Scientific Research Publishing

Agricultural Sciences, 2018, 9, 1470-1478
http://www.scirp.org/journal/as
ISSN Online: 2156-8561
ISSN Print: 2156-8553

Use of Observed Genomic Information to Infer Linkage Disequilibrium between Markers and QTLs

El Hamidi Hay[1]*, Romdhane Rekaya[2,3,4]

[1]USDA Agricultural Research Service, Fort Keogh Livestock and Range Research Laboratory, Miles City, MT, USA
[2]Department of Animal and Dairy Science, The University of Georgia, Athens, GA, USA
[3]Department of Statistics, The University of Georgia, Athens, GA, USA
[4]Institute of Bioinformatics, The University of Georgia, Athens, GA, USA
Email: *elhamidi.hay@ars.usda.gov

How to cite this paper: Hay, E.H. and Rekaya, R. (2018) Use of Observed Genomic Information to Infer Linkage Disequilibrium between Markers and QTLs. *Agricultural Sciences*, **9**, 1470-1478.
https://doi.org/10.4236/as.2018.911102

Received: October 19, 2018
Accepted: November 23, 2018
Published: November 26, 2018

Abstract

Conducting genomic selection in admixed populations is challenging and its accuracy in this case largely depends on the persistence of linkage disequilibrium between single nucleotide polymorphisms (SNP) and quantitative trait loci (QTL). Inferring linkage disequilibrium (LD) between SNP markers and QTLs could be important in understanding the change of SNP marker effects across different breeds. Predicting the change in linkage disequilibrium between markers and QTLs across two divergent breeds was explored using information from the genotype data. Two different models (M1, M2) that differ in the definition of the explanatory variables were used to infer the level of LD between SNP markers and QTLs using all markers in the panel or windows of fixed number of markers. Three simulation scenarios were conducted using different number of SNPs and QTLs. In the first scenario, the resulting coefficient of determination (R^2) was 0.65 and 0.52 using M1 and M2, respectively. In the second scenario, average R^2 equaled 0.12 using all markers in the panel and 0.25 using 100 marker windows. Across the three simulation scenarios, it was clear that a significant portion of the variation in the change in LD between SNP markers and QTLs could be explained by information already available in the observed SNP marker data.

Keywords

Genomic Selection, Linkage Disequilibrium, SNP

1. Introduction

Genomic selection is a type of marker assisted selection which involves the estimation of genomic breeding values (GEBV) based on a large number of markers across the genome [1]. Genomic selection relies on the assumption that all relevant quantitative loci (QTL) are in linkage disequilibrium (LD) with genotyped SNP markers. Thus, linkage disequilibrium or the non-random association of alleles at different loci [2] across genotyped markers and between the later and QTLs will fundamentally condition the efficiency of the association analysis and it is of great importance in QTL mapping, genomic selection and genome wide association studies. Although the strength of LD between genotyped SNP markers is easy to calculate, inferring the level of LD between SNP markers and QTLs is a complex problem due to the unavailability of QTL genotypes in the majority of genomic association studies. Although the knowledge of the QTL(s) genotypes or their LD with SNP markers in the panel is not needed in association studies, such information could be of great interest in some applications such as multi-breed and crossbred genomic selection.

Genomic selection has been successful in prediction of genomic breeding values. However this success did not extend to admixed breeds or crossbreds. Several studies showed that the structure of the reference population strongly impacts the accuracy of genomic predictions [3] [4] [5] [6]. Moreover, SNP marker estimates derived from one breed have little to no predictive power of GEBVs of animals in a different breed [4] [7]. A potential solution would be to use a pooled multi-breed reference population to predict GEBV of animals in other breeds or crossbred animals [8] [9] [10] [11] [12]. This method showed promising results in improving prediction accuracy in the case when a breed has a limited number of records. However, the performance of this approach, as expected, depends largely on the genetic similarity between components of the admixed population.

