A conditional multi-trait sequence GWAS of heifer fertility in tropically adapted beef cattle

M. Forutan[1*], B. Engle[1], M.E. Goddard[2], B.J. Hayes[1]

[1] Centre for Animal Science, Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, Brisbane, QLD 4072, Australia; [2]Agriculture Victoria, AgriBio, Centre for AgriBioscience, Bundoora, Victoria, Australia; [*] m.forutan@uq.edu.au

Abstract

Reproductive efficiency is major driver of profitability of beef production in northern Australia. Failure of heifers to achieve pregnancy results in a considerable loss to producers and accounts for most of the reproductive costs incurred. Availability of genomic information has provided a noteworthy opportunity to enhance the efficiency of selection, especially for traits such as fertility with low heritability. Here, we used imputed whole genome sequence data and up to 2,119 Brahman heifers in a stepwise conditional multi-trait GWAS (CM-GWAS) to identify pleiotropic putative causal variants and genes associated with fertility traits. We identified multiple genomic regions affecting fertility traits, some previously reported and some novel. Candidate gene findings included the genes *ARHGEF28, RNF150*, *EPHA6***.** When we investigated the genes closest to the most significant SNP, they included *CNTNAP4* gene that encodes a member of the neurexin protein family and *PLAG1* gene.

Introduction

Over the past decade, genome-wide association studies (GWAS) have been widely used to identify associations of single-nucleotide polymorphisms (SNPs) with complex traits in livestock, especially in dairy and beef cattle (Visscher *et al.* 2017). Using whole genome sequence data can improve the power of GWAS since the causal variants should be included in the sequence data. However, even using whole genome sequence data, the identification of the causal variants for a complex trait is still difficult. This is due to the small effect size of most causal variants and the linkage disequilibrium (LD) between variants. Consequently, there are too many variants in high LD, any one of which could be the cause of the variation in phenotype. Causal variants are often pleiotropic*, i.e.* affecting more than one trait, so multi-trait analysis might result in higher power to detect quantitative trait loci (QTL) and greater precision in mapping them (Bolormaa *et al.* 2014).

In this study, the stepwise conditional multi-trait GWAS (CM-GWAS) analysis was performed to identify significant variants and genes associated with four fertility-related traits, including a binary trait (heifer pregnancy status) and three continuously distributed traits including fetal age in weeks, measured via manual palpation at pregnancy diagnosis; heifer age at first calving, defined as the number of days between the birthdate and calving date; and days to calving, defined as the number of days between the date of bull turn out at the beginning of the breeding season and calving date.

Materials & Methods

Phenotypic Data

For this study, lifetime productivity of a Central Queensland Brahman cow herd was assessed. Born between 1981 and 2015, these cows and heifers were part of a stud herd that has been developed with a heavy emphasis on fertility, where failure to produce a calf was the primary culling criterion. Four fertility-related traits were considered. In these herds, heifer pregnancy status was recorded as a binary trait (1 = successful, 0 = unsuccessful) indicating whether a heifer was able to conceive prior to three years of age. Pregnancy success was determined at time of yearly pregnancy check for all heifers born in 2011 and later. For heifers born prior to 2011, calving records were used to determine pregnancy success. Among those heifers that had both a pregnancy test and calving record available, 6% experienced pregnancy loss after pregnancy check, making calving success a good approximation in cases where pregnancy check records were unavailable. Fetal age in weeks (hef_wks_preg) was recorded via manual palpation at pregnancy diagnosis for all heifers born in 2011 and later. Age at first calving (AFC) was only available for heifers with a recorded birth date and was calculated as the difference in days between first calving and birth. Days to calving (DTC) is a routinely recorded trait in Australia Brahmans and is defined as the number of days between the date of bull turn out at the beginning of the breeding season and calving date. Heifer days to calving was recorded as the number of days between first calving and bull exposure.

Genotyping and Quality Control

Heifers were genotyped with the BovineSNP50 BeadChip (Illumina, San Diego, CA). A detailed description of the genotype quality control was given by Hayes *et al.* (2019). Genotypes were imputed up to 728,785 SNPs (Bovine HD array) using the findhap4 software (VanRaden *et al.* 2013), and a panel of 4650 individuals from relevant breeds, including Brahman (300), Droughtmaster (300), Santa Gertrudis (250) and composites (1000) that were genotyped with the Bovine HD array. All genotypes were then imputed to 31,140,417 million whole-genome sequence variants using the 1000 Bull Genomes Run8, TaurIndicus reference (Hayes and Daetwyler 2019), with 600 Holsteins and 400 Simmental animals removed to avoid over-representation of these genomes in the imputation. Eagle (Loh *et al.* 2016) was used for phasing and Minimac3 software (Das *et al.* 2016) for imputation.

