Bayesian genomic prediction of junctional epidermolysis bullosa in sheep

Abstract: Junctional epidermolysis bullosa (JEP) is a heritable skin and mucosa disorders condition in association with mendelian mutations in sheep. The purpose of this investigation is to explore the relationship between different priors, linkage disequilibrium and single nucleotide polymorphisms (SNPs) selection methods to accuracy of Bayesian GP of JEP in sheep. 92 Spanish Churra sheep breed genotyped by 40668 SNP markers. Bayes Cπ shown to have slightly higher predicted accuracy [0.724 (0.113)] by unselected data. Prediction performance of the Bayesian GP models was found to be similar after correction for LD. There was a significant difference between predicted accuracies due to the SNPs selection by ranked p values of whole and training only dataset using linear model. The relevance of genetic architecture in conjugate to the prior distributions clearly supported by the unselected data. The most obvious finding emerge from this study is that preselection of SNPs referring to genetic architecture of the phenotype may lower the needs of computational load.

Key words: Bayesian models, genomic prediction, junctional epidermolysis bullosa

1. Introduction

Genomics have emerged as powerful platform in animal breeding and genetics due to decreased costs of molecular markers [1]. Genomic prediction (GP) is important for a wide range of scientific and industrial process in animal breeding including: detection of genes in connection with phenotypes and prediction of genomic breeding values [2, 3]. The main challenge faced by many researchers is the GP of disease statutes of the animals using single nucleotide polymorphisms (SNPs) to obtain clinical diagnostic systems [4].

Scholars have debated the impact of genetic architecture of the phenotype [5], preselection of SNPs [4], and differences between GP methods [4] for explaining the variation in GP accuracy. In the literature of Bayesian GP, the relative importance of prior distributions has been subject to considerable discussion [6]. Investigating genetic architecture of the phenotypes in terms of

prior distributions is a continuing concern to obtain higher GP accuracies. It has conclusively been shown that different priors lead to different genetic architectures in terms of number and action of the genes and level of linkage disequilibrium (LD) [7]. The GP research to date has tended to focus on quantitative phenotypes rather than binary disease statues.

Junctional epidermolysis bullosa (JEP) is a heritable skin and mucosa disorders condition in association with mendelian mutations in sheep [8, 9]. However genetic factors found to be influencing epidermolysis bullosa have been explored in several organisms including cattle [10], sheep [8, 11, 12], horse [13], dog, cats and rats [13]. Surveys in mammals such as that conducted by [10] have shown that hundreds of mutations in association with 18 genes have been molecularly characterized for epidermolysis bullosa. The purpose of this investigation is to explore the relationship between different priors, LD and SNPs selection methods to accuracy of Bayesian GP of JEP in sheep.

2. Materials and methods

92 (17 cases and 75 controls) Spanish Churra sheep breed genotyped by 40668 SNP markers. Phenotypes were assessed by visual inspection of the sheeps and recorded as a binary trait. More details about the dataset could be found at [8]. SNPs were analysed by PLINK [14] for quality control based on minor allele frequencies (<0.95), calling rate of SNPs (>0.90), Hardy-Weinberg proportions (P<1E-07), and optionally Linkage Disequilibrium (LD) (r^2>0.7).

The evolution of the Bayesian GP models was based on cross validation of the genotypic and phenotypic datasets over training and testing partions. Splitting the data as training (%80 of animals) and testing (% 20 of animals) are common for evolution of GP methods [4, 5]. Area under the curve (AUC) approach was used to obtain the accuracies over training and testing partions by using 10 fold cross validations of Bayesian GP methods as was defined in [4].

