ORIGINAL SCIENTIFIC PAPER

The estimation of linkage disequilibrium decay in Croatian Holstein cattle: potential for genomic selection

Marija Špehar[1], Zdenko Ivkić[1], Drago Solić[1], Ante Kasap[2]

[1]*Croatian Agency for Agriculture and Food, Svetošimunska 25, Zagreb, Croatia (marija.spehar@hapih.hr)*
[2]*University of Zagreb Faculty of Agriculture, Svetošimunska cesta 25, Zagreb, Croatia*

Abstract
The objective of this study was to estimate linkage disequilibrium (LD) in 101 Croatian Holstein cows based on Illumina Bovine50K SNP chip. The average autosomal whole-genome LD (r^2) was 0.16 and for SNPs 100 Kbp apart 0.30. The average within-chromosomal LD ranged from 0.12 for BTA28 to 0.23 for BTA20, with no straightforward relation between chromosome length and r^2. The decay of LD showed an exponential trend with physical distance showing a very steep decay of LD within the first 150 Kbp. The results implicate a sufficient density of the 50K SNP chip for successful implementation of genomic selection (GS) and genome wide association studies (GWAS) in the population of Holstein cattle in Croatia.

Key words: Holstein cattle, linkage disequilibrium, SNP, genomic selection

Introduction

The development of genotyping technology and reduced whole-genome (SNPs) genotyping costs in many livestock species strongly contributed to the interest in implementation of genomic selection (GS) (Meuwissen et al., 2001), genome-wide association studies (GWAS) (Meuwissen and Goddard, 2000; Hirschhorn and Daly, 2005) and understanding genomic architecture and historical population structure (Hayes et al., 2003). The success of these genomic techniques strongly depends on the extent of linkage disequilibrium (LD) and its rate of decline with distance between loci. LD refers to non-random association between alleles at different loci mostly due to physical proximity. Estimates of LD are affected by population history and evolutionary forces (Ardlie et al., 2002), sample size, marker type, marker density and distribution, and strictness of SNP filtering (Bohmanova et al., 2010). In livestock, the most common measures of LD are the squared correlation coefficient (r^2) (Hill and Robertson, 1968) which is equivalent to the covariance and the correlation between alleles at two different loci and the normalized D' (Lewontin, 1964). r^2 is less sensitive to sample (population) size than D', and D' tends to be inflated with small sample sizes and/or low allele frequencies (Bohmanova et al., 2010), making r^2 the preferred measure for LD.
Livestock populations have smaller effective population sizes than human and therefore LD is expected to extent larger genetic distances (Farnir et al., 2000). In cattle, LD in highly selected populations is extended over larger distances (Farnir et al., 2000) compared to moderately selected populations (Thevenon et al., 2007). Studies used SNPs for evaluating the extent of LD in intensively selected populations (such as Holstein) revealed moderate levels of LD ($r^2 \geq 0.2$) between markers with inter-marker distance up to 100 Kbp (McKay et al., 2007; Marques et al., 2008; Sargolzaei et al., 2008). Khatkar et al. (2008) reported $r^2 \geq 0.2$ between SNPs less than 60 Kbp apart in Australian Holstein. The objective of this study was to estimate the genome-wide level of LD in Croatian Holstein cattle in order to examine the potential of GWAS and GS.

Material and methods

The data included 101 Holstein females from 32 herds that were progenies of 45 sires. Females were genotyped using the Illumina BovineSNP50K BeadChip containing 52,445 SNPs. Only autosomal SNPs were included in the analysis. Markers with a minor allele frequency <5% (n=9,047), >10% missing genotypes (n=791), >10% missingness per animal (n=0), and strong deviation ($P<10^{-6}$) from Hardy-Weinberg equilibrium (n=11) were excluded from the analysis. This editing resulted in 40,763 SNPs that passed the above control and were used for further analysis. The PLINK (version 1.9: Purcell et al., 2007; Chang et al., 2015) was used to calculate LD between pairs of SNPs using the r^2 statistics. The dplyr (Wickham et al., 2021) and ggplot (Wickham, 2016) packages in the R programming environment (R Core Team, 2020) were used for post hoc statistical analysis of genome wide LD.

