https://doi.org/10.17221/89/2023-CJAS

Genomic characterization of Serbian Holstein-Friesian cattle population

Momčilo Šaran[1], Ljuba Štrbac[1], Dobrila Janković[1], Mihajla Djan[2], Snežana Trivunović[1], Minja Zorc[3]*

[1]Department of Animal Science, Faculty of Agriculture, University of Novi Sad, Novi Sad, Serbia
[2]Department of Biology and Ecology, Faculty of Sciences, University of Novi Sad, Novi Sad, Serbia
[3]Department of Animal Science, Biotechnical Faculty, University of Ljubljana, Ljubljana, Slovenia

**Corresponding author: momcilo.saran@stocarstvo.edu.rs*

Citation: Šaran M., Štrbac L., Janković D., Djan M., Trivunović S., Zorc M. (2023): Genomic characterization of Serbian Holstein-Friesian cattle population. Czech J. Anim. Sci., 68: 486–496.

Abstract: The use of genomic data makes it possible to examine genetic variability and calculate the genetic parameters of the population in an efficient and precise way. The aim of this research was to analyse linkage disequilibrium (LD), contemporary effective population size (Ne_c), haplotype block structure, minor allele frequency (MAF), observed (H_O) and expected heterozygosity (H_E), calculate the genomic relationship matrix and perform a principal component analysis (PCA) in the Serbian Holstein-Friesian cattle population using SNP data from the GGP Bovine 100K chip. After quality control (QC), 83 208 SNPs and 1 575 cows were retained for further analysis. LD on autosomes had an average value of ≥ 0.2 up to a distance of 50–60 kb (r^2 = 0.211), while on BTX r^2 ≥ 0.2 was represented at distances of 80–90 kb (r^2 = 0.211). LD differed between chromosomes. The average H_O for autosomes and X chromosome SNPs was 0.412 and 0.422, respectively. 81.30% of SNPs that passed QC had MAF > 0.2. The total number of haplotype blocks in the studied population was 15 642. On average, blocks contained 2.932 SNPs. The average block length was 32.657 kb and ranged from a minimum of 0.019 kb (BTA21 and BTA26) to a maximum of 999.562 kb (BTX). The estimated value of Ne_c in the this cattle population was 142. The results of PCA showed a significant variability of genotypes in the population, but there was no clear stratification of the population. The obtained results will serve as a basis for future genomic analyses such as the detection of QTLs for important economic traits and the implementation of genomic selection.

Keywords: single nucleotide polymorphism; genetic diversity; genomic analysis; linkage disequilibrium; haplotype block

The data obtained from SNP microarray chips are widely used in the analysis of population diversity and genetic improvement of domestic animals. Compared to conventional breeding methods, the use of genomic information in animal breeding enables to achieve higher selection accuracy and higher rates of genetic progress (Shin et al. 2017).

Genomic selection and Genome-Wide Association Studies (GWAS) are based on the assumption that there is linkage disequilibrium (LD) between SNPs and quantitative trait loci (QTL), which allows the use of SNPs for the detection of QTLs in the genome. Linkage disequilibrium is usually defined as a non-random association of al-

Supported by the Ministry of Science, Technological Development and Innovations, Republic of Serbia (Grant No. 451-03-47/2023-01/200117).

https://doi.org/10.17221/89/2023-CJAS

leles from different loci. One of the most important factors that affect the manifestation of LD is the distance between markers. In addition to distance, LD is also affected by genetic drift, selection, changes in population size and structure, and by the occurrence of mutations and recombination (Qanbari 2020). Based on LD data, it is possible to determine the effective population size (*Ne*) in previous generations (Hayes et al. 2003). *Ne* represents the size of a hypothetical ideal population that will generate the same amount of genetic drift as the population under analysis (Allendorf et al. 2022). Small values of *Ne* in a population indicate that the population has experienced greater genetic drift compared to populations with higher *Ne* (Qanbari 2020). *Ne* represents an important parameter based on which one can model breeding programs, manage the conservation of endangered populations, and perform studies of evolutionary changes at the molecular level (Charlesworth 2009).

Haplotype blocks are regions on chromosomes with low haplotype diversity and a high level of LD between SNPs, which are separated from each other by recombination hotspots (Gabriel et al. 2002). Long haplotype blocks may arise as a result of positive selection pressure on a causative variant increasing the LD between that variant and surrounding markers (Alhaddad et al. 2013). Therefore, it is important to examine what genes are found in such haplotype blocks.