Bayesian ridge regression (BRR), Bayesian (Least Absolute Shrinkage and Selection Operator) LASSO (BL), Bayes A, Bayes B and Bayes Cπ [7] are currently the most popular Bayesian GP methods for investigating animal breeding datasets. To obtain GP for animals $y = \sum_i^n X_i \hat{g}_i$ could be used where *y* is the phenotype (1 for case of JEP, 0 for healthy control), *n* is the number of SNPs , X_i is a design matrix connecting animals to genotypes at SNP *i*, and $\hat{g}_i$ is the predicted effect of the genotype at SNP i. $\hat{g}_i$ have been used to investigate the genetical properties on the phenotype with referring various prior assumptions regarding genetic architecture of the phenotype. BRR [15] assumes same additive genetic variance for all SNPs by using normal prior distribution. BL [16] assumes Laplace prior distribution for shrinking many of the SNPs towards to zero. Bayes A [15] assumes for the distribution of SNP effects is the Student's *t* distribution. However, there are certain drawbacks associated with the use of Bayes A including non-zero SNP effects over genome. The use of mixture models has a relatively long tradition within GP [17]. One advantage of Bayes B [15] is that it avoids the problem of non-zero SNPs effects by using mixture of two prior distributions:

$\sigma_{gi}^2 = 0$ with probability π,

$\sigma_{gi}^2 \sim \chi^{-2}(v, S)$ with probability (1-π)

where σ_{gi}^2 is the additive genetic variance of SNP *i*, with *v*=4.234, *S*=0.0429 [15], π (assumed to be 0.5) is the probability that the SNP has no effect on the phenotype. One possible improvement over Bayes B could be obtained by predicting π parameter in Bayes Cπ . Bayesian GP analysed by the BGLR package [18] with 52000 markov chain monte carlo iterations by 6000 burn-in period.

The literature on preselection of SNPs for GP has revealed the emergence of several contrasting themes [4, 19] referring genetic architecture of the phenotypes. SNPs were filtered for the stratified training and testing GWAS results set by P values (<0.05) and full data set ranked P

values (<0.05) of linear mixed model (LMM) [20]. Different from Bayes Cπ : BayesR assumes prior distributions with four mixture components of Gaussian distribution to model SNPs effects. LMM used a single SNP regression model hence only SNPs with large effects could be detected.

3. Results and Discussion

After quality control process: 40642 SNPs with 92 sheep obtained. After filtering out highly correlated SNPs from the genotypic dataset, 25254 SNPs obtained (Figure 1). Whole and training only LMM detected 2401 SNPs and 2120 SNPs in association with JEB respectively. Figure 2 and Table 1 compares the prediction accuracies obtained from Bayesian GP models under different experimental designs. This Table 1 is quite revealing in several ways. First, Bayes Cπ shown to have slightly higher predicted accuracy by unselected data. Prediction performance of the Bayesian GP models was found to be similar after correction for LD. There was a significant difference between predicted accuracies due to the SNPs selection by whole and training only LMM analyses. Interestingly, the full LMM based preselected data gave the highest GP accuracy with relatively smaller sampling size and smaller standard errors (Figure 2) compared with the other experimental designs in Table 1. Smallest sampling size (2120 SNPs) with relatively higher prediction accuracies were obtained by the preselection of SNPs using training only LMM. Prediction accuracies was found to be similar over different Bayesian GP models in each SNPs selection methods.

This study set out with the aim of assessing the importance of Bayesian GP of JEP in sheep under different experimental settings. The results of this study indicate that it is possible to predict JEP in sheep using SNPs data. Increased GP accuracy over LMM selected data (Table 1) in this study corroborates with hypothesis of mendelian inheritance pattern for JEP [9]. LMM provided the SNPs with largest set of mendelian effects due to its single SNP regression

algorithm. This finding broadly supports the work of other studies in this area linking mendelian mutations with JEP [8, 13]. Increased GP accuracy over, Bayes Cπ (in the unselected dataset) in this study corroborates these earlier findings of mendelian inheritance of JEP [8].

[21] demonstrated that postulated prior values for Bayesian GP became more important especially with small datasets. Surprisingly, no differences were found in Bayesian prediction accuracies over preselected datasets using different prior distributions (Table 1). These results could be explained by the fact that, filtering the SNPs for high LD or preselection of the SNPs for their effect sizes (by LMM) reduced the genotypic variation in the data and all Bayesian models started to give similar accuracies. These results reflect those of [22] who also found the flexibility and interpretability of GP obtained by BayesR under various simulation experiments including corrections for LD blocks. In accordance with the present results [4] reported their Bayesian GP accuracies were found to be similar due to unmatched genetic architecture of the phenotype and postulated prior distributions.