Genome-Wide Association Analysis

A linear mixed model was performed using the GCTA software (Yang *et al.* 2011), fitting each sequence variant as a covariate, one at a time, and testing for association with each trait as follows:

$$\boldsymbol{y} = \boldsymbol{1_n}\mu + \boldsymbol{X}\beta + \boldsymbol{Z}g + \boldsymbol{W_i}\alpha_i + \boldsymbol{e},$$

where **y** is the vector of phenotypic values of the animals, $\boldsymbol{1_n}$ is an n × 1 vector of 1s (n=number of animals with phenotypes), μ is the overall mean, **X** is an n × x matrix of fixed covariates, β is a length x vector of fixed effects, **Z** is a design matrix for the random additive genetic effects; g is a vector of random additive genetic effects assumed to be distributed as ~N(0, $\mathbf{G}\sigma_g^2$), where **G** is the genomic relationship matrix (**GRM**) calculated from high-density genotypes using the GCTA software. $\boldsymbol{W_i}$ is a vector of genotypes for each animal at the i-th variant, α_i is the corresponding variant effect, and **e** is a random vector of length n as ~ N (0, $\sigma_e^2\mathbf{I}$), where σ_e^2 represents non-genetic variance due to non-genetic effects assumed to be acting independently on animals. The choice of fixed covariate effect for continuous and binary traits was done using *lm* and *glm* function in R, respectively. For all four traits, year of birth and contemporary group were constantly considered as fixed covariates. Moreover, for trait AFC, calving success defined as 0 and 1, based upon whether they joined at two or three years of age was considered as fixed effect. Also, for trait DTC, the effect of heifer age of joining was fitted as continuous covariate fixed effect. The GRM was generated using variants with minor allele frequency (MAF) higher than 0.01 (609,878 SNPs) in the HD dataset.

We performed a conditional multi-trait meta-analysis (CM-GWAS) according to a previously described approach (Bolormaa *et al.* 2021) using whole genome sequence (WGS) variant effects estimated from four single-trait GWAS to identify pleiotropic variants that affected fertility trait. The multi-trait meta-analysis (M-GWAS) X^2 statistic with 4 degrees of freedom (equal to the number of traits analysed) was calculated as below:

$X^2 = \mathbf{t}_i\mathbf{V}^{-1}\mathbf{t}_i$,

where t_i is a vector of the signed t-values of the effects of the i-th sequence variants for the 4 traits and V^{-1} is the inverse of the 4 × 4 correlation matrix where the correlation was calculated over all estimated sequence variants effects (signed t-values) between each pair of traits. The CM-GWAS approach cycles back and forward between the single-trait GWAS for all traits and M-GWAS to re-test variants conditional on jointly fitting the most significant putative causal variants from independent QTL (where we defined significant as $P < 5\times10^{-6}$). To determine independent sequence variants, first the most significant M-GWAS variant from each chromosome was selected and added to the list of putative causal variants. If the pairwise LD between this variant and any other significant variant on the same chromosome was greater than 0.1, these other variants are considered as potentially tagging the same causal variant and were not considered as independent QTL for this cycle. Then, from the remaining significant variants in LD, r^2 less than 0.1, the next most significant variant was selected on each chromosome, LD was tested between these and the remaining significant variants and so on, until no more significant variants identified in this cycle.

Results and Discussion

There was no indication of inflation of the test statistic due to population structure or other confounding effects for any of the traits (Figure 1B). To be able to identify independent putative causal variants we used the CM-GWAS. We selected only the most significant ('top') variants from each cycle of the CM-GWAS and identified 144 independent putative causal variants across the genome (P<5×10-8). Candidate genes included the genes *ARHGEF28, RNF150*, *EPHA6* located in intronic region of chromosomes 20, 17, and 1, respectively. When we investigated the genes closest to the most significant SNP, they included *CNTNAP4* gene that encodes a member of the neurexin protein family. Members of this family function in the vertebrate nervous system as cell adhesion molecules and receptors. Moreover, we discovered the same mutation as the one Bouwman *et al.* (2018) identified in a very large meta-analysis of height in cattle using imputed sequence data on more than 50k animals in *PLAG1* gene (14: 23300304). In this study, new potential causative mutations affecting fertility traits have been identified that should be incorporated into commercial SNP arrays. This should increase the accuracy of genomic breeding values (GEBV) for these traits, just as a similar approach has increased accuracy of GEBV in sheep and dairy cattle (Moghaddar *et al.* 2019, Xiang *et al.* 2019).

