LIC report for the JDRC

Executive Summary

- A strategy for applying animals to cross-validation folds was identified and implemented.
- The relationship between animals within a fold was increased to twice that of animals outside the fold.
- The predictive power of the model as measured by the area under the ROC curve has decreased to 0.887 from 0.906.
- The model is proving to be relatively robust but predictions of susceptibility to Johnes' disease continue to be better than expected based on the estimated heritability of 0.22.
- Given the results the proposed validation study becomes more important in providing confidence in the technique. LIC is possibly interested in investing in this study to expand it.

Summary of progress

The JDRC Science review panel suggested that the predictive power of the test is too high given the estimated heritability of the trait (~0.22). The suspected cause was having closely related animals in both the training & test groups. To address this issue, k-means clustering of a decomposed A matrix was used to optimise the allocation of animals to the 10 folds such that animals within a fold are more related to each other than animals in the other 9 folds. A linear programming approach was then taken to re-balance the folds to the same size.

The results of this allocation (Tables 1 & 2) show that, relative to randomly assigning animals to folds, the balanced k-means approach was able significantly increase the relationships between animals in the same fold (0.073 vs 0.041), and decrease the relationships between animals in a fold and its complement (0.037 vs 0.031).

Table 1. Fold size and average relationship of a fold with the fold's complement by strategy for allocating animals to test folds.

	Fold size			Average relationship with other folds		
Fold	Random	k-means	Balanced k-means	Random	k-means	Balanced k-means
1	840	605	840	0.041	0.046	0.040
2	840	305	840	0.041	0.044	0.045
3	840	883	840	0.041	0.045	0.045
4	840	664	840	0.040	0.047	0.042
5	840	164	840	0.040	0.018	0.043
6	840	2087	840	0.040	0.018	0.019
7	840	416	840	0.041	0.027	0.023
8	840	728	840	0.041	0.043	0.042

9	840	2186	840	0.040	0.041	0.046
10	844	366	844	0.041	0.022	0.026

Table 2. Degree of relationship between each fold and its complement averaged over 10 folds.

	Relationship	
	between folds	within fold
Random	0.041	0.041
k-means	0.035	0.120
Balanced k-means	0.037	0.073

The tenfold cross-validation study was then re-run to determine the accuracy with which the data could be used to predict Johne's status based on an animal's genomic profile. The software package GenSel (Fernando and Garrick, Iowa State University), was used to fit a Bayes B model using 1 megabase windows across the genome. The model estimated the genomic merit for the different combinations of SNP in each window for each of the 10 training populations and then used those estimates to predict the actual Johne's status of the animals in the corresponding 10 test populations.

Figures 1 & 2 show that there is no clear relationship between predicted susceptibility to Johnes' and either age or Jersey proportion, suggesting that these factors have been successfully accounted for in the analysis and that the predicted merit is not just a proxy for breed or age.

