Whole Genome Association Analysis of Susceptibility to Paratuberculosis in Holstein Cattle

B. W. Kirkpatrick[*] , X. Shi[*], G. E. Shook[†], M. T. Collins[‡]

Introduction

Paratuberculosis, commonly called Johne's disease, is a chronic infection of the small intestine caused by *Mycobacterium avium,* ssp. *paratuberculosis (MAP)*. The disease occurs in cattle and other ruminants with signs including diarrhea, severe weight loss, and decreased milk production. Cattle normally become infected with *MAP* as calves, but clinical signs of infection usually do not appear until the cow's second or third lactation. There is no cure for the disease and infected animals ultimately become emaciated and are removed from the herd prematurely. Paratuberculosis is a good candidate for genetic selection because a) an effective vaccine is not available, b) the disease is not curable, c) it causes significant economic losses and d) it is potentially zoonotic. Selective breeding to reduce disease susceptibility would be a low cost, sustainable practice. The current study was undertaken to identify genetic markers and genomic regions associated with susceptibility to infection by *MAP*.

Materials and methods

Animal resources. Two resource populations of approximately 5,000 cows each were used to identify genomic regions associated with susceptibility to infection by *MAP*. Population 1 consisted primarily of daughters of twelve Holstein sires. Collection of these samples has been described previously (Gonda *et al*. 2006). Samples were obtained mainly during 2001 to 2003 from 300 herds from across the US. Cows were specifically chosen to be in second or later lactation to increase the likelihood of identifying cows manifesting evidence of infection. Population 2 consisted of cows from six commercial Holstein herds in Wisconsin that were cooperators in a Johne's disease control project. For the current project, blood samples for disease testing and DNA extraction were obtained from all cows in these herds over a period of 15 months in 2006-7; these samples were collected near the end of the six-year disease control project.

Diagnosis of *MAP* infection was based on fecal culture of *MAP* (Collins *et al*. 1990) or evidence of antibody titer to *MAP* as based on an ELISA test (Shin *et al*. 2008) of blood samples. In both populations samples with the highest sample:positive ratios were preferentially genotyped. Case definition for Population 1 was a positive test either with the

[*] Department of Animal Sciences, University of Wisconsin-Madison, Madison, WI 53706
[†] Department of Dairy Science, University of Wisconsin-Madison, Madison, WI 53706
[‡] Department of Pathobiological Sci., School of Veterinary Medicine, Univ. of Wisconsin-Madison, Madison, WI

fecal culture or ELISA (8.8% of samples). Case definition for Population 2 was based only on ELISA test of blood samples (9.7% of samples ELISA positive).

Genotyping and data triage. DNA samples from both populations were genotyped with the Illumina Bovine SNP50 Bead Chip. A total of 248 infected cows from Population 1 and 307 infected cows from Population 2 were genotyped with the SNP50 Bead Chip. Animals with fewer than 95% successfully scored genotypes and markers that were successfully scored for fewer than 90% of the samples in either of the two resource populations were removed prior to statistical analyses. In addition, SNPs with unknown genomic location or with minor allele frequencies below 5% (Cole *et al*. 2009) were not included in analyses. After exclusion for these various reasons, a total of 35,772 SNPs remained.

Reference samples for a case-reference analysis were an extensive sample of Holstein bulls used as artificial insemination (AI) sires in the US. Bull genotype data based on the SNP50 Bead Chip was obtained from the USDA Bovine Functional Genomics Laboratory and Cooperative Dairy DNA Repository (CDDR) cooperators. Bulls serving as the reference group for Population 1 were chosen based on their inclusion as sires of Holstein dams calving between 1994 and 1998 based on national calving records. Bulls serving as the reference group for Population 2 were chosen based on their usage as sires and maternal grandsires within the six cooperating herds. Given the known paternal half-sib family structure in Population 1, female samples were checked for paternity relative to potential sires using a subset of 200 SNPs with high minor allele frequency. Of 233 females, 205 were verified as daughters of project sires.