Although simple in its concept, the multi-breed reference population approach makes strong genetic and population structure assumptions. In its most basic formulation, it assumes a genetically homogenous population where SNP marker effects are constant across sub-populations or breeds. Further, it assumes that linkage disequilibrium (LD) between SNPs and QTLs is the same across the reference and validation populations. Although that is the case for within breed genomic selection, such assumption is often violated when breeds with different genetic structure and background are being considered. This genetic difference between breeds is manifested by varying allele frequencies for markers and QTLs, change in LD strength and structure, and linkage phase [13] [14] [15] [16]. Furthermore, several studies have evaluated LD blocks in various population structures and reported differences in the extent of LD. For example, Shifman *et al.* (2003) showed that LD was several folds higher in isolated population than out bred populations very likely due to higher inbreeding [17]. Similarly, Lindbladtoh *et al.* (2005) reported, as expected, larger LD blocks within breeds than across breeds [18]. Hay and Rekaya (2015) showed that accommodating the potential change in SNP effects between the different components of an admixed population, increased accuracies of genomic prediction [19] [20]. When change in SNP effects was directly modeled, substantial increase in accuracies was observed compared to the classical pooled data approach. Unfortunately, such model suffers from high dimensionality and numeral instability especially in presence for large number of SNPs. Their indirect approach to account for change in SNP effects was based on heuristically developed structural model using available information on marker genotypes. Although it remedies the problems associated with the direct approach and yields better results than the classical pooled data model, its performance are significantly lower than the direct approach. These results indicate that change in the distribution of SNP marker genotypes between sub-populations is likely to carry relevant information about change of LD structure and strength between markers and QTLs across components of the admixed population that could be garnished to account for change in SNP effects. Since genomic selection largely depends on LD structure, it is of great importance to be able to evaluate and infer the magnitude of change in LD between SNP markers and QTLs in different populations. This information might shed some light on the change of SNP effects across different breeds or lines and how to adjust for this change. The objective of this study is to evaluate and infer the change of LD between markers and QTLs across two breeds using simulated data sets.

2. Materials and Methods

As indicated in the introduction section, genetic heterogeneity between sub-populations leads to change in estimates of SNP effects due to change in LD between observed markers and putative QTLs. The foundation of genome wide associations is that QTL effects can be inferred indirectly through their correlation with genotyped markers. Across sub-population, LD structure between markers (LD_{M-M}) as well between markers and QTLs (LD_{M-Q}). changes. Consequently, it is reasonable to postulate that change in LD between SNP markers across two sub-populations (ΔLD_{M-M}) could explain, at least partially, the change in LD between markers and QTLs (ΔLD_{M-Q}).

In order to evaluate this hypothesis, several small-scale simulations were carried out. In these simulations, the genotypes of the QTL(s) and associated SNPs markers were all assumed known. Thus, LD between SNP markers and QTL(s) was available. In all cases our goal was to test the ability of ΔLD_{M-M} to predict ΔLD_{M-Q}.

<u>Simulation scenarios:</u> Three simulation scenarios with varying number of SNP markers and QTLs were carried out to test the postulated hypothesis. In all cases, two divergent sub-populations for a trait with heritability equal to 0.5 were generated. A full description of the simulation parameters are presented in the next section. Two models (M1, M2) were evaluated in their ability to predict the change in ΔLD_{M-Q}:

$$\Delta LD_{M_k-Q} = a_0 + a_1 M_k + a_2 S_k + e_k \qquad \text{(M1)}$$

$$\Delta LD_{M_k-Q} = b_0 + b_1 MR_k + b_2 SR_k + e_k \tag{M2}$$

where ΔLD_{M_k-Q} is the difference of LD between marker *k* and the QTL across the two sub-populations, M_k and S_k are the mean and standard deviation of the difference of LD between marker *k* and all the remaining SNPs or a 100 adjacent SNP markers, respectively. MR_k and SR_k are the same as M_k and S_k, except they represent the relative mean and standard deviation of the difference in LD, a_j and $b_j\left(j=0,1,2\right)$ are unknown regression coefficients. To evaluate the fit of the model, the coefficient of determination R^2 was calculated.