Figure 1. Genomic merit for by age

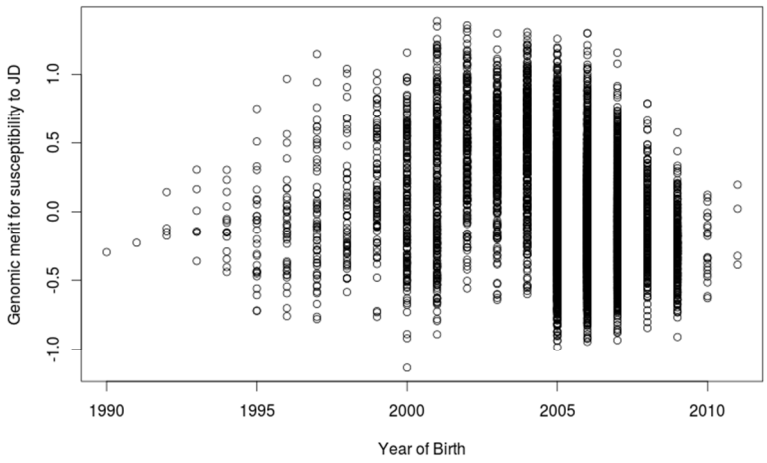
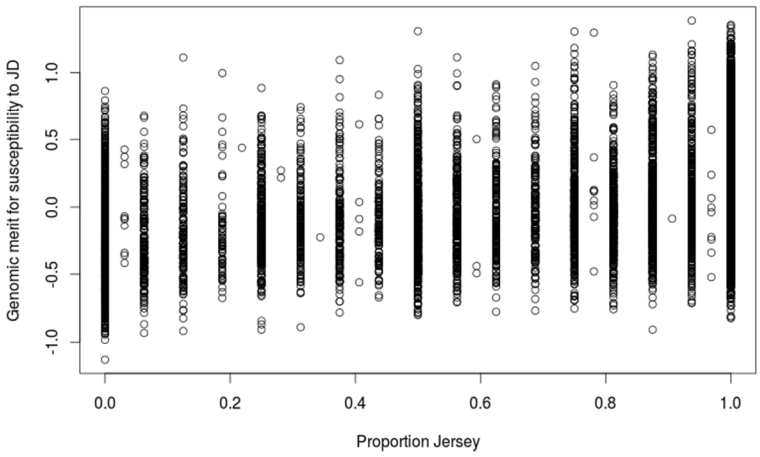


Figure 2. Genomic merit for susceptibility to Johnes' disease by Jersey proportion.



Figures 3 & 4 show that the JD+ group has a significantly higher mean for predicted merit for susceptibility to JD than the Control group, but that the populations overlap.

Figure 3. Distribution of predicted genomic merit of Control and Johne's disease positive (JD+) animals for JD susceptibility

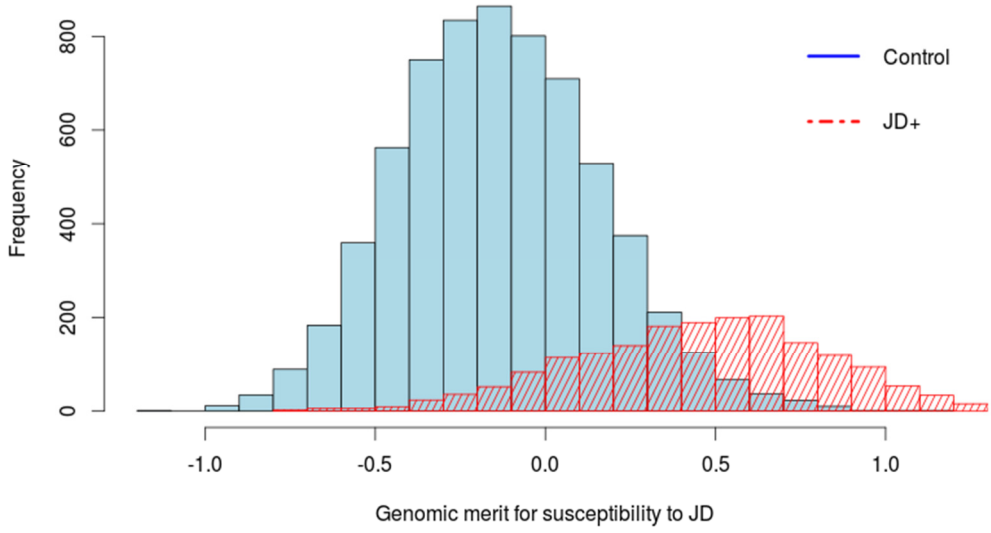


Figure 4. Notched boxplot of the predicted genomic merit of Control and Johne's postive animals for Johne's susceptibility.

