REGIONS IMPACTING INBREEDING DEPRESSION AND THEIR ASSOCIATION WITH ADDITIVE GENETIC EFFECTS FOR JERSEY CATTLE FROM THE UNITED STATES OF AMERICA AND AUSTRALIA

J. T. Howard[1], J. E. Pryce[2,3], M. Haile-Mariam[2], C. Maltecca[1]

[1]Department of Animal Science, North Carolina State University, Raleigh, NC 27695-7627, USA
[2]Department of Economic Development, Jobs, Transport and Resources and Dairy Futures Cooperative Research Centre, 5 Ring Road, Bundoora Victoria, 3083, Australia
[3]La Trobe University, Bundoora, Victoria 3086, Australia

SUMMARY

The advent of dense genotype platforms allow for the identification of specific regions that give rise to inbreeding depression and characterize their relationship with the additive effect of that region. Utilizing a run of homozygosity (ROH) metric the first study objective is to identify regions having an impact on inbreeding depression for United States (US) and Australia (AU) Jersey cows. The second objective is to determine the relationship between the additive and ROH SNP effects. Genotyped cows with phenotypes on milk yield traits (US=6751; AU=3974) and calving interval (US=5816; AU=3905) were utilized. A ROH based metric (ROH4Mb) was calculated across the genome. Residuals from a model that accounted for the fixed and additive genetic effects were regressed on ROH4Mb using a single marker regression or a machine-learning tree based model. The relationship between ROH4Mb and additive effect was characterized based on sliding window (500kb) direct genomic value derived from a Bayesian-LASSO analysis. Genomic regions across multiple traits were found to be associated with ROH4Mb for the US on BTA13, BTA23 and BTA25 and AU on BTA3, BTA7 and BTA17. Multiple potential epistatic interactions were characterized. The covariance between ROH4Mb and the additive effect depended on the genomic region.

INTRODUCTION

High levels of inbreeding result in a reduction in fitness and overall performance at the phenotypic level (Leroy, 2014). Previous research has shown that inbreeding depression is heterogeneous across founders (Gulisija *et al.* 2006), which implies that the genetic load is not distributed evenly among genomes. Utilizing a run of homozygosity (ROH) metric, Pryce *et al.* (2014) confirmed heterogeneity in inbreeding depression by identifying multiple regions that resulted in reduction in milk yield and fertility traits in Australian (AU) Holstein and Jersey cows. Furthermore, regions with multiplicative effects, which individually may have a minor effect but which have significant impact on fitness when combined, might provide clues about the previously identified non-linear relationship of inbreeding depression (Gulisija *et al.* 2007).

Characterizing regions that give rise to inbreeding depression in dairy cattle is advantageous due the increasingly large number of genotyped cows and the extensive list of recorded phenotypes. It has been shown by Howard *et al.* (2015) that ROH frequency differs across the Australian (AU) and United States (US) Jersey populations, which could potentially give rise to different regions that have an impact on inbreeding depression. Also, Howard *et al.* (2015) found that regions of the genome with high ROH frequency are most likely the result of directional selection. Here we hypothesize that the covariance between the additive effect and ROH status of a SNP is variable across the genome, and characterizing this may provide clues about the relationship between the two metrics. Therefore the

objectives are: 1. Identify regions that have an impact on inbreeding depression for US and AU Jersey cows utilizing ROH metrics; 2. Determine the relationship between the additive and ROH status effect of a SNP.

MATERIALS AND METHODS

Data. Two populations of genotyped cows born in the US and AU were utilized to identify regions that when homozygous cause a reduction in dairy yield traits including milk (MY), fat (FY) and protein (PY; n = 6751 US; n = 3974 AU) and fertility measured as calving interval (FERT; n = 5816 US; n = 3905 AU). Phenotypic information for the AU population was provided in the form of yield deviations. In order to make comparison similar, yield deviations were also constructed for the US population based on the same model outlined by Howard *et al.* (2015). A total of 32,431 QC SNP were used for the analysis. The ROH status of a SNP was defined based on whether the SNP was within an ROH of at least 4 Mb in length (ROH4Mb; 1 if the SNP was in a ROH and 0 otherwise).

