Genome-wide association study and gene network analysis of fertility, retained placenta, and metritis in US Holstein cattle

J.B. Cole[1], K.L. Parker Gaddis[2], D.J. Null[1], C. Maltecca[3], and J.S. Clay[4]

[1] Animal Genomics and Improvement Laboratory, ARS, USDA, Bldg 005, BARC-West, 10300 Baltimore Avenue, Beltsville, MD 20705
[2] Council on Dairy Cattle Breeding, 4201 Northview Drive, Bowie, MD 20716, USA
[3] Department of Animal Science, College of Agriculture and Life Sciences, North Carolina State University, Campus Box 7621, Raleigh, NC 27695
[4] Dairy Records Management Systems, 313 Chapanoke Road, Suite 100, Raleigh, NC 27603

Summary

The objectives of this research were to identify genes, genomic regions, and gene networks associated with three measures of fertility (daughter pregnancy rate, DPR; heifer conception rate, HCR; and cow conception rate, CCR) and two measures of reproductive health (metritis, METR; and retained placenta, RETP) in US Holstein using producer-reported data. A five-trait mixed model analysis was used to perform a genome-wide association study (GWAS) to identify significant SNP located within 25 kbp of genes in bull and cow predictor populations. Gene ontology (GO) and medical subject heading (MeSH) analyses were used to identify pathways and processes over-represented compared to a background set of all annotated *Bos taurus* genes. An adaptive weight matrix was used to identify significant associations among genes. GWAS results identified different sets of SNP in the two predictor populations, with SNP affecting protein processing, cell-cell signaling, sex differentiation, and embryonic development. Significant GO and MeSH terms also differed between predictor populations, but terms associated with reproductive processes were identified in both cases. The degree of nodes in the network analysis did not deviate from expectations, but fertility-related terms were identified, and several of the most-connected genes were associated with male or female fertility and embryo size and morphology in mice or humans, most notably *ITPR1*, *SETB1*, *LMNB1*, *NEO1*, and *DGKA*. None of the 100 SNP explaining the most variance in the GWAS were among the connected genes in the networks. While this study identified genes and interactions among them clearly related to fertility, no obvious associations with peripartum reproductive health were found. A more powerful experimental design, such as a case-control study, may be needed to identify relationships among fertility and reproductive tract health.

Keywords: dairy cattle, fertility, health, reproduction

Introduction

Parker Gaddis *et al.* (2016) recently used single- and multiple-trait genome-wide association studies (GWAS) in all-bull, all-cow, and mixed predictor populations to dissect three fertility traits. Their results showed that gene network analysis was able to identify several important genes that were not identified by ordinary GWAS. The US will soon introduce genetic evaluations for 6 health traits in Holstein cattle, including retained placenta and metritis as

measures of reproductive health (Parker Gaddis *et al.*, 2017). Cows beginning a lactation with retained placenta or metritis have longer days open and lower conception rates than cows that do not (e.g., Fourichon *et al.*, 2000). However, it is not known to what degree susceptibility to reproductive tract diseases and cow fertility are influenced by common sets of genes. The objectives of this research were to identify genes, genomic regions, and gene networks associated with three measures of fertility (daughter pregnancy rate, DPR; heifer conception rate, HCR; and cow conception rate, CCR) and two measures of reproductive health (metritis, METR; and retained placenta, RETP) in US Holstein cows.

Materials and methods

Phenotypic and genotypic data

Genomic evaluations for DPR, HCR, and CCR from the December 2016 proofs calculated by the Council on Dairy Cattle Breeding (CDCB; Bowie, MD, USA) were combined with evaluations of METR and RETP calculated from on-farm health event data provided by Dairy Records Management Systems (Raleigh, NC, USA) as described in Parker Gaddis *et al.* (2014, 2017). Genotypes included 60,671 SNP used in the routine U.S. evaluations. Holstein bull and cow predictor populations were formed by selecting animals with reliabilities of predicted transmitting ability (PTA) for lifetime net merit greater than the reliability of their parent average. All 35,724 bulls in the predictor set were retained, and a random sample of 35,000 cows was drawn from the 112,895 cows in the predictor population. Only animals with PTA for all traits were included in the analysis.

Genome-wide association studies

The five-trait multivariate genome-wide association study (GWAS) used the model:

where **Y** is an $n \times 5$ matrix offor n individuals, $\boldsymbol{\mu}$ is the intercept, is an n-vector of marker genotypes, is a vector of marker effect sizes for the 5 , **U** is an $n \times 5$ matrix of random effects, and **E** is an $n \times 5$ matrix of errors. The random effects matrix, **U**, was where **K** is a known relatedness matrix and $\mathbf{V}_{\boldsymbol{\varepsilon}}$ is a symmetric matrix of genetic variance components. The error matrix, **E**, was , where is an identity matrix and is a symmetric matrix of residual variance components (Zhou, 2014).