Linkage disequilibrium across SNP markers and between SNP markers and QTLs in both lines was calculated using the r^2 coefficient as proposed by [2] using the following general equation.

$$r^2 = \frac{D^2}{f\left(A\right)f\left(a\right)f\left(B\right)f\left(b\right)}$$

where *D* is calculated as $D = f\left(AB\right) - f\left(A\right)f\left(B\right)$ and $f\left(AB\right), f\left(A\right), f\left(a\right)$, $f\left(B\right)$ and $f\left(b\right)$ are observed frequencies of haplotype AB and of alleles A, a, B, and b, respectively. The higher r^2, the stronger the linkage disequilibrium.

For all cases and for both models, unknown coefficients were estimated using the proc glm of SAS software [21].

Data simulation: QMSim software [22] was used for data simulation. A historical population of unrelated individuals was simulated and used as a base population for two pure breeds (A and B). Breeds A and B consisted of 1677 and 1668 individuals respectively. The simulated genome consisted of 1 chromosome, with varying number of QTLs and varying number of SNP markers with equal spacing of an average 50 Kb. Minor allele frequency was set to 0.05. QTL additive effects were sampled from a gamma distribution with shape and scale parameter equal to 0.4. Phenotypes were simulated based on a heritability of 0.5. Three simulation scenarios were carried out. In the first scenario, 10 SNP markers and 1 QTL were considered. The QTL was positioned in close proximity to SNP marker 5. In the second scenario the number of markers was increased to 300 SNP markers and also increased the number of QTLs to 3. Finally, in the last simulation scenario, the number of SNP markers was increased to 3000 SNPs and the number of QTLs increased to 30. These QTLs were randomly positioned across the genome. All SNP markers were used in the inference of ΔLD_{M-Q}. In both statistical models (M1, M2), LD between marker *k* and all the remaining SNPs or 100 adjacent SNP marker windows were implemented.

3. Results and Discussion

Linkage disequilibrium between the SNP markers and the QTL for lines A and B as well as ΔLD_{M-Q} for the first simulation scenario are presented in **Table 1**. Since the QTL was placed in the center of the simulated segment, the LD_{M-Q} was, as expected, higher for markers 4, 5 and 6. **Figure 1** shows the trend of LD between the SNP markers and QTL for the two lines. Similarly, the LD between

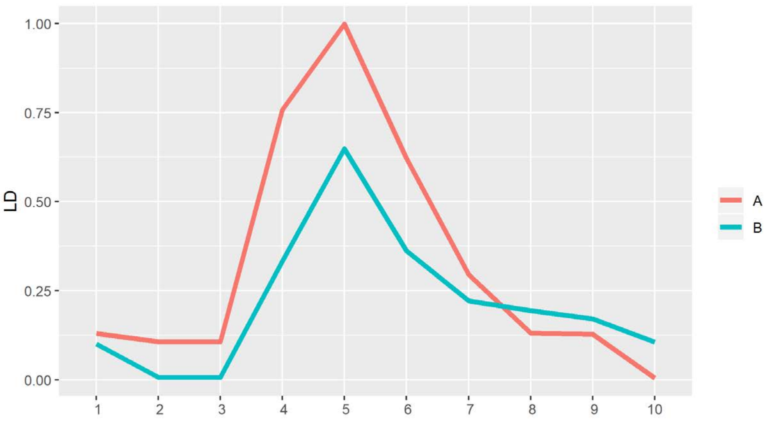

Figure 1. Linkage disequilibrium between markers and QTL for breeds A and B.

Table 1. Linkage disequilibrium between markers and QTL for breed A and B in the first simulation scenario.