Results and discussion

This study provides an overview of LD in Croatian Holstein dairy cattle using a 50K SNP panel. We used r^2 value to estimate the extent of LD due to small sample size. According to Bohmanova et al. (2010), estimates of r^2 are not influenced by sample size when if at least 55 animals were used in the calculation. The overall average genome-wide LD was 0.16. To examine the decay of LD with physical distance, syntenic SNP pairs on autosomes were sorted into bins (intervals) based on their inter-marker distance. The average r^2 was calculated for each bin (Table 1). A moderate level of r^2 (0.30) was estimated for distances shorter than 100 Kbp. The average r^2 has been declined from 0.15 to 0.10 when moving from 100 to 800 Kbp. LD comparisons between different studies are difficult due to differences in sample size, LD measures, marker types and density of markers (Bohmanova et al., 2010). The results of this study revealed similar values at a distance of 100 Kbp as reported MacKay et al. (2007) for Holstein in their study of eight cattle breeds (Angus, Charolais, Dutch Black and White Dairy, Holstein, Japanese Black, Limousin, Brahman, and Nelore). Espigolan et al. (2013) genotyped Nellore cattle and found a mean LD (r^2) between adjacent markers of 0.17. Marques et al. (2008) studied BTA14 in Holstein cattle and found a moderate LD level (0.2) separated by 100 Kbp. An average r^2 estimate of 0.16 (in 60-100 Kbp bins) was reported for North American Holstein bulls (Bohmanova et al., 2010). In the study by El Hou et al. (2021) conducted on three French beef cattle breeds (Charolaise, Blonde d'Aquitaine, Limousine), the average r^2 varied from 0.50 (distances smaller than 15 Kbp) to less than 0.10 (distances greater than 120 Kbp). On the other hand, r^2 ranged between 0.231 (distances between 0 and 50 Kbp) and 0.065 (distances between 150 and 200 Kbp) in Korean Hanwoo cattle (Lee and Kong, 2021). LD values decrease as the distance between markers on the genome increases. Bovine studies using SNP data have shown that the average LD was close to zero for distances between markers greater than 500 Kbp (McKay et al., 2007).

Table 1. Genome wide linkage disequilibrium (LD) classified by inter-marker distance

Distance (Kbp)	N	avg (r^2)	sd (r^2)	Distance (Kbp)	N	avg (r^2)	sd (r^2)
[0,100]	77153	0.30	0.32	(500,600]	25619	0.11	0.14
(100,200]	65967	0.15	0.19	(600,700]	15121	0.11	0.14
(200,300]	62968	0.12	0.16	(700,800]	8818	0.10	0.13
(300,400]	55611	0.11	0.14	(800,900]	5229	0.10	0.13
(400,500]	40370	0.11	0.14	(900,1000]	2952	0.09	0.13

Graphical representation of LD decay was obtained by plotting the average LD (expressed as r^2) vs. average between-marker distance of bins spanning 100 Kbp (Figure 1). Decay of LD showed a clear exponential trend with physical distance. The changes of r^2 revealed a very steep decay of LD within the first 150 Kbp and intermediate decay from 150 to 250 Kbp. Decay after 250 Kbp was negligible.

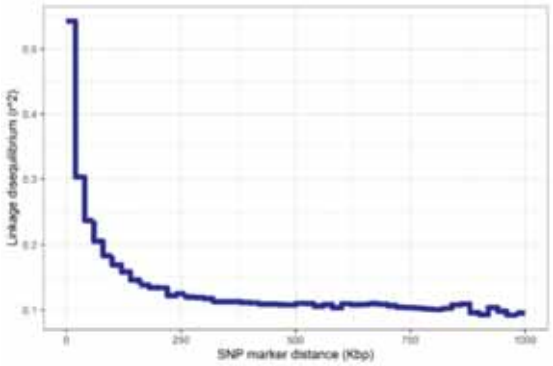

Figure 1. Decay of r^2 as a function of physical distance

Table 2 shows the mean r^2 of the 29 individual autosomes (BTA). Some chromosomes had higher LD than others, however there was no clear relationship between chromosome length and r^2. A linear relationship was observed between LD (r^2) and chromosome length in the study of Sahiwal cattle (Mustafa et al., 2018). On the other hand, Bohmanova et al. (2010) found no association between LD level and chromosome size. In this study, the average r^2 between adjacent SNP ranged from 0.12 (BTA28) to 0.23 (BTA20). The median for r^2 was 0.15 and was observed for BTA3, BTA5, BTA9, BTA17, BTA22, BTA23, BTA26, and BTA29. In the study conducted by Bohmanova et al. (2010), the average r^2 was slightly higher, ranging from 0.17 (BTA27, BTA28 and BTA 29) to 0.25 (BTA 7 and BTA14). The average LD between SNPs ranged from 0.002 to 0.19 for r^2 across all autosomes in the study conducted by Mustafa et al. (2018) in Sahiwal cattle. The extent of LD was considerably higher in this population than in many other populations of small ruminants, especially sheep (e.g., French Lacaune, Baloche et al., 2014; Chinese Merino, Liu et al., 2017), which is logical considering the history of the breeds under consideration (different selection intensities).