SNP and enrichment analyses

Each autosomal marker was assigned to the closest gene within 25,000 bp using BEDTools version 2.21.0 (Quinlan and Hall, 2010). Gene information was taken from the Bovine UMD3.1

genome assembly (Zimin et al., 2009). After merging with the annotated gene data 36,435 markers were available for subsequent analysis. SNP from the GWAS whose *P*-value from a Wald test exceeded a threshold of 5×10^{-8} from the five-trait multivariate analysis were selected for further analysis and gene function was determined by a review of the literature. Gene ontology (GO; Ashburner *et al.*, 2000) and medical subject heading (MeSH; Morota *et al.*, 2015) enrichment analyses were used to compare all SNP with *P*-values less than 0.05 against a background of all annotated genes in the bovine genome. GO and MeSH term analyses were carried out in R v. 3.4.0 using the "GOSTATS" v. 1.5.3 and "meshr" v. 1.12.0 packages as distributed in Bioconductor v. 3.5.

Gene network construction

An association weight matrix (AWM) was constructed following the procedures previously implemented by Fortes *et al*. The construction of the AWM started with the selection of relevant SNP from those identified as significant in the association analyses. Each column in the AWM corresponded to a trait, and each row corresponded to a SNP. Each cell in the matrix corresponded to the *z*-score normalized effect size for the SNP. When more than one SNP was mapped to the same gene, the most significant SNP was retained and the others dropped. Row-wise partial correlations were computed on the AWM using the PCIT algorithm in R which produced an *m* symmetric adjacency matrix. Each cell in the adjacency matrix corresponded to a partial correlation between gene *i* and gene *j*. When partial correlations were not significant the value in the cell was set to 0. The significant correlations can be interpreted as significant gene-gene interactions. These interactions were used to construct bull (Figure 1) and cow gene networks. In order to avoid spurious connections, the bull and cow networks were reduced to sub-networks including only connections with a partial correlation ≥ 0.98. Correlation networks were visualized using Cytoscape version 3.2.1 (Shannon *et al.*, 2003).

Results and discussion

Genome-wide association studies

The were 43 significant SNP in the bull predictor population, and 11 in the cow population. The five SNP with the largest effects in each population are described in Table 1. There was no clear pattern among gene functions, but developmental, cell-signaling, and protein modification processes were represented in both populations. The top SNP between the bull and cow populations did not overlap.

Table 1. The five SNP with the largest effects on in a multivariate analysis using bull and cow genotypes.

Group	SNP	Chrome[1]	Location	Gene	Function	$-\log_{10}(P)$
Bulls	BTB-00790451	20	57,373,160	*FBXL7*	Ubiquitination	44.67
	ARS-BFGL-NGS-64415	18	48,486,442	*ECH1*	Fatty acid degradation	41.43
	ARS-BFGL-NGS-72630	6	118,871,663	*SORCS2*	Nervous system	21.88

					development	
	BTB-00259343	6	62,642,435	*BEND4*	Longevity	15.02
	Hapmap55409-rs29022997	4	33,236,485	*CROT*	Lipid metabolism	12.77
Cows	ARS-BFGL-NGS-23066	6	92,153,394	*CDKL2*	Sex differentiation	13.26
	BTB-00062715	1	135,269,426	*EPHB1*	Cell signaling	9.08
	BTB-00176697	4	40,934,520	*SEMA3C*	Embryonic development	8.02
	ARS-BFGL-NGS-111133	4	119,341,142	*UBE3C*	Ubiquitination	7.96
	ARS-BFGL-NGS-36082	17	55,916,203	*KDM2B*	Ubiquitination	7.45

[1]Chrome = chromosome number.

GO and MeSH term enrichment analyses

Significantly enriched GO and MeSH terms for the bull and cow populations are presented in Table 2. Gene ontology terms were taken from the Biological Processes category and identify pathways that involve the activities of many gene products. Bulls were enriched for processes including spermatogenesis and DNA processing, while cows were enriched for a broad array of pathways including embryonic development and gene expression. Medical subject heading terms identify enriched processes based on literature reports. As in the case of GO terms, many different processes were identified in bulls, while cows had only two significant terms.

Table 2. Gene ontology (GO) and medical subject heading (MeSH) terms with significant effects on in a multivariate analysis using bull and cow genotypes.

	GO[1]			MeSH[2,3]		
Group	GO ID	Term	P-value	MeSH ID	Term	P-value
Bulls	0006270	DNA replication initiation	0.005	D002970	Cleavage stage, ovum	0.004
	0007288	sperm axoneme assembly	0.014	D003599	Cytoskeleton	0.032
	0051661	maintenance of centrosome location	0.014	D009210	Myofibrils	0.035
	1902979	mitotic DNA replication termination	0.014	D013116	Spinal cord	0.035
	0007283	spermatogenesis	0.036	D042541	Intracellular space	0.036
Cows	2000738	positive regulation of stem cell differentiation	0.016	D002823	Chorion	0.034
	0070126	mitochondrial translational termination	0.024	D009092	Mucous membrane	0.043

2000637	positive regulation of gene silencing by miRNA	0.024	—	—	—
0048701	embryonic cranial skeleton morphogenesis	0.039	—	—	—
0060147	regulation of posttranscriptional gene silencing	0.039	—	—	—

[1]Biological processes (BP) category.
[2]Anatomy (A) category.
[3]Only two MeSH terms were significantly enriched in the cow population.

