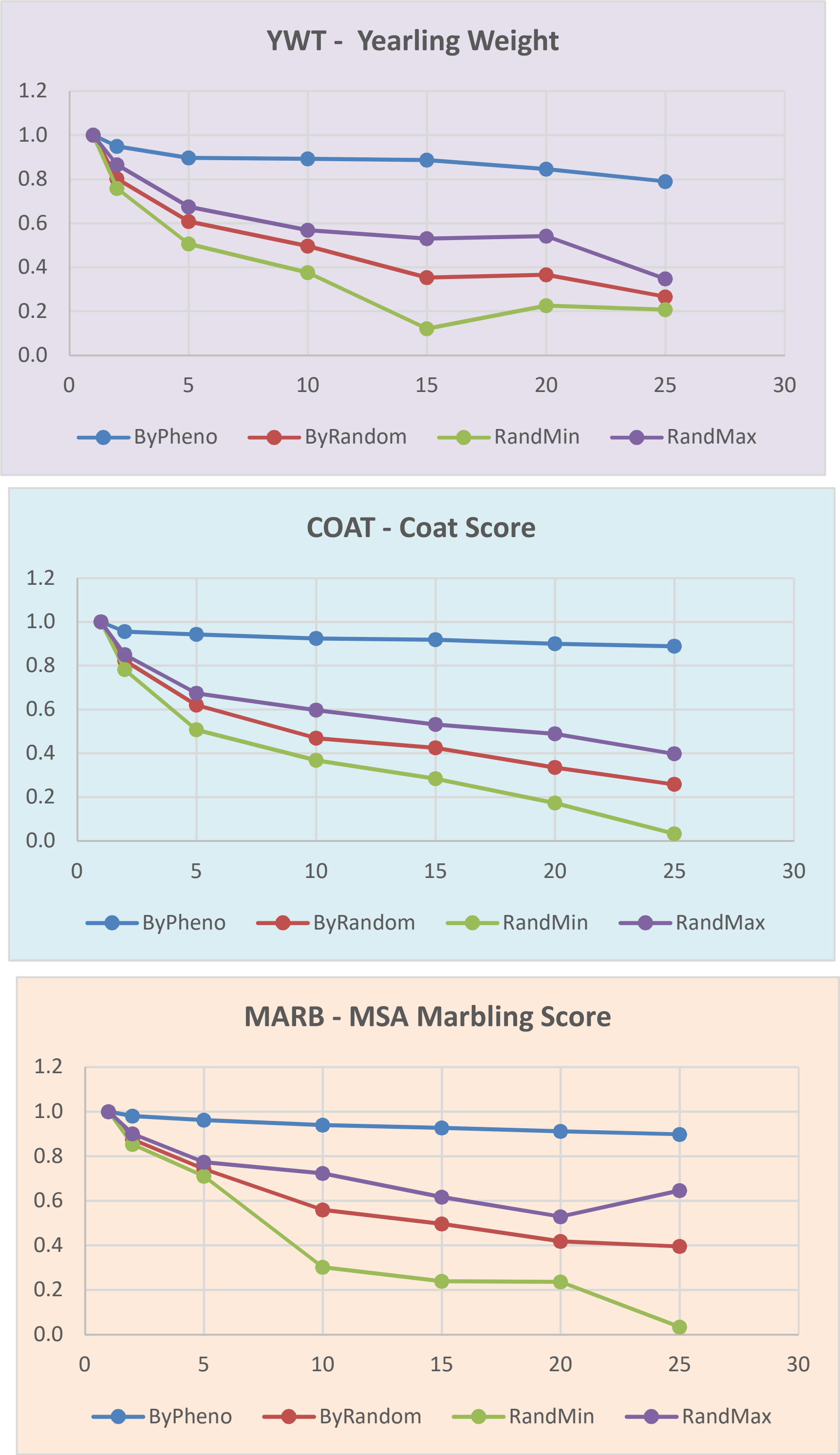


Fig. 1: Correlation between Sire GEBV based on DNA pools of progeny of various sizes from 1 (no pooling) to 2, 5, 10, 15, 20 and 25 for Yearling Weight (top), Coat score (middle) and MSA marbling score (bottom) and when pooling was either based on phenotype (blue trend) or at random with average (red), minimum (green) and maximum (purple) of 10 random replicates.