Statistical analysis. Analysis of data from Population 1 accounted for the paternal half-sib family structure in the population. Inheritance of paternal and maternal haplotypes in Population 1 was determined using a Fortran program (de Roos *et al*. 2008) that compared sire and offspring genotypes for each SNP marker. Paternally inherited haplotypes at each marker bracket were evaluated for deviation from a frequency of 0.5 expected under the null hypothesis of no linkage with a disease susceptibility locus using a z test.

Frequency of maternally inherited SNP alleles from daughters in paternal half-sib families were used for a case-reference analysis, in combination with allele frequency estimates from 28 infected cows which were not daughters of the 12 project sires. Maternally inherited allele frequencies were estimated using a single locus, maximum likelihood estimator. Bull allele frequency for Population 1 was a weighted estimate with weighting in proportion to the individual bull's representation as a sire of dam's in the national calving records for 1994-1998. SNP association with susceptibility to *MAP* infection was tested by comparing allele frequency difference between case (maternally inherited alleles of infected cows) and reference (Holstein allele frequency based on bull data) groups. Results from the two separate statistical tests (linkage, linkage disequilibrium, ie case-reference) were subsequently combined to yield a combined linkage-linkage disequilibrium result for Population 1. SNP association with susceptibility to *MAP* infection for Population 2 was likewise tested by comparison of allele frequencies between infected cows and an estimate of population allele frequency. Allele frequencies were estimated separately by herd.

Development of a multi-marker model for prediction. The most significant autosomal markers from separate and combined case-reference and linkage-linkage disequilibrium analyses (nominal P<0.001; n=197) were used in logistic regression analysis to identify a subset of markers that could be used in genomic selection. The data set was comprised of the 521 genotyped infected cows from Populations 1 and 2 and 1,025 Holstein AI sires that served as reference animals in the analysis. These 1,546 samples were randomly assigned to ten groups. For model development and cross-validation, nine of the ten groups were combined to comprise a training data set, and the model developed from the training data set was applied in prediction using the remaining group (ie. testing data set). Model efficacy was evaluated by determining percent concordance. Models were constructed using a forward-stepwise approach with a minimum probability for SNP entry of P<0.005 and a minimum probability for continued inclusion in the model of P<0.001.

Results and discussion. Strong linkage signals ($P<5x10^{-5}$) were observed on chromosomes 6 and 20 (Figure 1), in the latter case strengthening and refining a previous observation based on a subset of the population and within-family linkage analysis of microsatellite marker data (Gonda *et al*. 2007). Tests of allele frequency difference (association test) in Population 1 surpassed a moderate significance level ($P<5x10^{-5}$) for SNPs on BTA17, 20 and 23 (Figure 1, top panel). The location of the most significant association test results on BTA20 corresponded very well with the peak of the linkage analysis results for this chromosome. As a consequence, the combined linkage-linkage disequilibrium plot showed the most striking and significant result for BTA20 (Figure 1, bottom panel).

In general, results from analysis of Population 2 were less significant than Population 1 (Figure 2). Association test results in Population 2 surpassed a moderate significance level ($P<5x10^{-5}$) for SNPs on BTA14 and 26. As in Population 1, no individual marker associations surpassed a more stringent level of 5×10^{-7} adopted for strong association.

The number of markers included in the models developed in the cross-validation analyses ranged from 12 to 20 and the models produced concordance values ranging from 66.3 to 77.5%. Cross-validation analysis identified 18 SNPs that appeared in four or more of the ten models developed with the various subsets of the data. Based on the concordance of observed and predicted values in the cross-validation testing sets, a concordance of approximately 71% could be expected. Potentially, these estimates of concordance are conservative in the sense that they may underestimate the true ability of the model to discern between genetically susceptible and less susceptible groups. The reference bulls represent the overall contemporary population and were of unknown disease status rather than a test-negative group.

Conclusion

The SNP associations reported here provide preliminary information that can be useful in initial efforts aimed at genomic selection against genetic susceptibility to *MAP* infection. Validation of these results by additional testing in independently derived Holstein populations is highly desirable.