$LD_{M\text{-}Q}(A)$[1]	$LD_{M\text{-}Q}(B)$[2]	$\Delta LD_{M\text{-}Q}$[3]
0.131	0.101	0.029
0.107	0.008	0.026
0.107	0.008	0.024
0.758	0.333	0.419
0.999	0.649	0.350
0.622	0.363	0.259
0.296	0.222	0.074
0.132	0.195	−0.063
0.128	0.172	−0.043
0.005	0.106	−0.051

[1]LD between marker and QTL for breed A; [2]LD between marker and QTL for breed B, [3]Difference in marker and QTL LD between breed A and B.

markers (LD_{M-M}) for the two lines as well as the difference in LD ΔLD_{M-M} were calculated. To infer ΔLD_{M-Q} between the two breeds, the mean and standard deviation of ΔLD_{M-M} were calculated and later used as explanatory variables in the regression model (Table 2). Fitting model M1 resulted in an R^2 of 0.65; indicated that the mean and standard deviation of ΔLD_{M-M} explained around two thirds of the variation in ΔLD_{M-Q} between breeds A and B. On the other hand, fitting model M2 resulted in 25% decrease in R^2 (0.52). Although M2 resulted in a decrease in R^2, the model still was able to explain a significant portion of the variation in ΔLD_{M-Q} across the two breeds. When the number of SNP markers and QTLs were increased to 30 and 3, respectively (second simulation scenario), the coefficients of determination tended to decrease using either all the SNP markers (300) or fixed size widows of 100 SNPs to calculate the parameters of the regression model. Table 3 shows the resulting coefficients of

Table 2. Mean and standard deviation of change of LD between markers in the first simulation scenario[1].

$\Delta LD_{M\text{-}M}$[1]	
mean	SD
0.010	0.187
0.018	0.166
0.019	0.167
−0.044	0.198
0.014	0.259
−0.091	0.225
−0.134	0.197
−0.082	0.095
−0.087	0.088
−0.072	0.120

[1]Difference in LD of marker and marker between breeds A and B.

Table 3. Coefficient of determination for models M1 and M2 in the second simulation scenario.

$\Delta LD_{M\text{-}Q}$	M1		M2	
	All markers	100 marker window	All markers	100 marker window
QTl_1	0.14	0.26	0.07	0.03
QTL_2	0.12	0.24	0.02	0.02
QTL_3	0.12	0.27	0.01	0.01

determination (R^2) for models M1 and M2 using all markers and using fixed windows of 100 SNPs. Using M1 resulted in R^2 equal to 0.14, 0.12 and 0.12 for QTLs 1, 2 and 3 respectively using all 300 markers. In the case of using 100 marker windows, R^2 increased to 0.26 for QTL 1, 0.24 for QTL 2, and 0.27 for QTL 3. This increase in R^2 is due for at least two reasons: 1) a QTL was positioned in each 100 SNP marker window, and 2) including all 300 SNP markers where a large portion of them has no LD with the QTL, resulted in a less informative mean and standard deviation of ΔLD_{M-M} to explain variation in ΔLD_{M-Q}. The highest increase in R^2 was for QTL 3, from 0.12 to 0.27. Using M2, a substantial decrease in R^2 was observed across all QTLs using either 100 marker windows or all markers. Table 4 shows the average R^2 across all 3 markers, it is clear that M1 performed better than M2 in this simulation scenario.

In the third simulation scenario, a larger SNP panel (3000 SNPs), and a higher number of QTLs (30) were simulated. Table 4 shows the average R^2 obtained using M1and M2. Clearly, M1 performed notably better than M2 using either all markers or 100 marker windows. For example, fitting M1 using all markers resulted in an average R^2 of 0.27 compared to 0.01 for M2. It should be mentioned

Table 4. Average coefficient of determination over all QTLs for models M1 and M2 in the second and third simulation scenarios.