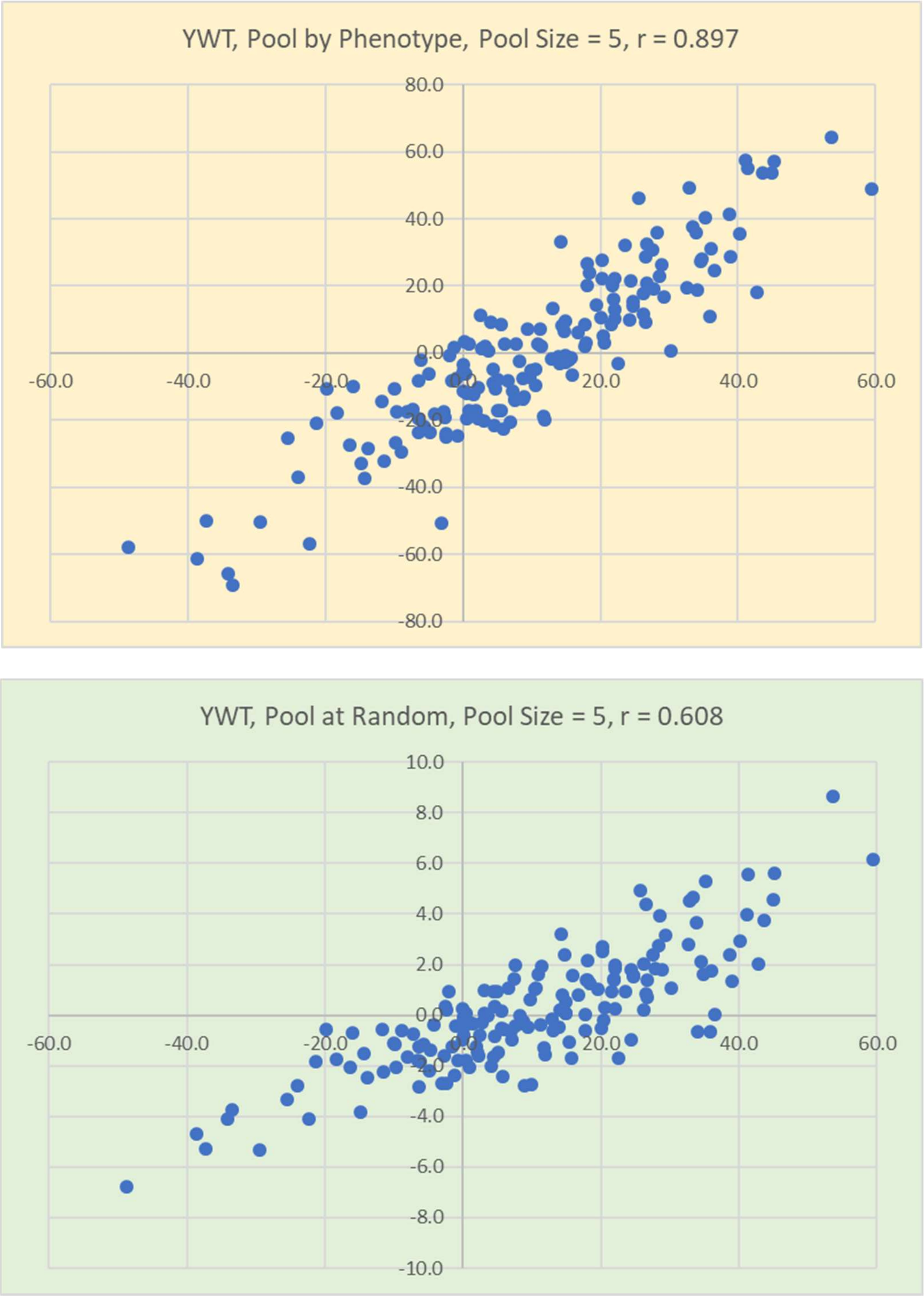