Gene networks

Sex-specific gene networks included 824 genes in bulls and 856 genes in cows. Their number of connections (the degree of the vertices induced by the PCIT algorithm) ranged between 1 and 1,049 in bulls, and 1 and 1,240 in cows. The two networks shared 139 genes in common. The number of connections between nodes in biological networks usually follows a Power-law distribution. We used a Kolmogorov-Smirnov test to validate this assumption, and the null hypothesis of the networks being drawn from a Power-law distribution was not rejected. Several genes identified as the top connected in the networks were associated with either male or female fertility and embryo size and morphology in mice or humans, most notably *ITPR1*, *SETB1*, *LMNB1*, *NEO1*, and *DGKA*. None of the 100 SNP explaining the largest amount of variance in the GWAS were among the most connected genes in the networks.

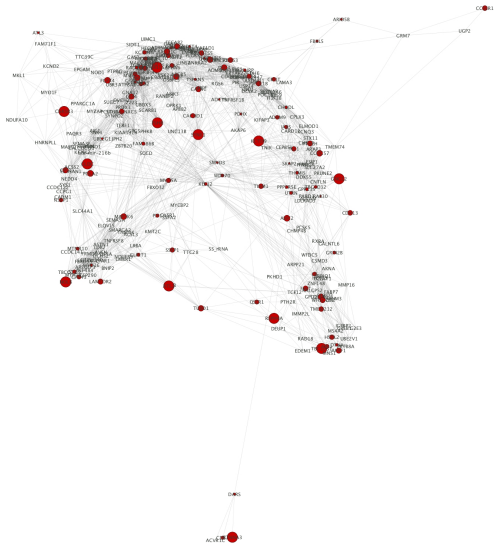

Figure 1. Gene network based on the bull predictor population constructed using edges from thewith partial correlation ≥ 0.98. Node sizes are proportional to their degree.

Conclusions

As expected, these analyses identified individual SNP associated with fertility, and enriched pathways also included some fertility terms. Bull- and cow-specific gene networks similarly included genes with known effects on fertility. However, no significant loci had any obvious associations with reproductive tract health as measured by METR and RETP. This may be due to the . A case-control study using paired animals could provide greater power for identifying SNP and coexpression networks associated with both reproductive health and fertility.

Acknowledgments

The authors thank Dairy Records Management Systems (Raleigh, NC) for providing health data. The Council on Dairy Cattle Breeding (Bowie, MD) provided fertility evaluations, and the Cooperative Dairy DNA Repository (Madison, WI) provided genotypes.

List of References

Ashburner M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, *et al.*, 2000. Gene Ontology: tool for the unification of biology. Nature Genet. 25:25–29.

Fourichon, C., H. Seegers, & X. Malher, 2000. Effect of disease on reproduction in the dairy cow: a meta-analysis. Theriogenol. 53:1729–1759.

Garrick, D.J., J.F. Taylor, and R.L. Farnando, 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet. Sel. Evol. 41:55.

Morota, G., F. Peñagaricano, J.L. Petersen, D.C. Ciobanu, K. Tsuyuzaki, *et al.*, 2015. An application of MeSH enrichment analysis in livestock. Anim. Genet. 46(4):381–387.

Parker Gaddis, K.L., D.J. Null, and J.B. Cole, 2016. Explorations in genome-wide association studies and network analyses with dairy cattle fertility traits. J. Dairy Sci. 99(8):6420–6435.

Parker Gaddis, K.L., J.B. Cole, J.S. Clay, & C. Maltecca, 2014. Genomic selection for producer-recorded health event data in US dairy cattle. J. Dairy Sci. 97(5): 3190–3199.

Parker Gaddis, K.L., M.E. Tooker, J.R. Wright, J.H. Megonigal, J.S. Clay, *et al.*, 2017. Development of genomic evaluations for direct measures of health in U.S. Holsteins and their correlations with fitness traits. J. Dairy. Sci. 100(Suppl. 2):378(abstr. 377).

Quinlan, A.R., and I.M. Hall, 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26(6):841–842.

Servin, B., & M. Stephens, 2007. Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet. 3:e114.

Shannon, P., A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, & T. Ideker, 2003. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res. 13:2498–2504.

Zhou, X., P. Carbonetto, & M. Stephens, 2013. Polygenic modeling with Bayesian sparse linear mixed models. PLoS Genet. 9:e1003264.

Zhou, X., and M. Stephens, 2014. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat. Methods 11(4):407–409.

Zimin, A.V., A. L. Delcher, L. Florea, D.R. Kelley, M.C. Schatz, *et al.* 2009. A whole-genome assembly of the domestic cow, Bos taurus. Genome Biol. 10:R42.