Genome wide association study (GWAS). A two-stage analysis was performed within each population. Stage one involved generating residuals from an animal model that accounted for the additive genetic effects. The second stage involved using the residuals from the first stage as a phenotype to regress on ROH4Mb status. For this analysis two models were utilized. Model 1 was a single marker regression model, which regressed the phenotype on the ROH4Mb status for each SNP. Significance (P-value < 0.001) was declared by using a permutation test (n= 2,500 samples). The second model utilized a machine-learning tree based regression algorithm, referred to as a gradient boosted machine (GBM), to identify higher order ROH interactions. The "gbm" R package (Ridgeway 2010) was used to carry out the analysis. Based on a 4-fold cross-validation the final model was constructed from 1200 trees at an interaction depth of 5 and a shrinkage parameter of 0.0075. Within each chromosome, SNP with a correlation exceeding 0.1, as outlined by Lubke *et al.* (2013), were removed and only SNP with the largest impact based on single marker regression analysis were included in the final analysis. The final number of SNP utilized was 115 and 81 for the US dataset and 100 and 105 for the AU dataset for yield traits and fertility, respectively. The variable importance value (Ridgeway 2010) was used as measure to assess the importance of a ROH4Mb status of a SNP on a given phenotype.

The identification of epistatic interactions between the ROH4Mb status of a SNP was carried out by counting the number of times two SNP were a descendent pair, as described by Yao *et al.* (2013). Briefly, based on Figure 1, assume SNP B and D have a large epistatic interaction. The SNP pairs are represented based on the levels at which they appear, such that SNP D was derived from a split produced by SNP B, and therefore the combination represents a parent (SNP B) and child (SNP D) descendent pair. The SNPs B and D will appear more frequently in the same branch of a tree due to the pair having an epistatic interaction. Lower level descendent pairs such as parent (SNP A) grandchild (SNP D) will also be referred to as a descendent pair. The identification of SNP with independent effects, such as SNP B and C won't be tagged as descendent pairs due to SNP B and C being on separate branches. The significance (P-value < 0.001) of the frequency of a descendent pair and variable importance value was declared based on a permutation test (n= 2,500 samples).

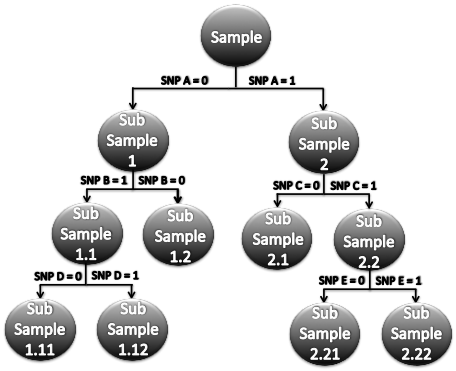

Figure 1. An example of a regression tree generated by a Gradient Boosted Machine algorithm.

Additive and ROH4Mb relationship. The second objective of the study was to characterize the relationship between the additive and ROH4Mb effect of a SNP. Both the additive and ROH4Mb marker values were obtained within each population and trait using a Bayesian LASSO marker regression (Park and Casella 2008). Estimates of the additive effect of a SNP were obtained using yield deviations as phenotypes. Estimates of the ROH4Mb effect of a SNP were obtained using the same phenotype as in the two-stage approach. The LASSO analysis was performed using the 'BLR' package in R (de los Campos *et al.* 2013). To characterize the relationship between the additive genetic and ROH4Mb SNP effect across the genome, 500 kb overlapping windows were used to estimate the GEBV (co)variance for a given window for both analyses.