Including preselected SNPs in the model improved prediction accuracies using GWAS results of whole genotypic dataset (Table 1). Preselection of SNPs beneficial not only for improvement of prediction accuracies but also for reduction of dimension of the genotypic dataset [23]. Instead of 40642 SNPs: LMM SNPs selection models GPs obtained by 2401 SNPs. This finding is consistent with that of [24] who reported advantages of using BayesR in GP for take into account of variants with large effects. It is possible to hypotheses that employing SNPs selection models in conjugate with genetic architecture of the phenotype would be more efficient compared with other data reduction techniques as such principal component analyses. In accordance with the present results [25] point out that Bayesian GP with whole genome sequence data far more computationally expensive with millions of SNPs. However, it has been demonstrated that a multiple chain markov chain monte carlo methods for Bayesian GP results computationally cost effective and accurate predictions [26]. Similarly [27] found that updating

the right hand side of the Bayesian GP equations over multiple SNPs may reduce the need of memory allocations and computing time. However, it could be argued that the inflated accuracies were due to SNPs of GWAS results obtained from whole dataset. These results therefore need to be interpreted with caution. In order to correct for this bias, we performed the GWAS based on only training samples (Table 1), to select SNPs. As shown in Table 1 the accuracies were found to be lower compared with the results of the full dataset.

The present study was designed to determine the effect of prior distributions in Bayesian GP in terms of prediction accuracy under different experimental designs. The relevance of genetic architecture in conjugate to the prior distributions clearly supported by the unselected data. The most obvious finding emerge from this study is that preselection of SNPs referring to genetic architecture of the phenotype may lower the needs of computational load. Consistent with the literature, SNPs selection process should be exercised on training populations in order to avoid falsely inflated accuracies.