Holstein-Friesian (HF) is the most numerous and important cattle breed in the Autonomous Province of Vojvodina (APV), which represents the northern part of Serbia. The original population of Holstein-Friesian cattle in Serbia was formed by importing cows and heifers from various European countries (Netherlands, Germany, Denmark, Belgium), and then from the USA, as well as by importing the semen of progeny-tested bulls (Djedovic et al. 2023). HF cattle were also crossed with locally present breeds in order to improve their productivity. Breeding by crossing with HF bulls has been done over generations, so the percentage of HF genes in crossbred cattle has increased over time, and therefore they can now be considered HF. Nowadays, cows of the HF breed are mainly inseminated with the semen of bulls from the countries of North America and Western Europe. The accuracy of prediction using genomic selection is highly dependent on the size and genetic structure of the reference population and the relatedness between animals from the reference population and candidate animals (Calus 2010). Therefore, before applying genomic selection, taking into account the origin of the Serbian HF population, it is important to examine the degree of genetic relationship and parameters of genetic diversity in the population. The use of a genomic relationship matrix (GRM), calculated based on SNP markers, to estimate breeding values, compared to a matrix calculated based on pedigree information, allows for a more accurate estimation of the relationship between animals because GRM takes into account Mendelian sampling, which leads to deviations between actual and expected relationships (Veerkamp et al. 2011). One of the most significant reasons for the occurrence of false positive results in GWAS is the unidentified stratification of the population, which arises because subpopulations differ from each other in allele frequency, due to different genetic ancestry. One of the methods to overcome this problem is to calculate GRM, principal component analysis (PCA), and include them in the statistical model.

The aim of this research was to analyse LD, contemporary *Ne*, observed (*Ho*) and expected heterozygosity (H_E), minor allele frequency (MAF), the structure of haplotype blocks, calculate GRM and perform PCA in the Serbian HF population using GGP Bovine 100K chip.

MATERIAL AND METHODS

Ethical approval

The procedure with animals used in this research was approved by the Ministry of Agriculture, Forestry and Water Management of the Republic of Serbia – Veterinary Directorate (protocol code 323-07-10977/2019-05 from 19.11.2019).

Sample collection and genotyping

Samples of 1 600 cows of the Holstein-Friesian (HF) breed were collected from eight farms from different parts of the APV during the year 2022. Cows were selected at random and they all were registered in the Herd Book. Sample collection was performed by plucking approximately 40 hairs with follicles from the tail and placing them in hair cards. Genotyping was performed using the GGP Bovine 100K chip

https://doi.org/10.17221/89/2023-CJAS

(Neogen Inc., Lincoln, NE, USA) at the Neogen Europe Ltd laboratory (Ayr, Scotland, UK). The positions in the genome of SNPs from the chip were mapped according to the ARS-UCD1.2 assembly.

Data quality control

Quality control (QC) of data obtained after genotyping was performed using Plink v1.9 software (Chang et al. 2015) and the R programming language (R Core Team 2022). Pairs of animals that had an identity by descent (IBD) relationship ratio greater than 90% and animals that had an SNP call rate below 90% were excluded from further analyses. SNPs were excluded based on the following criteria: (1) SNPs that were located at the same positions; (2) markers which represented insertions and deletions; (3) that were not located on autosomes or the X chromosome; (4) that had a minor allele frequency (MAF) < 0.05; (5) with deviation from Hardy-Weinberg equilibrium with *P*-values < 0.000 001 and (6) SNP markers that had a call rate of less than 90%. After QC, the imputation of missing SNPs was performed using Beagle v5.4 software (Browning et al. 2018). After imputation, additional control was performed, where SNPs with MAF < 0.05 were excluded. The distribution of SNPs on chromosomes was graphically displayed using the *CMplot* R package (Yin et al. 2021)

Linkage disequilibrium

Linkage disequilibrium (LD) was calculated using the r^2 method, which represents the squared correlation between alleles at different loci (Hill and Robertson 1968), using the following formula:

$$r^2 = \frac{[f(AB) \times f(ab) - f(Ab) \times f(aB)]^2}{f(A) \times f(a) \times f(B) \times f(b)} \quad (1)$$

where:

$f(AB), f(ab), f(Ab), f(aB)$ – genotype frequencies;
$f(A), f(a), f(B), f(b)$ – allele frequencies.

LD values between SNPs were calculated with PLINK v1.90 software (Chang et al. 2015), using the command: *--ld-window-kb 15000 --ld-window 99999 --ld-window-r2 0*. For LD analysis, all SNPs from autosomes and X chromosome that passed QC were pooled into two separate groups, and calculations were performed only for syntenic SNP pairs. LD was calculated for pairs of SNPs that were 0 to 15 Mb apart. SNP pairs, depending on their mutual distance, were categorized into bins of increasing length of 1 kb, starting from 0 to 15 Mb, after which, using the R programming language (R Core Team 2022), the average r^2 value was calculated for each bin. Using the *ggplot2* R package (Wickham 2016), graphs were made showing LD decay, i.e. LD value as a function of physical distance between SNPs at distances of 0–15 Mb, while the 0–0.5 Mb region was zoomed in using the *ggforce* R package (Pedersen 2022). Each point in these graphs represents the average value of r^2 for a bin of increasing length of 1 kb. Summary statistics were calculated for r^2 values at the following distances: 0–10 kb, 10–20 kb, 20–30 kb, 30–40 kb, 40–50 kb, 50–60 kb, 60–70 kb, 70–80 kb, 80–90 kb, 90–100 kb, 100–500 kb, 0.5–1 Mb, 1–5 Mb, 5–10 Mb, 10–15 Mb and 0–15 Mb. Average LD values and standard deviations (SD) were also calculated for each chromosome separately for SNP pairs, at the following distances: 0–10 kb, 10–25 kb, 25–50 kb, 50–100 kb, 100–500 kb, 0.5–1 Mb, 1–5 Mb, 5–10 Mb, 10–15 Mb, and 0–15 Mb. All intervals in this study were left-closed, while the last interval was closed. All values after the ± sign represented the SD.