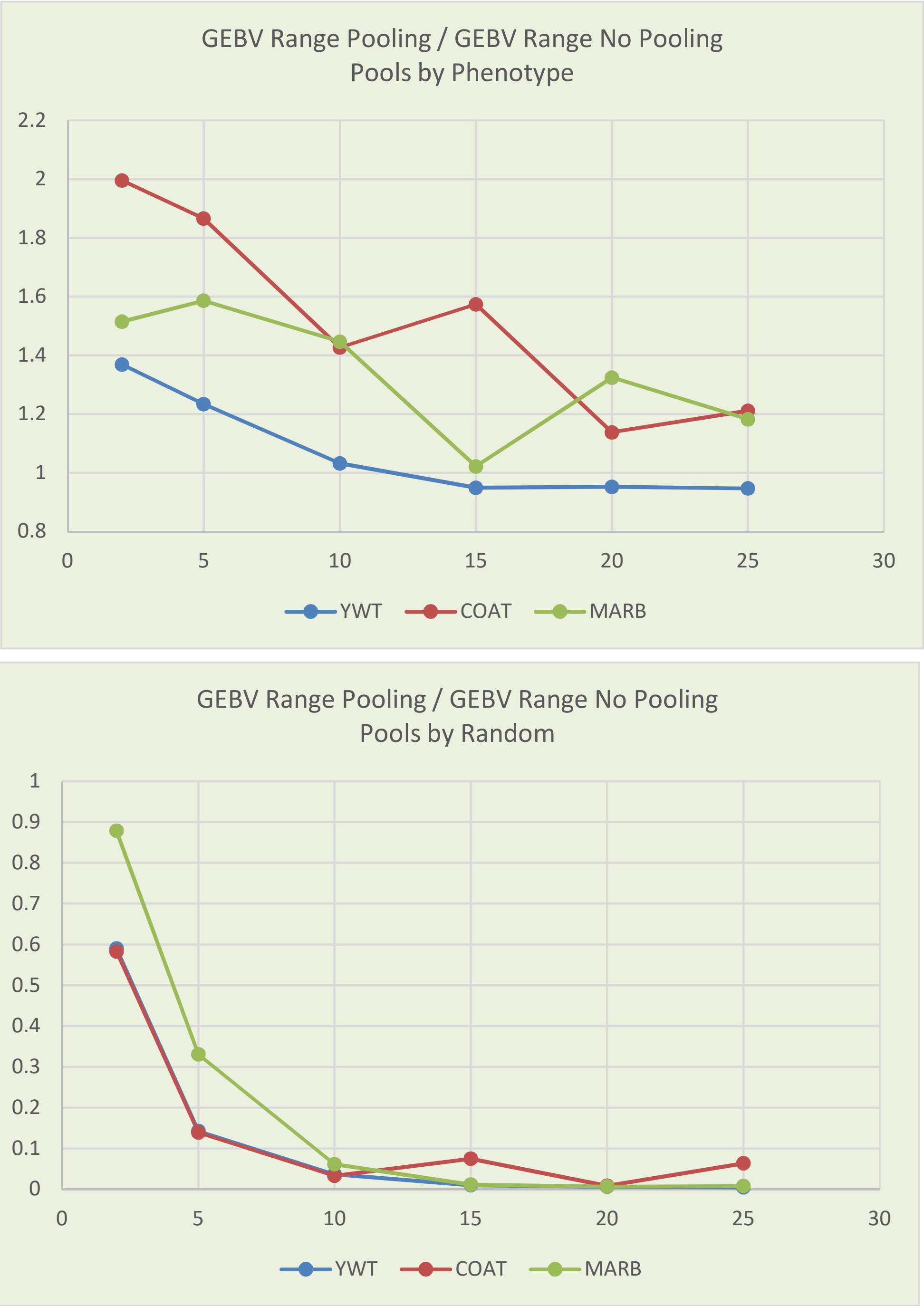