RESULTS AND DISCUSSION

Within a population, regions that had an effect across multiple traits included BTA13 (19.3-19.9 Mb; MY-PY), BTA23 (32.7-33.3; MY-FY-PY) and BTA25 (24.8-30.7; MY-PY) for the US population and BTA3 (113.4-114.6; FY-PY), BTA7 (6.6-16.7; FY-PY) and BTA17 (68.9-75.0; MY-FY-PY) for the AU population. In both countries none of these regions had an effect on calving interval. Strikingly, no regions were identified that were significant in both populations. Both models ranked regions comparably with a rank correlation of 0.48 to 0.65 across all traits and populations. A list of the regions and candidate genes is outlined in Table 1.

Multiple regions of the genome were found to display potential interactions based on their frequency of being a descendent pair. The majority of all significant descendent pairs were associated with at least one SNP that also had a large variable importance score. Additionally, a gene network analysis revealed network associations including shared protein domain (HNF1B-LBX2) genetic interactions (HNF1B-LOXL3; HNF1B-RPL17). Additional work will be needed to validate regions found across both the single marker regression and GBM analyses in other populations.

The relationship between the additive genetic effect and the ROH4Mb effect displayed positive and negative covariance across the genome. Regions on BTA3, BTA7, BTA20 and BTA26 had a positive covariance between the additive and ROH4Mb effect of a SNP across both populations, and

have previously been found to be under positive directional selection (Howard *et al.* 2015). Homozygosity at certain regions of the genome is beneficial such that homozygosity based on the ROH4Mb status gives rise to a higher additive genomic estimated breeding value. Conversely, homozygosity at certain regions is unfavorable giving rise to lower additive genomic estimated breeding values. The majority of the regions with the largest absolute covariance value across traits were positive, which is not surprising due to a low frequency of ROH4Mb status for regions with a large ROH4Mb effect (mean ROH4Mb frequency = 0.089) in comparison to the regions that displayed a large positive covariance (mean ROH4Mb frequency = 0.235). These results provide evidence that it is possible to distinguish between two individuals that have the same inbreeding coefficient, but different overall genomic loads. This would have in turn important consequences in the management of genomic diversity and the implementations of effective mating design.

Table 1. Regions of the genome associated with inbreeding depression for milk traits.

Country	Traits	BTA (Region Mb)	Frequency	Candidate Gene
United States	MY-PY	13 (19.3-19.9)	0.10	PARD3
	MY-FY-PY	23 (32.7-33.3)	0.18	ALDH5A1
	MY-PY	25 (24.8-30.7)	0.05	IL4R, CALN1
Australia	FY-PY	3 (113.4-114.6)	0.06	UGT1A1
	FY-PY	7 (6.6-16.7)	0.17	NOTCH3
	MY-FY-PY	17 (68.9-75.0)	0.04	IGLL1

REFERENCES

de Los Campos G., Pérez P., Vazquez A.I. and Crossa J. (2013) *Methods Mol Biol* **1019**: 299.

Gulisija, D., Gianola D. and Weigel K.A. (2006) *Livest. Sci.* **104**: 244.

Gulisija, D., Gianola D. and Weigel K.A. (2007) *J. Dairy Sci.* **90**: 493.

Howard J.T., Maltecca C., Haile-Mariam M., Hayes B.J. and Pryce J.E. (2014) *BMC Genomics* **16(1)**: 1352.

Leroy, G. (2014) *Anim. Genet.* **45**: 618.

Lubke G.H., Laurin C., Walters R., Eriksson N., Hysi P., Spector T.D., Montgomery G.W., Martin N.G., Medland S.E. and Boomsma D.I. (2013*) J Data Mining Genomics Proteomics* **4**:10.4172/2153-0602.1000143.

Park T. and Casella G. (2008) *J. Am. Statist. Assoc.* **103**: 681.

Pryce, JE, Haile-Mariam M., Goddard M.E. and Hayes B.J. (2014) *Genet. Sel. Evol.* **46**: 71.

Ridgeway G. (2010) *R package version* 1.6-3.1.

Yao C., Spurlock D.M., Armentano L.E., Page Jr C.D., Van de Haar M.J., Bickhart D.M. and Weigel K.A. (2013) *J. Dairy Sci.* **96**:6716.