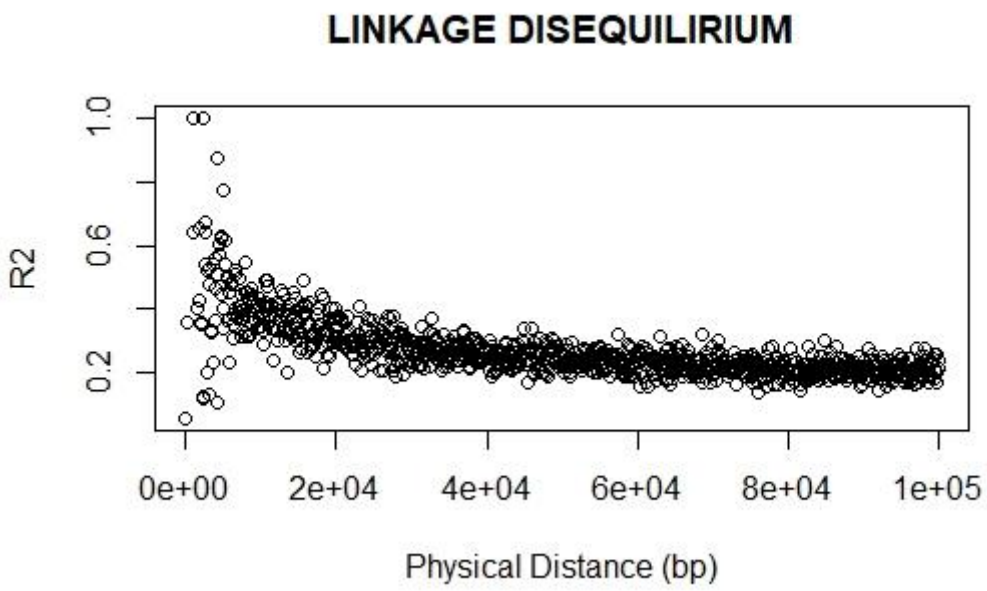
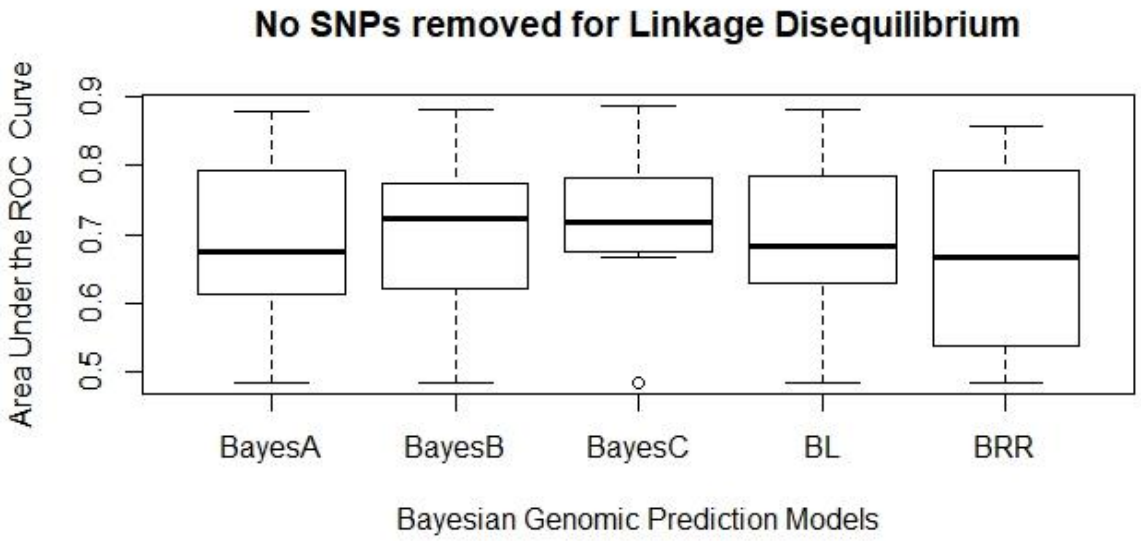
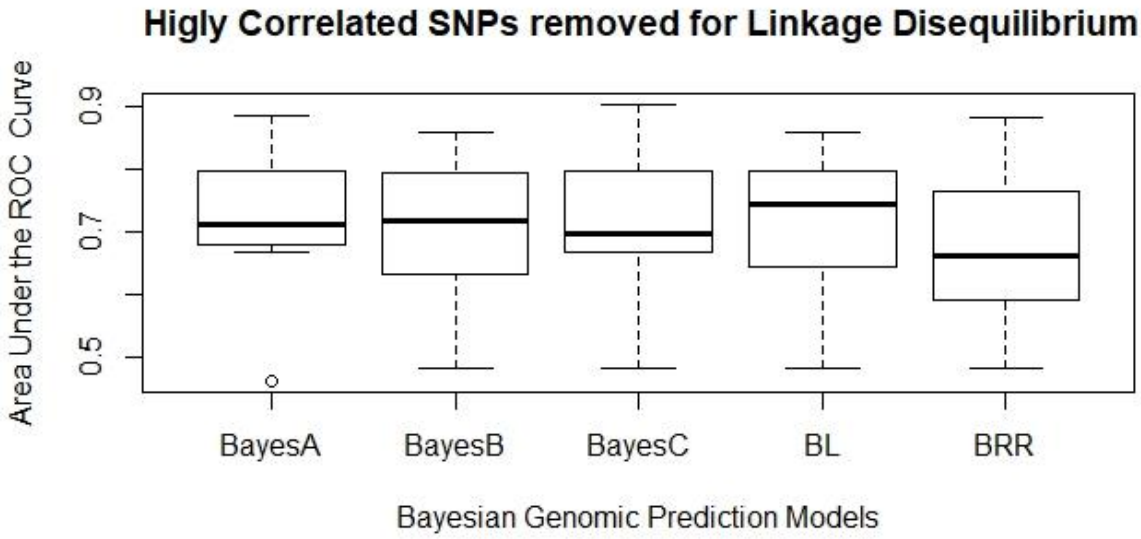
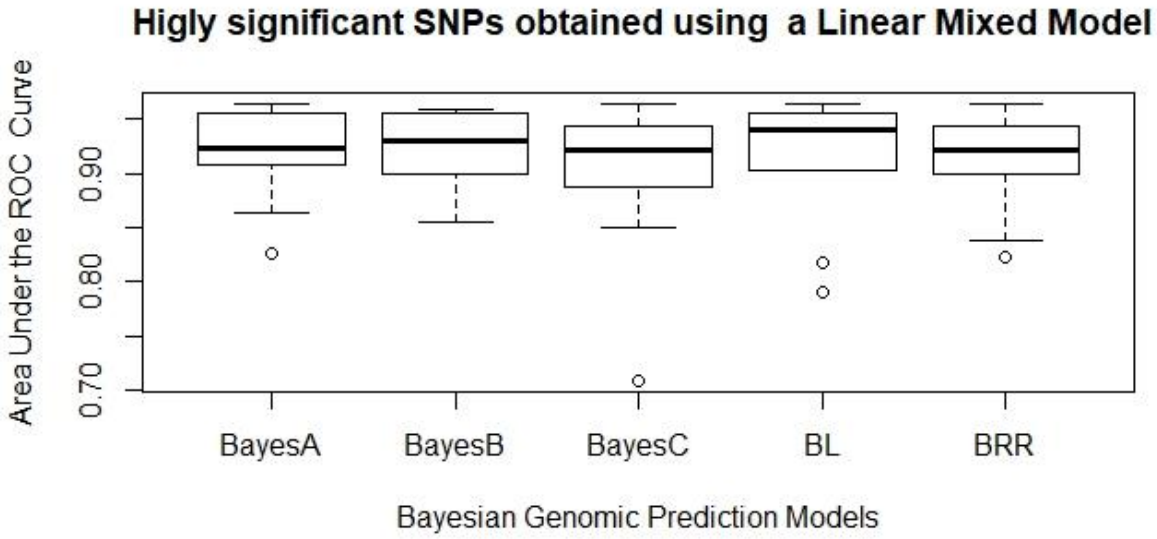