Fig. 2: Relationship between sire GEBV for yearling weight (YWT) GEBV without pooling (x-axis) and GEBV after pooling (y-axis) with a pool size of 5 and either by phenotype (top panel) or at random (bottom panel).

In the top panel of Fig. 3 it is demonstrated that DNA pooling by phenotype resulted in GEBV ranges wider than without pooling particularly with small pools (pool sizes ≤ 10). The opposite is true when pools were made at random: the resulting GEBV were narrower in range than those without pooling, particularly for large pools (pool sizes > 10; Fig. 3, bottom panel).


GEBV Range Pooling / GEBV Range No Pooling
Pools by Phenotype

GEBV Range Pooling / GEBV Range No Pooling
Pools by Random

Fig. 3: Ratio of GEBV range based on DNA pooling over EBV range without pooling when the pools are based on phenotype (top panel) or at random (bottom panel) and for three traits: Yearling weight (YWT, blue profile), coat score (red profile) and MSA marbling score (green profile).

Table 6: Pooling by Phenotype: Estimates of environmental (Ve), genetic (Vg) and phenotypic (Vp) and heritability (h2) at various pooling sizes and for the three traits.

Trait	Pool Size[A]	Ve	Vg	Vp	h^2	Exp(Vp)[B]	Vp Inflation
YWT	**1**	485.54	637.03	1,122.57	0.567	1,122.57	1.000
YWT	**2**	61.93	1,664.29	1,726.21	0.964	1,587.55	1.538
YWT	**5**	45.52	2,418.35	2,463.87	0.982	2,510.14	2.195
YWT	**10**	38.99	2,819.11	2,858.10	0.986	3,549.88	2.546
YWT	**15**	34.21	3,334.50	3,368.71	0.990	4,347.69	3.001
YWT	**20**	33.16	3,759.94	3,793.10	0.991	5,020.28	3.379
YWT	**25**	32.56	4,204.91	4,237.47	0.992	5,612.85	3.775
COAT	**1**	0.17	0.12	0.29	0.420	0.29	1.000
COAT	**2**	0.02	0.44	0.46	0.965	0.40	1.610
COAT	**5**	0.01	0.66	0.67	0.981	0.64	2.358
COAT	**10**	0.01	0.78	0.79	0.986	0.90	2.762
COAT	**15**	0.01	0.94	0.95	0.989	1.10	3.319
COAT	**20**	0.01	1.05	1.06	0.989	1.28	3.721
COAT	**25**	0.01	1.02	1.03	0.990	1.43	3.609
MARB	**1**	5,638.33	5,421.23	11,059.57	0.490	11,059.57	1.000
MARB	**2**	1,544.40	15,743.79	17,288.18	0.911	15,640.59	1.563
MARB	**5**	1,320.65	22,246.38	23,567.03	0.944	24,729.94	2.131
MARB	**10**	1,123.31	27,202.23	28,325.54	0.960	34,973.42	2.561
MARB	**15**	866.08	29,694.40	30,560.48	0.972	42,833.51	2.763
MARB	**20**	826.93	38,636.06	39,462.99	0.979	49,459.88	3.568
MARB	**25**	839.81	43,368.48	44,208.29	0.981	55,297.83	3.997

[A] Pool Size = 1 for no pooling.

[B] Exp(Vp) = Expected Vp based on Vp times sqrt(*n*), where *n* is the pool size.

DNA pooling by phenotype resulted in an inflation of the estimate of phenotypic variance (Vp) resulting from an overestimate of the genetic variance and h^2 ($h^2 > 0.95$ in all pool sizes and traits). This overestimation was attributed to pools being more phenotypically consistent than individual observations and likely to capture relatives. The observed inflation in Vp was in-line with what could be the expected based Vp without pooling times the square root of the size of the pool. This approximation was reasonably accurate up until pools of size 10. Beyond that, the estimated Vp was less than the expected and attributed to variance within pools becoming relatively much smaller than the variance between pools.

Table 7: Pooling at Random[A]: Estimates of environmental (Ve), genetic (Vg) and phenotypic (Vp) and heritability (h2) at various pooling sizes and for the three traits.

Trait	Pool Size[B]	Ve	Vg	Vp	h^2	Exp(Vp)[C]	Vp Deflation
YWT	1	485.54	637.03	1,122.57	0.567	1,122.57	1.000
YWT	2	197.32	518.26	715.58	0.724	561.28	0.637
YWT	5	149.60	145.00	294.59	0.492	224.51	0.262
YWT	10	91.25	45.02	136.27	0.330	112.26	0.121
YWT	15	62.56	22.47	85.03	0.264	74.84	0.076
YWT	20	46.72	13.85	60.57	0.229	56.13	0.054
YWT	25	38.10	17.03	55.13	0.309	44.90	0.049
COAT	1	0.17	0.12	0.29	0.420	0.29	1.000
COAT	2	0.08	0.09	0.17	0.516	0.14	0.601
COAT	5	0.04	0.04	0.08	0.507	0.06	0.277
COAT	10	0.02	0.01	0.04	0.330	0.03	0.125
COAT	15	0.01	0.03	0.04	0.732	0.02	0.144
COAT	20	0.01	0.01	0.02	0.293	0.01	0.067
COAT	25	0.01	0.02	0.03	0.759	0.01	0.102
MARB	1	5,638.33	5,421.23	11,059.57	0.490	11,059.57	1.000
MARB	2	340.23	9000.22	9340.45	0.964	5,529.78	0.845
MARB	5	763.39	4009.96	4773.35	0.840	2,211.91	0.432
MARB	10	855.68	768.10	1623.77	0.473	1,105.96	0.147
MARB	15	697.45	229.36	926.81	0.247	737.30	0.084
MARB	20	542.75	229.35	772.10	0.297	552.98	0.070
MARB	25	492.05	223.27	715.32	0.312	442.38	0.065