Table 2. Linkage disequilibrium (LD) classified by inter-marker distance of 100 Kbp on 29 autosomes (BTA)

Chr	avg (r^2)	sd (r^2)	Chr	avg (r^2)	sd (r^2)	Chr	avg (r^2)	sd (r^2)
1	0.16	0.22	11	0.14	0.19	21	0.18	0.26
2	0.17	0.22	12	0.13	0.18	22	0.15	0.21
3	0.15	0.20	13	0.20	0.26	23	0.15	0.21
4	0.14	0.20	14	0.18	0.22	24	0.14	0.18
5	0.15	0.20	15	0.13	0.17	25	0.13	0.17
6	0.17	0.23	16	0.19	0.24	26	0.15	0.20
7	0.16	0.21	17	0.15	0.20	27	0.14	0.22
8	0.16	0.22	18	0.13	0.18	28	0.12	0.17
9	0.15	0.21	19	0.21	0.28	29	0.15	0.21
10	0.17	0.21	20	0.23	0.28			

Evaluation of genome-wide LD and pattern of LD decay can be used to estimate SNP density required for GWAS studies and implementation of GS (Ardlie et al., 2002; Hayes et

al., 2008). The consensus is that roughly r^2>0.3 was considered useful LD for GWAS (Ardlie et al., 2002; Khatkar et al., 2008). For the effective GS with satisfied accuracy of 85%, the threshold of the suggested LD level (r^2) should be 0.20 (Meuwissen et al., 2001). The obtained values of LD in this population are very close to the proposed benchmarks and would probably reach them with the increased sample size. Therefore, the determined LD in this population implies a good potential for GS and GWAS, as has been confirmed in many other Holstein populations worldwide (Hayes et al., 2009).

Conclusions

This moderate extend of LD level determined in this population implies that the 50K SNP chip should be sufficient for the GS and GWAS in Croatian Holstein cattle. However, more samples should be used to get more information on this issue.

Literature

Ardlie K. G., Kruglyak L., Seielstad M. (2002). Patterns of linkage disequilibrium in the human genome. Nature Reviews Genetics. 3(4): 299-309.

Baloche G., Legarra A., Sallé G., Larroque H., Astruc J. M., *et al.* (2014). Assessment of accuracy of genomic prediction for French Lacaune dairy sheep. Journal of Dairy Science 97(2):1107-1116.

Bohmanova J., Sargolzaei M., Schenkel F. S. (2010). Characteristics of linkage disequilibrium in North American Holsteins. BMC Genomics. 11:421.

Chang C. C., Chow C. C., Tellier L. C., Vattikuti S., Purcell S. M., *et al.* (2015). Gigascience. 4(1): s13742-015-0047-8.

El Hou A., Rocha D., Venot E., Blanquet V., Philippe R. (2021). Long-range linkage disequilibrium in French beef cattle breeds. Genetics Selection Evolution. 53: 63.

Espigolan R., Fernando B., Arione A. B., Fabio R. P. S., Daniel G.M.G., Tonussi R.L., Cardoso D.F., Oliveira H.N., Tonhati H., Sargolzaei M., Schenkel F.S., Carvalheiro R., Ferro J.A., Albuquerque L.G. (2013). Study of whole genome linkage disequilibrium in Nellore cattle. BMC Genomics. 14: 305.

Farnir F., Coppieters W., Arranz J. J., Berzi P., Cambisano N., Grisart B., Karim L., Marcq F., Moreau L., Mni M., Nezer C., Simon P., Vanmanshoven P., Wagenaar D., Georges M. (2000). Extensive genome-wide linkage disequilibrium in cattle. Genome Research. 10: 220–227.

Hayes B. J., Visscher P. M., McPartlan H. C., Goddard M. E. (2003). Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Research. 13: 635–643.

Hayes B. J., Lien S., Nilsen H., Olsen H. G., Berg P., Maceachern S., Potter S., Meuwissen T. H. E. (2008). The origin of selection signatures on bovine chromosome 6. Animal Genetics. 39:105–111.

