Genome Wide Association Analysis of Lung Lesions in Cattle using Sample Pooling

J. W. Keele, L. A. Kuehn, T. G. McDaneld, S. A. Jones, T. P. L. Smith, S. D. Shackelford, D. A. King and T. L. Wheeler
U. S. Meat Animal Research Center, Clay Center, NE, USA

ABSTRACT: Lung samples were collected from 11,520 young cattle from a large beef processing plant in central Nebraska. Lung samples with lesions (cases) and healthy lungs (controls) were collected when both phenotypes were in close proximity on the viscera table. Lung samples for 96 animals with the same phenotype (case or control) were placed in a pool; 60 pools each for case and control, and the Bovine HD array (770K SNP) was run on all pools. Seven SNP on BTA 3, 7, 9, 11, 14 and 15 were significant at the genome wide experiment wise error rate of 5 % ($P \leq 1.49 \times 10^{-7}$). Eighty-four SNP on 28 chromosomes achieved a false discovery rate of 5 % ($P \leq 5.47 \times 10^{-6}$). Significant SNP were near (+/- 100 Kb) genes involved in tissue repair and regeneration, tumor suppression, control of organ size, and immunity.
Keywords: bovine respiratory disease; lung lesions; genome wide association study; beef cattle

Introduction

Bovine respiratory disease complex (BRDC) is the most expensive disease in cattle costing more than $1 billion annually in the U.S. alone (Griffin (1997)). Several vaccines are available to prevent this disease but many cattle acquire BRDC in spite of rigorous vaccination schedules. Sub-therapeutic use of antibiotics is effective in controlling and preventing BRDC and therapeutic use is effective in treating disease but there are concerns that extensive use of broad spectrum antibiotics in livestock production could lead to antibiotic resistance in human pathogens like tuberculosis. Genetic selection for host resistance to BRDC might be an effective way of generating resistance in the host population which would be expected to reduce incidence in the feedlot. Unfortunately, commercial feedlot cattle are exposed to a wider variety of feedlot environments than seedstock cattle which would compromise selection response. Marker assisted selection based on genetic markers could be effective provided that the collection of markers used accounts for sufficient variation in BRDC resistance/susceptibility to make selection viable. Previous studies indicate that the heritability of BRDC is low (5 to 15 %; Schneider et al. (2010); Snowder et al. (2006)).

The low heritability of BRDC indicates that sample size requirements are very large for detecting SNP associations with BRDC. BRDC results in pneumonia in many cattle and pneumonia results in lung lesions as evidence of infection long after the cattle have recovered. Hence, collection of lung lesion data in slaughter plants offers a way of acquiring large samples of cattle with and without lung lesions to investigate the genetics underlying BRDC. Unfortunately, previous studies indicate that the correlation between lung lesions at slaughter and previous BRDC in the feedlot is low. This low correlation may be because cattle showing signs of BRDC are usually treated with antibiotic and antibiotic treatment in many cases might prevent pneumonia and the subsequent formation of lung lesions. Also many cattle have severe lung lesions that have never shown signs of BRDC; these may be silent sufferers. Pneumonia is part of the progression of BRDC; hence, we expect overlap between the genetic footprints of lung lesions in the packing plant and BRDC in the feedlot; however, treatment and silent cases have obscured the relationship between BRDC and lung lesions when looked at as a linear correlation.

Our objective was to exploit the large number of lung samples available in commercial beef processing plants to increase our sample size and enable the detection of SNP associations. We use sample pooling of extreme animals to reduce the cost of genotyping. We aim to identify regions of the genome that underlie variation in BRDC. Here we investigate lung lesions and compare our findings to previous research on BRDC. Indeed we find common genetic footprints between lung lesions and BRDC.

Materials and Methods

Lung Sampling. We sampled 11,520 lungs from a large central Nebraska beef processing plant over 38 days. Sampling days were not contiguous but were intermittent over 5 months. Lungs were sampled from shifts in which cattle belonged to 1 of 2 production/marketing programs, *Natural* or age and source verified. Both of these programs had relatively high incidences of lung lesions compared to other categories. *Natural* cattle received no added hormones, were traceable to place of birth, had not received antibiotics and had not consumed animal byproducts. *Natural* cattle could receive non-steroidal anti-inflammatory drugs (NSAIDS) and they could be vaccinated. If cattle targeted for *Natural* programs became sufficiently ill in the feedlot as to require therapeutic antibiotics then they were removed from the *Natural* program. Ranch of birth and birth date were known by the processing plant for age and source verified cattle.