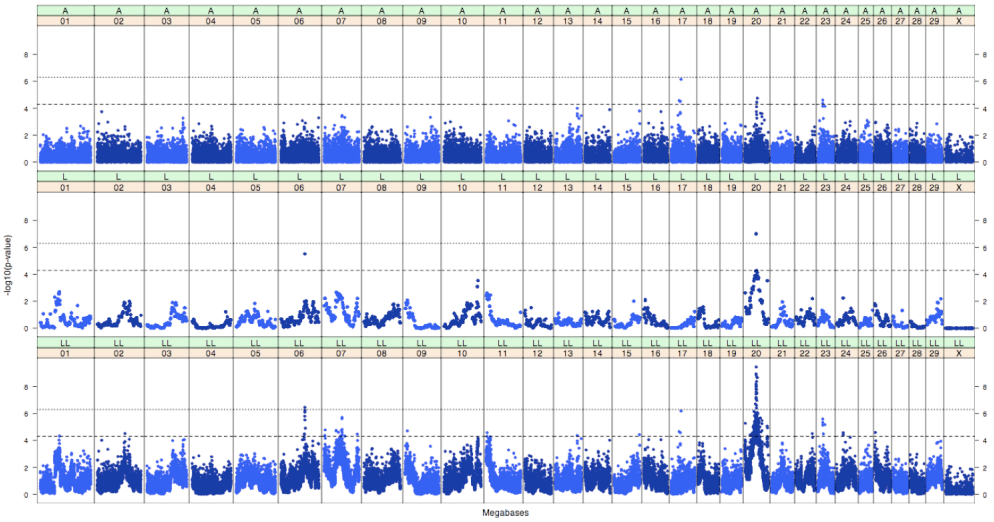

Figure 1: Results of whole genome scan of Population 1 for genetic marker association with susceptibility to infection of cattle by *MAP*. Points represent the $-\log_{10}$ of the P-value (y-axis) from association (top; "A"), linkage (center; "L") and combined linkage-linkage disequilibrium (bottom; "LL") analyses, relative to SNP genomic location (x-axis). Dashed and dotted lines represent p-values of $5x10^{-5}$ and $5x10^{-7}$, respectively.

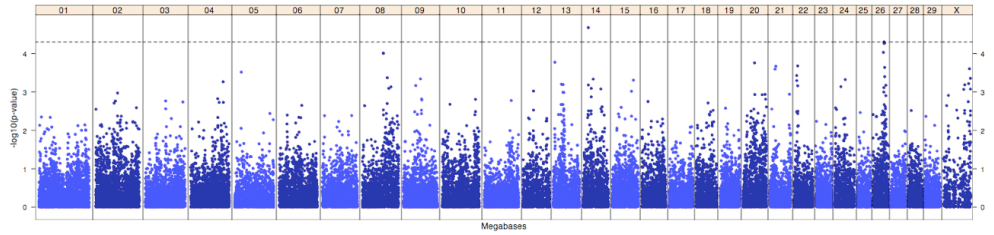

Figure 2. Results of whole genome scan of Population 2. The dashed line represents a P-value of $5x10^{-5}$ corresponding to moderate significance.

References

Cole J.B., Van Raden P.M., O'Connell J.R. *et al*. (2009). *J. Dairy Sci*. 92:2931-2946.

Collins M.T., Kenefick K.B., Sockett D.C., *et al*. (1990). *J. Clin. Micro*. 28:2514-2519.

de Roos A.P., Hayes B.J., Spelman R.J. *et al*. (2008). *Genetics* 179:1503-12.

Gonda M.G., Chang Y.M., Shook G.E. *et al*. (2006). *J. Dairy Sci*. 89:1804-1812.

Gonda M.G., Kirkpatrick B.W., Shook G.E. *et al*. (2007). *Anim. Genet*. 38:389-396.

Shin S.J., Cho D. & Collins M.T. (2008). *Clin. Vac. Immun*. 15:1277-1281.