Hayes B. J., Bowman P. J., Chamberlain A. J., Goddard M. E. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. Journal of Dairy Science. 92: 433-443.

Hill W. G., Robertson A. (1968). Linkage disequilibrium in finite populations. Theoretical and Applied Genetics. 38: 226–31.

Hirschhorn J. N., Daly M. J. (2005). Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics. 6(2):95-108.

Khatkar M. S., Nicholas F. W., Collins A. R., Zenger K. R., Cavanagh J. A., Barris W., Schnabel R.D., Taylor J.F. Raadsma H.W. (2008). Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a highdensity SNP panel. *BMC* Genomics. 9:187.

Liu S., He S., Chen L., Li W., Di J., Liu M. (2017). Estimates of linkage disequilibrium and effective population sizes in Chinese Merino (Xinjiang type) sheep by genome-wide SNPs. Genes Genomics. 39(7): 733-745.

Lee G. H., Kong H. S. (2021). Linkage Disequilibrium Analysis of Hanwoo in Gyeonggi Region using Hanwoo SNP Chip. Journal of Animal Breeding and Genomics. 5(4): 235-242.

Marques E., Schnabel R. D., Stothard P., Kolbehdari D., Wang Z., Taylor J.F., Moore S.S. (2008). High density linkage disequilibrium maps of chromosome 14 in Holstein and Angus cattle. BMC Genetics. 9: 45.

McKay S. D., Schnabel R. D., Murdoch B. M., Matukumalli L. K., Aerts J., Coppieters W., Crews D., Neto E.D., Gill C.A., Gao C., Mannen H., Stothard P., Wang Z., Tassell C.P.V., Williams J.L., Taylor J.F., Moore S.S. (2007).Whole genome linkage disequilibrium maps in cattle. BMC Genetics. 8: 74.

Meuwissen T. H. E., Hayes B. J., Goddard M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics. 157: 1819-1829.

Meuwissen T. H. E., Goddard M. E. (2000). Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. Genetics. 155: 421–430.

Mustafa H., Ahmad N., Heather H. J., Eui-soo K., Khan W. A., Pasha T.N., Ali A., Kim J.J., Sonstegard T.S. (2018). Whole genome study of linkage disequilibrium in Sahiwal cattle. South African Journal of Animal Science. 48: 353–360.

Purcell S. N. B., Neale B., Todd-Brown K., Thomas L., Ferreira M. A. R., Benderba D., Mallerba J., Sklarbaa P., Bakkerba P.I.W., Dalyba M.J., Sham P.C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American journal of human genetics. 81(3): 559-575.

R Core Team (2020). R: A Language and Environment for Statistical Computing; R Foundation for Statistical Computing: Vienna, Austria. Available at: https://www.R-project.org.

Sargolzaei M., Schenkel F. S., Jansen G. B., Schaeffer L. R. (2008). Extent of linkage disequilibrium in Holstein cattle in North America. Journal of Dairy Science. 91: 2106-2117.

Thevenon S., Dayo G. K., Sylla S., Sidibe I., Berthier D., Legros H., Boichard D., Eggen A., Gautier M. (2007). The extent of linkage disequilibrium in a large cattle population of western Africa and its consequences for association studies. Animal Genetics. 38: 277–286.

Wickham H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, N.York.

Wickham H., François R., Henry L., Müller K. (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.7.

Procjena neravnoteže povezanosti genomskih markera kod holstein goveda u Republici Hrvatskoj: potencijal za provedbu genomske selekcije

Cilj ovog rada je bio procijeniti neravnoteže povezanosti genomskih markera (LD) u populaciji Holstein krava (n=101) koristeći Illumina Bovine50K SNP čip. Prosječni r^2 je bio 0,16, a 0,30 za markere međusobno udaljene 100 Kbp. Prosječni kromosomalni r^2 je bio u rasponu od 0,12 (BTA28) do 0,23 (BTA20), bez jasne veze između r^2 i veličine kromosoma. Utvrđeno je eksponencijalno opadanje LD sa porastom fizičke udaljenosti markera, s najvećim padom unutar prvih 150 Kbp. Dobiveni rezultati upućuju na dostatnu gustoću 50K SNP čipa za uspješnu provedbu genomske selekcije (GS) i genomske asocijacijske studije (GWAS) u populaciji holstein pasmine goveda u Republici Hrvatskoj.

Ključne riječi: Holstein, neravnoteža povezanosti markera, SNP, genomska selekcija