Genome	M1		M2	
	All markers	100 marker window	All markers	100 marker window
300 SNP,3 QTLs	0.12	0.25	0.05	0.03
3000 SNP, 30 QTLs	0.27	0.10	0.03	0.01

that M2 did not explain any variation in the change of ΔLD_{M-Q} across breed A and B.

Across the three simulation scenarios, it is clear that a significant portion of the variation in variation in ΔLD_{M-Q} could be explained by information already available in the observed SNP marker data. Furthermore, the statistical model as well as the extent of the window of SNPs considered in the calculation of the parameters of the regression line plays a crucial role in estimating change in LD between markers and QTLs in both breeds. Based on the results of this simulation study and the structure of LD generated, it seems that small windows are preferable. This is true because including large number of SNPs with little to no LD with the QTL(s) will render the mean and standard deviation non-informative about the variation in ΔLD_{M-Q}. Using real data, the situation will be more complex due to a larger number of SNP markers and QTLs where the latter have a random and unknown distribution. In such case, information about LD blocks should be used in determining the length of SNP windows to be used. Additionally, the relationship between ΔLD_{M-Q} and the observed information in the SNP genotypes could be non-linear and cannot be approximated well with simple regression models.

4. Conclusion

In this simulation study, inferring change of linkage disequilibrium between marker and QTL between two pure breeds proved to be possible. This might help in inferring the change of SNP marker effects when having different breeds or lines in the population. Whether or not this could be used in genomic selection in the case of admixed populations, further testing and research is required.

Statement

Mention of trade names or commercial products in this publications solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA.

USDA is an equal opportunity provider and employer.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Meuwissen, T., Hayes, B. and Goddard, M. (2001) Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics*, **157**, 1819-1829.

[2] Hill W. and Robertson, A. (1968) Linkage Disequilibrium in Finite Populations. *Theoretical and Applied Genetics*, **38**, 226-231. https://doi.org/10.1007/BF01245622

[3] VanRaden, P.M. (2008) Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science*, **91**, 4414-4423. https://doi.org/10.3168/jds.2007-0980

[4] Hayes, B.J., Bowman, P.J., Chamberlain, A.C., Verbyla, K. and Goddard, M.E. (2009) Accuracy of Genomic Breeding Values in Multi-Breed Dairy Cattle Populations. *Genetics Selection Evolution*, **41**, 51. https://doi.org/10.1186/1297-9686-41-51

[5] VanRaden, P., Van Tassell, C., Wiggans, G., Sonstegard, T., Schnabel, R., Taylor, J., *et al.*, (2009) Invited Review: Reliability of Genomic Predictions for North American Holstein bulls. *Journal of Dairy Science*, **92**, 16-24. https://doi.org/10.3168/jds.2008-1514

[6] Erbe, M., Hayes, B., Matukumalli, L., Goswami, S., Bowman, P., Reich, C., *et al.*, (2012) Improving Accuracy of Genomic Predictions within and between Dairy Cattle Breeds with Imputed High-Density Single Nucleotide Polymorphism Panels. *Journal of Dairy Science*, **95**, 4114-4129. https://doi.org/10.3168/jds.2011-5019

[7] Pryce, J., Gredler, B., Bolormaa, S., Bowman, P., Egger-Danner, C., Fuerst, C., *et al.*, (2011) Genomic Selection Using a Multi-Breed, Across-Country Reference Population. *Journal of Dairy Science*, **94**, 2625-2630. https://doi.org/10.3168/jds.2010-3719

[8] Heringstad, B., Guosheng, S., Solberg, T., Guldbrandtsen, B., Svendsen, M. and Lund, M.S. (2011) Genomic Predictions Based on a Joint Reference Population for Scandinavian Red Breeds. 62*nd Annual Meeting of the European Federation of Animal Science*, *EAAP*, 29-29.