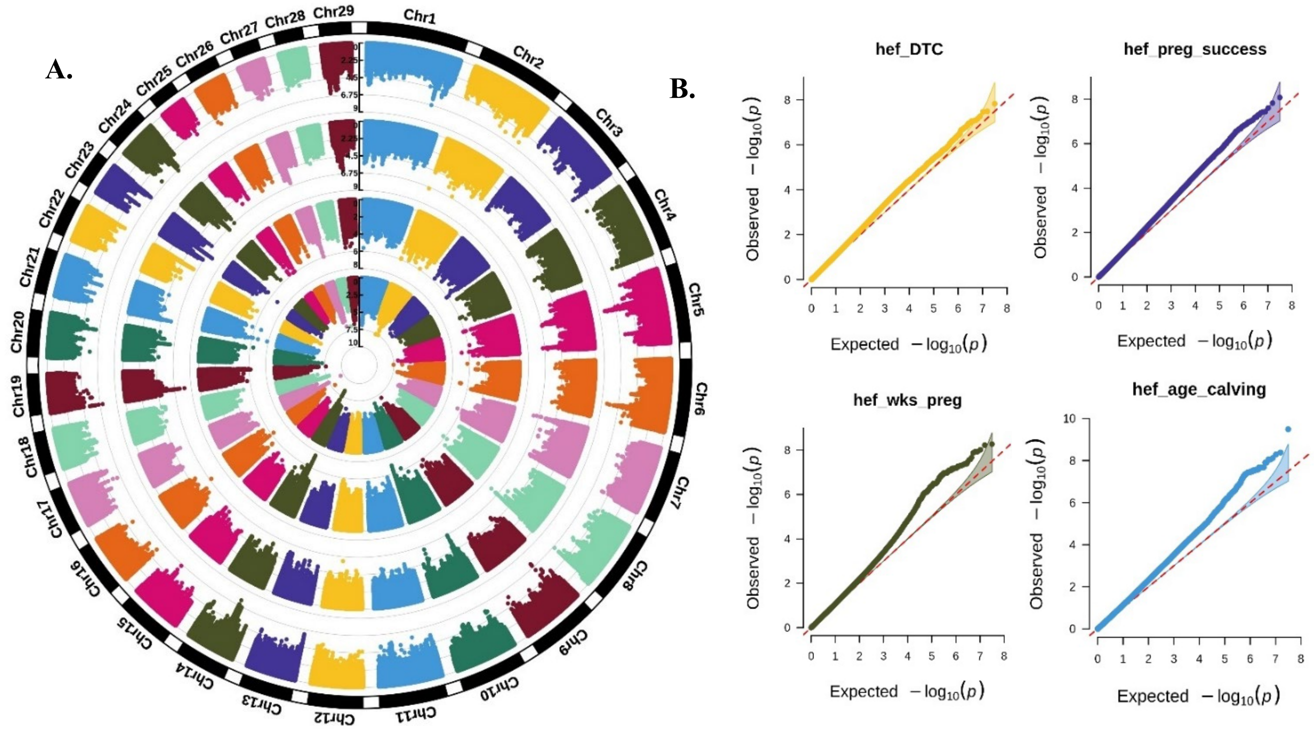

Figure 1. A. Manhattan plot of single-trait GWAS analysis for four fertility traits of heifer day to calving(her_DTC), heifer pregnancy success (hef_preg_success), heifer weeks pregnant (hef_wks_preg) and heifer age of calving (hef_age_calving) in Bos indicus cows. **B.** Quantile-quantile (QQ) plot of the single-trait GWAS shown in the Manhattan plot for 4 fertility traits.

References

Bolormaa S., Pryce J.E., Reverter A., Zhang Y., and Barendse W., et al. (2014) PLoS Genet 10(3): e1004198.

Bolormaa S., Swan A.A., Stothard P. et al. (2021) Genet Sel Evol 53, 58.

Bouwman A.C., and Daetwyler H.D. et al. (2018) Nat Genet. 50(3):362-367.

Das S., Forer L., Schönherr S., Sidore C., and Locke A.E., et al. (2016) Nat Genet. 48:1284–7.

Hayes B.J., and Daetwyler H.D. (2019) Annual review of animal biosciences 7: 89-102.

Hayes B.J., Corbet N.J., Allen J.M., Laing A.R., and Fordyce G. et al. (2019) J Anim Sci. 97:55–62.

Loh P.R., Danecek P., Palamara P.F., Fuchsberger C., and Reshef Y.A. et al. (2016) Nat Genet. 48:1443–8.

Moghaddar N., Khansefid M., van der Werf J.H.J, Bolormaa S., and Duijvesteijn N., et al. (2019) Genet Sel Evol. 5:51(1):72.

VanRaden P.M., Null D.J., Sargolzaei M., Wiggans G.R., and Tooker M.E., et al. (2013) J Dairy Sci. 96(1):668-78.

Visscher P.M., et al. (2017) Am. J. Hum. Genet. 101, 5–22.

Xiang R., Berg I.V.D., MacLeod I.M., Hayes B.J., Prowse-Wilkins C.P., et al. (2019) PNAS 116(39):19398-19408.

Yang J., Lee S.H., Goddard M.E., and Visscher P.M. (2011) Am. J. Hum. Genet. 88:76-82.