Lungs were sampled from the moving viscera table when both healthy lungs and lungs with lesions were visible in close proximity to one another. There were large gaps between lots (cattle purchased from the same source); hence, requiring variation in lung lesions in close proximity on the table before extreme samples were collected increased the probability that lungs with lesions (cases) and healthy lungs (controls) were matched by source. Lung lesions were identified by SAJ, a veterinarian with

extensive experience in postmortem evaluation of lung pathology in cattle having expired from respiratory disease and in slaughter plants.

Sample Preparation and Pooling. Five to 10 gram lung samples were placed on ice immediately after collection and frozen (-20 ^{o}C) within 10 hours after collection and stored frozen until aliquots could be taken for pooling. Sixty lung tissue pools were constructed for each phenotype (case or control) for a total of 120 pools. For each sample pool, cores (1.2 mm diameter × 5 mm length) were taken from partially thawed lung samples from 96 animals of the same phenotype (case or control). DNA was isolated from the pooled sample by a standard salt extraction procedure. In a pilot project, we compared this procedure of pooling lung samples prior to extraction to pooling of DNA following extraction of individual lung samples. We found that pool construction variation was not affected by method used to construct pools (data not shown).

Sample pools were sent to Neogen (Lincoln, NE) for analysis with the Bovine HD array (770K SNP).

Statistical analyses. Pooling allele frequency (PAF; Peiris et al. (2011)) was computed as X / (X + Y), where X is normalized red intensity and Y is normalized green intensity. Hence, PAF is a pooling estimate of the frequency of the A allele. Population stratification and outliers were visualized using a neighbor joining tree constructed from Euclidean distances among pools computed across all SNP. Euclidean distances among pools were computed using dist() of R. Pools that did not fall within a cluster and were at least twice as distant from any cluster compared to inter-pool distances within cluster were considered outliers and were culled from further analysis.

Effects of phenotype (case or control) on individual SNP were estimated and P Values computed based on methods similar to McDaneld et al. (2014). Briefly, average covariance matrix ($\boldsymbol{A}$) among pools was computed across all SNP using cov() of R to account for population stratification and technical errors common to all SNP on an individual pooled sample on an SNP array. Using mixed model methodology, effects were estimated assuming that PAF for a given SNP ($\boldsymbol{y}$) was distributed as multivariate normal with mean dependent on whether the pool was case or control and variance proportional to $\boldsymbol{A}$; hence $y \sim MVN(\boldsymbol{X\beta}, \boldsymbol{A}\sigma^2)$, where $MVN()$ is the multivariate normal distribution and σ^2 is an SNP specific multiplication factor estimated from the individual SNP data. $\boldsymbol{X}$ is a 114 × 2 incidence matrix with the first column having a value of 1 for case pools and 0 for controls; whereas the second column has a value of 1 for control pools and 0 for cases. The mean PAF for cases and controls, respectively, are in $\boldsymbol{\beta}$. The test used to compute *P* values was F with 1 numerator degrees of freedom and 112 denominator degrees of freedom.

We estimated the effective number of tests (M_{eff}) genome wide using Simple M (Gao et al. (2008)) to derive the family (or experiment) wise error rate at the 5 % level to adjust for multiple testing of ~770k SNP and linkage disequilibrium (LD) among SNP on the same chromosome. We estimated M_{eff} using individual genotype Bovine HD data from 1,530 animals which were a combined data set of hapmap animals (n=718; Porto-Neto et al. (2013)) and USMARC animals (n=812). The effective number of tests based on individual genotype data from the Bovine HD array on 1,530 animals was 343,497 which meant that a 5 % genome wide error rate corresponded to a nominal P value of 1.49×10^{-7}.

In addition, we estimated the false discovery rate which utilizes discrepancies from the uniform distribution of *P* values as evidence of multiple true positive SNP associations with the phenotype. The nominal P value corresponding to a false discovery rate (FDR) of 5 % was estimated as nominal $P \leq 5.47 \times 10^{-6}$ (Benjamini and Hochberg (1995)).