[9] Daetwyler, H., Kemper, K., Van der Werf, J. and Hayes, B. (2012) Components of the Accuracy of Genomic Prediction in a Multi-Breed Sheep Population. *Journal of Animal Science*, **90**, 3375-3384. https://doi.org/10.2527/jas.2011-4557

[10] Olson, K., VanRaden, P. and Tooker, M. (2012) Multibreed Genomic Evaluations Using Purebred Holsteins, Jerseys, and Brown Swiss. *Journal of Dairy Science*, **95**, 5378-5383. https://doi.org/10.3168/jds.2011-5006

[11] Zhou, L., Heringstad, B., Su, G., Guldbrandtsen, B., Svendsen, M., Grove, H., *et al.* (2014) Genomic Predictions Based on a Joint Reference Population for the Nordic Red Cattle Breeds. *Journal of Dairy Science*, **97**, 4485-4496. https://doi.org/10.3168/jds.2013-7580

[12] Hoze, C., Fritz, S., Phocas, F., Boichard, D., Ducrocq, V. and Croiseau, P. (2014) Efficiency of Multi-Breed Genomic Selection for Dairy Cattle Breeds with Different Sizes of Reference Population. *Journal of Dairy Science*, **97**, 3918-3929. https://doi.org/10.3168/jds.2013-7761

[13] Goddard, M. (2009) Genomic Selection: Prediction of Accuracy and Maximisation of Long Term Response. *Genetica*, **136**, 245-257. https://doi.org/10.1007/s10709-008-9308-0

[14] De Roos, S.A., Hayes, B.J., Spelman, R. and Goddard, M.E. (2008) Linkage Disequilibrium and Persistence of Phase in Holstein Friesian, Jersey, and Angus Cattle. *Genetics*, **179**, 1503-1512. https://doi.org/10.1534/genetics.107.084301

[15] Kizilkaya, K., Fernando, R. and Garrick, D. (2010) Genomic Prediction of Simulated Multibreed and Purebred Performance Using Observed Fifty Thousand Single

Nucleotide Polymorphism Genotypes. *Journal of Animal Science*, **88**, 544-551. https://doi.org/10.2527/jas.2009-2064

[16] Wientjes, Y.C., Veerkamp, R.F. and Calus, M.P. (2013) The Effect of Linkage Disequilibrium and Family Relationships on the Reliability of Genomic Prediction. *Genetics*, **193**, 621-631. https://doi.org/10.1534/genetics.112.146290

[17] Shifman, S., Kuypers, J., Kokoris, M., Yakir, B. and Darvasi, A. (2003) Linkage Disequilibrium Patterns of the Human Genome across Populations. *Human Molecular Genetics*, **12**, 771-776. https://doi.org/10.1093/hmg/ddg088

[18] Lindblad-Toh, K., Wade, C.M., Mikkelsen, T.S., Karlsson, E.K., Jaffe, D.B., Kamal, M., *et al.* (2005) Genome Sequence, Comparative Analysis and Haplotype Structure of the Domestic Dog. *Nature*, **438**, 803. https://doi.org/10.1038/nature04338

[19] Hay, E.H. and Rekaya, R. (2015) A Multi-Compartment Model for Genomic Selection in Multi-Breed Populations. *Livestock Science*, **177**, 1-7. https://doi.org/10.1016/j.livsci.2015.03.027

[20] Hay, E.H. and Rekaya, R. (2015) A Structural Model for Genetic Similarity in Genomic Selection of Admixed Populations. *Livestock Science*, **181**, 72-76. https://doi.org/10.1016/j.livsci.2015.10.009

[21] S. Institute (1990) SAS/STAT User's Guide: Version 6. Vol. 2.

[22] Sargolzaei, M. and Schenkel, F.S. (2009) QMSim: A Large-Scale Genome Simulator for Livestock. *Bioinformatics*, **25**, 680-681. https://doi.org/10.1093/bioinformatics/btp045