Figure 1. Decay of linkage disequilibrium over physical distance





1



2

3

4 Figure 2 Prediction accuracies obtained for Bayesian genomic prediction models from 10-fold
5 cross validations.

6

Table 1: Results of Bayesian learning models obtained from different experimental settings over 10 fold cross validation procedure.

Model	No. SNPs	AUC (SD)
No SNPs removed for Linkage Disequilibrium		
BayesA	40642	0.684 (0.119)
BayesB	40642	0.710 (0.117)
BayesC	40642	0.724 (0.113)
BL	40642	0.700 (0.125)
BRR	40642	0.673 (0.130)
Highly Correlated SNPs removed for Linkage Disequilibrium (r^2>0.7)		
BayesA	25254	0.727 (0.118)
BayesB	25254	0.699 (0.132)
BayesC	25254	0.721 (0.119)
BL	25254	0.711 (0.134)
BRR	25254	0.679 (0.128)
SNPs selected by using Linear Mixed Model results (P<0.05)		
BayesA	2401	0.919 (0.044)
BayesB	2401	0.922 (0.039)
BayesC	2401	0.898 (0.074)
BL	2401	0.915 (0.061)
BRR	2401	0.910 (0.047)
SNPs selected by using Linear Mixed Model results (P<0.05) obtained from training populations		
BayesA	2120	0.613 (0.126)
BayesB	2120	0.624 (0.124)
BayesC	2120	0.694 (0.081)
BL	2120	0.619 (0.114)
BRR	2120	0.673 (0.103)

References

1. VanRaden PM. Symposium review: How to implement genomic selection. Journal of dairy science 2020; 103.6: 5291-5301.

2. Gutierrez-Reinoso MA, Aponte PM, Garcia-Herreros M. Genomic Analysis, Progress and Future Perspectives in Dairy Cattle Selection: A Review. Animals 2021; 11.3: 599.

3. Meuwissen T, Hayes B, Goddard M. Genomic selection: A paradigm shift in animal breeding. Animal frontiers 2016; 6.1: 6-14.

4. Baker LA, Momen M, Chan K, Bollig N, Lopes FB, Rosa G J, Muir P. Bayesian and machine learning models for genomic prediction of anterior cruciate ligament rupture in the canine model. G3: Genes, genomes, genetics 2020; 10.8: 2619-2628.

5. Grinberg NF, Orhobor OI, King RD. An evaluation of machine-learning for predicting phenotype: studies in yeast, rice, and wheat. Machine Learning 2020; 109.2: 251-277.