Results and Discussion

Whole genome PAF for 6 outlier samples were excessively divergent from all other samples so they were culled from further analysis leaving 114 pools for GWAS rather than the targeted 120. Seven SNP achieved genome wide significance at the 5 % level ($P \leq 1.49 \times 10^{-7}$; Figure 1. In order from most significant to least, genome wide significant SNP were at positions (chromosome: position expressed as kilobases (Kb) from the most centromic SNP) 15:66,932; 14:183; 9:72,751; 3:60,357; 7:40,473; 11:62,650; and 15:22,840. Eighty-four SNP including the 7 listed above achieved a false discovery rate (FDR) of 5 % (nominal $P \leq 5.47 \times 10^{-6}$) which included SNP on all autosomes and the X chromosome with the exception of BTA16 and 19. Numbers of SNP achieving FDR ≤ .05 by chromosome were 8 for BTA15; 7 for BTA3; 6 for BTAX; 5 each for BTA4, 7, and 21; 4 for BTA12; 3 each for BTA 2, 8, 9, 10, 14, and 27; 2 each for BTA1, 5, 6, 13, 20, 23, 24, 25, 26, 28, and 29; and 1 each for BTA11, 17, 18, and 22.

The most significant SNP at 15:66,932 is 1.4 Kb from the gene four-jointed (FJX1). FJX1 is a component of the planar cell polarity and Hippo pathways (Happé et al. (2011)). These pathways are involved in epithelial repair, tissue regeneration, and organ size control. Furthermore, the planar cell polarity and Hippo pathways are involved in lung development and maintenance of healthy adult lung (Chung et al. (2013); Yates et al. (2011)).

The second place SNP at 14:183 is not near (+/- 100 Kb) an annotated gene.

The third place SNP at 9:72,751 is 14.6 Kb away from Eyes Absent 4 (EYA4). EYA4 has been associated with familial lung cancer risk and has a role in modulating apoptosis and DNA repair (Wilson et al. (2013)). EYA4 is a tumor suppressor gene and has a clear role in maintenance of healthy lung tissue.

The fourth place SNP at 3:60,357 is within the TTLL7 gene. The TTLL7 gene is required for the normal development of components of the nervous system (Ikegami et al. (2006)).

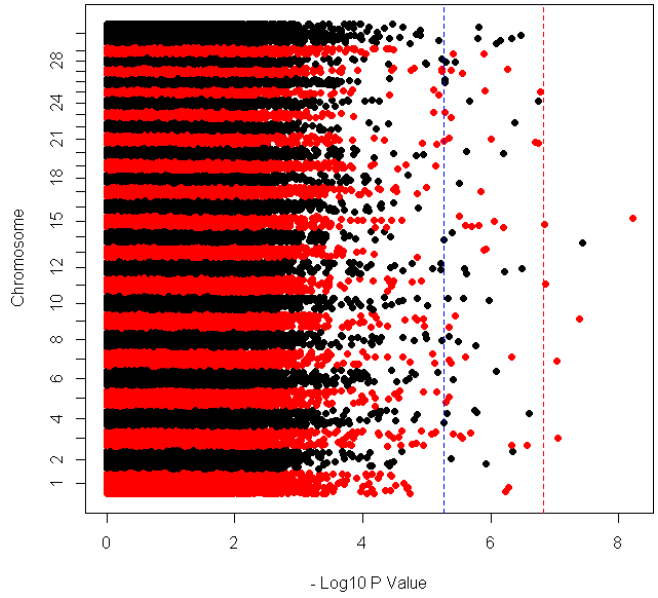

Figure 1: Manhattan plot for genome wide association on Lung Lesions. Blue dashed line is the false discovery rate of 5 % threshold and the red dashed line corresponds to the genome wide experiment wise error rate of 5 %.

The fifth place SNP at 7:40,473 is 2.6 Kb from *TMED9*. The *TMED9* gene is a protein transporter which is targeted by the microRNA, *miR-296* (Robson et al. (2012)). *miR-296* is imprinted so it imposes parental specific modulation of expression of *TMED9*.

The sixth place SNP at 11:62,650 is not near an annotated gene.

The seventh place SNP at 15:22,840 is 2.0 Kb from beta carotene oxygenase 2 (*BCO2*). *BCO2* has been associated with daily milk and protein yield and somatic cell counts in Czech Fleckvieh dairy cattle (Bartonova et al. (2012)). Mutations in *BCO2* have been shown to affect milk color in cattle (Berry et al. (2009)). Furthermore, *BCO2* can affect Vitamin A availability and a deficiency affects immune function in cattle (Spears (2000)) and humans (Thornton et al. (2014)).