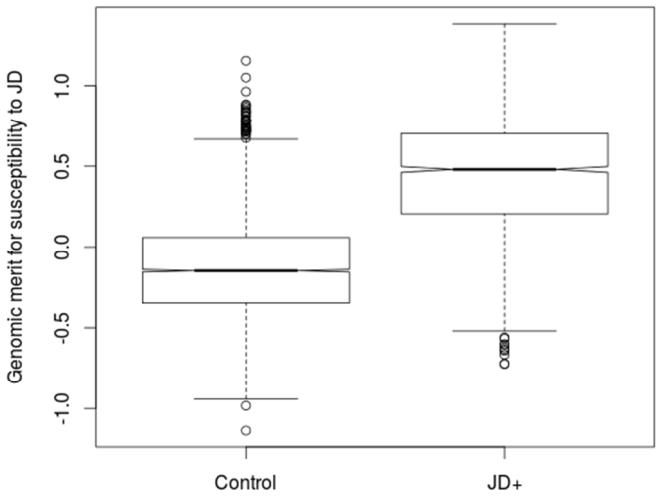
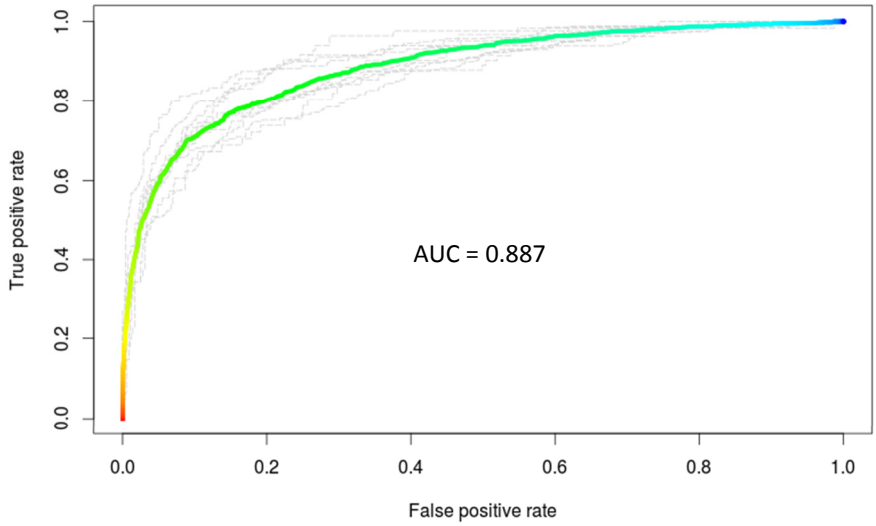


Figure 5. Receiver operating characteristics curve for tenfold cross-validation. The coloured line represents the average of the 10 training-testing datasets shown in grey.



The receiver operating characteristic curves (Figure 5) show that the allocation of more-related animals within each fold has increased the variability of the training-test datasets relative to the previous analysis and has decreased the Area Under the Curve (AUC) from 0.906 to 0.887. If a true positive and a true negative animal were chosen at random, the test would correctly rank them 88.7% of the time.

The effect of setting thresholds on the data from Figure 5 is shown in Table 3. A lower threshold value of -0.22 would correctly identify 40% of animals as non-susceptible, while incorrectly identifying 3.9% of JD+ animals as non-susceptible.

An upper threshold of 0.38 would correctly identify 60% of JD+ animals as susceptible, while identifying approximately 4% of control animals as susceptible. The number of incorrectly assigned

control animals will be less than 4% as some of them would likely have tested positive for JD (prevalence of JD in NZ is about 1%).

The above thresholds would have classified 49% of the population as either susceptible or not-susceptible to JD, while leaving 51% of the population uncategorised.

Table 3. Proportion of true-positive (TP), false-positive (FP), true-negative (TN) and false-negative (FN) animals classified by example thresholds.

Threshold		Proportion classified				
Lower	Upper	TP	FP	TN	FN	All
-0.22	0.38	0.60	0.04	0.40	0.04	0.49
-0.03	0.27	0.70	0.08	0.65	0.10	0.75