[A] Values in table represent averages across 10 random replicates

[B] Pool Size = 1 for no pooling.

[C] Exp(Vp) = Expected Vp based on Vp devided by *n*, where *n* is the pool size.

DNA pooling at random resulted in a deflation of the estimate of phenotypic variance (Vp) resulting from an underestimate of the residual and genetic variances, while h^2 was overestimated at pool size of 2 and 5 (except yearling weight) and rapidly decreased at pool sizes ≥ 10. The deflation of Vp could be approximated by dividing the estimate of Vp without pooling by the size of the pool. This approximation worked particularly well at pool size ≥ 5.

4.3 Benefit cost of pooling strategies

The cost of genotyping was assessed for each strategy. Creating the same number of pools genotyping based on phenotype or randoml pooling costs the same. However, it must be considered that pooling by phenotype is associated with increased labour and requires a structured sampling and/or pooling process. When pools are formed by phenotype, blood samples must be identified with the specific phenotype (e.g. coat score 3) at the time of sampling or samples are “thrown” into a container for coat score 3. Such a process is more difficult if the phenotype is continuous and the phenotype must be written on the label of the blood sample. When pools are formed at random, blood tubes do not need to be labelled at sampling and blood samples can easily be pooled at any stage in the process just based on the required number of samples in the pool.

Table 8: Cost of genotyping for various pool sizes.

Pool size	Number of pools and/or required genotypes	Cost of genotyping	% reduction in genotyping by pooling
1	2,026	60,780	--
2	1,013	30,390	50%
5	405	12,125	80%
10	202	6,060	90%
15	135	4,050	93%
20	101	3,030	95%
25	81	2,430	96%

Pooling reduces the cost of genotyping .The cost of pooled genotyping as percentage of the cost for individual genotyping equates to

% reduction in genotyping by pooling = 100-[(1/pool size)*100]

In our data set, 2,026 individual genotypes would have cost $60,780. Pools of 10 DNA samples were identified as representing a good compromise between loss of accuracy (~10-15%) and cost savings (~90%) from genotype assays. Pooling by phenotype is the best approach to implementing genomic evaluation using commercial herd data, particularly when pools of 10 individuals are evaluated. However, the increased logistics of the pooling by phenotype needs to be considered.

5 Conclusions/recommendations

The results from the study using real data are consistent with findings using simulated data. We conclude that pools of 10 individuals were identified as representing a good compromise between loss of accuracy (~10-15%) and cost savings (~90%) from genotype assays. In particular, pooling by phenotype is a very useful and cost effective approach to implementing genomic evaluation using commercial herd data, particularly when pools of 10 individuals are evaluated.

This study was initiated with the aim of exploring DNA pooling as a suitable tool to use commercial data on characteristics that are relevant to heat tolerance in Australian Angus, such as coat score in combination with a production trait. DNA pooling could provide a cost effective strategy to obtain GEBV for Angus sires for heat tolerance based on their commercial progeny performance.

6 Bibliography

Alexandre PA, Porto-Neto LR, Karaman E, Lehnert SA, Reverter A (2019) Pooled genotyping strategies for the rapid construction of genomic reference populations. *Journal of Animal Science* **pii**: skz344. doi: 10.1093/jas/skz344. https://www.ncbi.nlm.nih.gov/pubmed/31710679

Bell A, Henshall J, McCulloch R and Kijas J (2014) Evaluating sires from commercial progeny data using pooled DNA. *Proceedings 10thWorld Congress of Genetics Applied to Livestock Production*. Online.

Miguel Pérez-Enciso M, Misztal I (2011) Qxpak.5: Old mixed model solutions for new genomics problems. *BMC Bioinformatics* **12**: 202. doi: 10.1186/1471-2105-12-202