There is not sufficient space here to go into detail on all SNP achieving FDR $\leq$ .05; however, a few are note worthy. An SNP at 2:128,834 ($P = 4.46 \times 10^{-7}$) is within a 12.8 Mb region identified by Niebergs et al. (2011) associated BRDC and persistent infection with BVDV. The 12.8 Mb is the distance between flanking microsatellites surrounding their peak association signal. Our SNP is 15.2 Kb from *SRRM1*. *SRRM1* is a spliceosomal protein and is a member of the serine-arginine protein kinase family. SRRM1 was differentially expressed in the placental proteome between mothers with gestational diabetes and healthy controls (Lapolla et al. (2013)).

An SNP at 3:119,882 ($P = 4.63 \times 10^{-7}$) is 5.4 Kb from *CD1A* which is an antigen-presenting molecule that presents lipids instead of peptides to T cells. A polymorphism in human *CD1A* has been associated with susceptibility to tuberculosis (Seshadri et al. (2014)).

Conclusions

Seven SNP achieved genome wide significance at the 5 % level and 84 SNP achieved a FDR level threshold of 5 %. Generally, genes near significant SNP were involved in tissue repair and regeneration, tumor suppression, control of organ size, and immunity. These findings indicate that a broad diversity of gene functions employed are involved in resisting and recovering from bovine respiratory disease. In the future, we plan to look at BRDC using a similar approach but feedlot samples are more difficult to come by so sample size will not be as large.

Mention of a trade name, proprietary product, or specific equipment does not constitute a guarantee or warranty by the USDA and does not imply approval to the exclusion of other products that may be suitable. USDA is an equal opportunity provider and employer.

Literature Cited

Bartonova, P., Vrtkova, I., Kaplanova, K. et al. (2012). Genet. Mol. Res. 11(2):1058-1063.

Benjamini, Y., Hochberg, Y. (1995). J. R. Statis. Soc. 57:289-300.

Berry, S. D., Davis, S. R., Beattie, E. M. et al. (2009). Genetics 182:923-926.

Chung, C., Kim, T., Kim, M. et al. (2013). PNAS 110:7732-7737.

Gao., X., Starmer, J. D., Martin, E. R. (2008). Genetic Epidemiology 32:361-369.

Griffin, D. (1997). Food Anim. Pract. 13:367–377.

Happé, H., van der Wal, A. M., Leonhard, W, N. et al. (2011). J. Pathol. 224(1):133-142.

Ikegami, K., Mukai, M., Tsuchida, J. et al. (2006). J. Biol. Chem. 281(41)30707-30716.

Lapolla, A., Porcu, S., Roverso, M. et al. (2013). Eur. J. Mass. Spectrom. 19:211-213.

McDaneld, T.G., Kuehn, L. A., Thomas, M. G. et al. (2014). J. Anim. Sci. (submitted).

Peiris, B. L., Ralph, J., Lamont, S. J. et al. (2011). Anim. Genet. 42:113-116.

Porto-Neto, L. R., Sonstegard, T. S., Liu, G. E. et al. (2013). BMC Genomics 14:876.

Schneider, M. J., Tait, R. G. Jr., Ruble, M. V. et al. (2010). J. Anim. Sci. 88:1220-1228.

Robson, J. E., Eaton, S. A., Underhill, P. et al. (2012). RNA 18(1);135-144.

Seshadri, C., Thuonq N. T., Yen, N. T. et al. (2014). Genes Immun. (2/6/2014, Epub ahead of print).

Snowder, G. D., Van Vleck, L. D., Cundiff, L. V. et al. (2006). J. Anim. Sci. 84:1999-2008.

Spears, J. W. (2000). Proc. Nutr. Soc. 59:587-594.

Thornton, K.A., Mora-Plazas, M., Marin, C. et al. (2014). J. Nutr. (2/5/2014, Epub ahead of print)

Wilson, I. M., Vucic, E. A., Enfield, K. S. et al. (2013). Oncogene (2013/10/07/online).

Yates, L. L., and Dean, C. H. (2011). Organogenesis 7(3): 209–216.