6. Shi S, Li X, Fang L, Liu A, Su G, Zhang Y, Zhang S. Genomic Prediction Using Bayesian Regression Models With Global–Local Prior. Frontiers in Genetics 2021; 12: 426.

7. Gianola D. Priors in whole-genome regression: the Bayesian alphabet returns. Genetics 2013; 194.3: 573-596.

8. Suárez-Vega A, Gutiérrez-Gil B, Benavides J, Perez V, Tosser-Klopp G, Klopp C, Arranz JJ. Combining GWAS and RNA-Seq approaches for detection of the causal mutation for hereditary junctional epidermolysis bullosa in sheep. PLoS One 2015; 10.5: e0126416.

9. Mömke S, Kerkmann A, Wöhlke A, Ostmeier M, Hewicker-Trautwein M, Ganter M, Distl O. A frameshift mutation within LAMC2 is responsible for Herlitz type junctional epidermolysis bullosa (HJEB) in black headed mutton sheep. PloS one 2011; 6.5: e18943.

10. Sartelet A, Harland C, Tamma N, Karim L, Bayrou C, Li W, Charlier C. A stop-gain in the laminin, alpha 3 gene causes recessive junctional epidermolysis bullosa in Belgian Blue cattle. Animal genetics 2015; 46.5: 566-570.

11. Kerkmann A, Ganter M, Frase R, Ostmeier M, Hewicker-Trautwein M, Distl O. Epidermolysis bullosa in German black headed mutton sheep. Berliner und Munchener Tierarztliche Wochenschrift 2010; 123.9-10: 413-421.

12. Ostmeier M, Kerkmann A, Frase R, Ganter M, Distl O, Hewicker-Trautwein M. Inherited junctional epidermolysis bullosa (Herlitz type) in German black-headed mutton sheep. Journal of comparative pathology 2012; 146.4: 338-347.

13. Milenkovic D, Chaffaux S, Taourit S, Guérin G. A mutation in the LAMC2 gene causes the Herlitz junctional epidermolysis bullosa (H-JEB) in two French draft horse breeds. Genetics Selection Evolution 2003; 35.2: 249-256.

14. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American journal of human genetics 2007; 81.3: 559-575.

15. Meuwissen TH, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics 2001; 157.4: 1819-1829.

16. Park T, Casella G. The bayesian lasso. Journal of the American Statistical Association 2008; 103.482: 681-686.

17. Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM. Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. PLoS genetics 2015; 11.4: e1004969.

18. Pérez P, de Los Campos G. Genome-wide regression and prediction with the BGLR statistical package. Genetics 2014; 198.2: 483-495.

19. Schulz-Streeck T, Ogutu JO, Piepho HP. Pre-selection of markers for genomic selection. In BMC proceedings 2011; 5.3: 1-4. BioMed Central.

20. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nature genetics 2012; 44.7: 821-824.

21. Calus MPL, De Haas Y, Veerkamp RF. Combining cow and bull reference populations to increase accuracy of genomic prediction and genome-wide association studies. Journal of Dairy Science 2013; 96.10: 6703-6715.

22. Fanny M, Andrea R, Pascal C. Evaluating the Interpretability of SNP Effect Size Classes in Bayesian Genomic Prediction Models. In HUMAN HEREDITY 2021; 85.2: 86-86.

23. Bang SJ, Kim YG, Park T. Joint selection of SNPs for improving prediction in genome-wide association studies. In 2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops 2012: 852-858.

24. Xiang R, Breen E, Prowse-Wilkins C, Chamberlain A, Goddard, M. Bayesian genome-wide analysis of cattle traits using variants with functional and evolutionary significance. bioRxiv 2021.

25. Meuwissen T, van den Berg I, Goddard M. On the use of whole-genome sequence data for across-breed genomic prediction and fine-scale mapping of QTL. Genetics Selection Evolution 2021; 53.1: 1-15.

26. Guo P, Zhu B, Niu H, Wang Z, Liang Y, Chen Y, Li J. Fast genomic prediction of breeding values using parallel Markov chain Monte Carlo with convergence diagnosis. BMC bioinformatics 2018; 19.1: 1-11.

27. Calus MP. Right-hand-side updating for fast computing of genomic breeding values. Genetics Selection Evolution 2014; 46.1: 1-11.