Contemporary effective population size estimation

Estimation of contemporary effective population size (Ne_c) was performed using GONE software (Santiago et al. 2020), which uses the algorithm (Mitchell 1998) and LD structure between SNPs at different genetic distances for calculation. The default parameters of the software were used for the calculation. The estimate for the most recent generation, which was for one generation ago, was taken as the estimate for Ne_c.

Genetic diversity parameters

In this study, MAF, observed (H_O) and expected heterozygosity (H_E) were calculated for each SNP using the Plink v1.9 software (Chang et al. 2015). For each of these values, the average and SD per

https://doi.org/10.17221/89/2023-CJAS

chromosome were calculated. The GRM was calculated using the GAPIT R Software package (v3) (Wang and Zhang 2021), according to the formula developed by VanRaden (2008):

$$GRM = \frac{\mathbf{ZZ}'}{2\sum p_i(1-p_i)} \quad (2)$$

where:

GRM – genomic relationship matrix;

Z – the genotype matrix which is centered using the observed allelic frequencies for each locus (the **Z** matrix has dimensions n x m, where n is the number of animals and m is the number of SNP markers);

p_i – represents the frequency of the second allele at locus i (VanRaden, 2008).

Principal component analysis (PCA) was performed with the same software based on the previously calculated GRM. PCA was used to examine the genetic relatedness between cows from different farms. The results of the PCA analysis were plotted using the *ggplot2* R package (Wickham 2016), with cows marked with different colours depending on the farm they came from.

Haplotype block structure

The structure of haplotype blocks in the population was analysed with Plink v1.9 software (Chang et al. 2015), using the haplotype block definition proposed by Gabriel et al. (2002). The following command was used for the analysis: *--blocks no-pheno-req --blocks-max-kb 1000*, while default settings were used for the analysis of other parameters. The maximum length of the blocks analysed was limited to 1 Mb. Then, summary statistics of the results were calculated using the R programming language (R Core Team 2022) for each chromosome separately and for the entire group of results. In this research, it was assumed that the longest haplotypes are regions that were subjected to selection pressure during the development of the breed, therefore, for the top 0.5 percentile (> 307.959 kb) of the longest haplotype blocks, we examined which genes they overlap with, using the GALLO R package (Fonseca et al. 2020). Gene position data were based on the ARS-UCD1.2 assembly and were downloaded from the Ensembl database: (http://www.ensembl.org/index.html).

RESULTS

Data quality control and SNP marker statistics

From the initial number of genotyped SNPs (95 256), 85 996 SNPs of them were retained after QC, which is 90.28% of the initial number. Of these, 82 851 SNPs were on autosomes, and 3 145 were on the X chromosome. After QC, 1 587 animals were retained for further analysis. The average distance between adjacent SNPs on all *Bos taurus* autosomal (BTA) chromosomes was 0.03 Mb with a SD of 0.029 Mb, while the average distance between adjacent SNPs on the *Bos taurus* X chromosome (BTX) was 0.044 Mb with a SD of ± 0.081 Mb. The examined SNPs covered about 2.62 Gb of the bovine genome. The largest distance between two adjacent SNPs was on BTX (2.205 Mb), and the smallest on BTA1 (0.004 kb). The longest chromosome was BTA1 (157.873 Mb), and the shortest was BTA25 (42.24 Mb). The largest number of SNP markers (n = 5 165) was located on BTA1, and the smallest was on BTA25 (n = 1 583). Complete SNP statistics are shown in Table S1. The distribution of SNPs on chromosomes divided into 1 Mb bins is shown in Figure 1.

Linkage disequilibrium

In this research, LD was examined between all combinations of SNP pairs that were at a distance of 0 to 15 Mb, whereby LD was calculated for a total of 37 373 064 syntenic SNP pairs from autosomal chromosomes and 1 035 189 pairs from the X chromosome. The average LD value for all examined autosomal SNP pairs at distances of 0–15 Mb was 0.034 ± 0.065, while on the X chromosome, it was 0.033 ± 0.07. A total of 5 611 autosomal and 506 BTX SNP pairs had an LD value of r^2 = 1, while 421 BTA and 6 BTX pairs had r^2 = 0. Average r^2 values for 1 500 bins for increasing 1 kb intervals are shown in Table S2. The decrease in LD with increasing distance between pairs of SNPs is shown in Figure 2, where LD values at distances of 0–15 Mb and a zoomed section at distances of 0–0.5 Mb are shown. The figure shows that at shorter distances, r^2 between pairs on BTX was higher compared to BTA, as well as that r^2 from BTX had a greater variability compared to r^2 from BTA. By comparing the first 10 bins of size 10 kb, it was determined that the highest average r^2 value

https://doi.org/10.17221/89/2023-CJAS

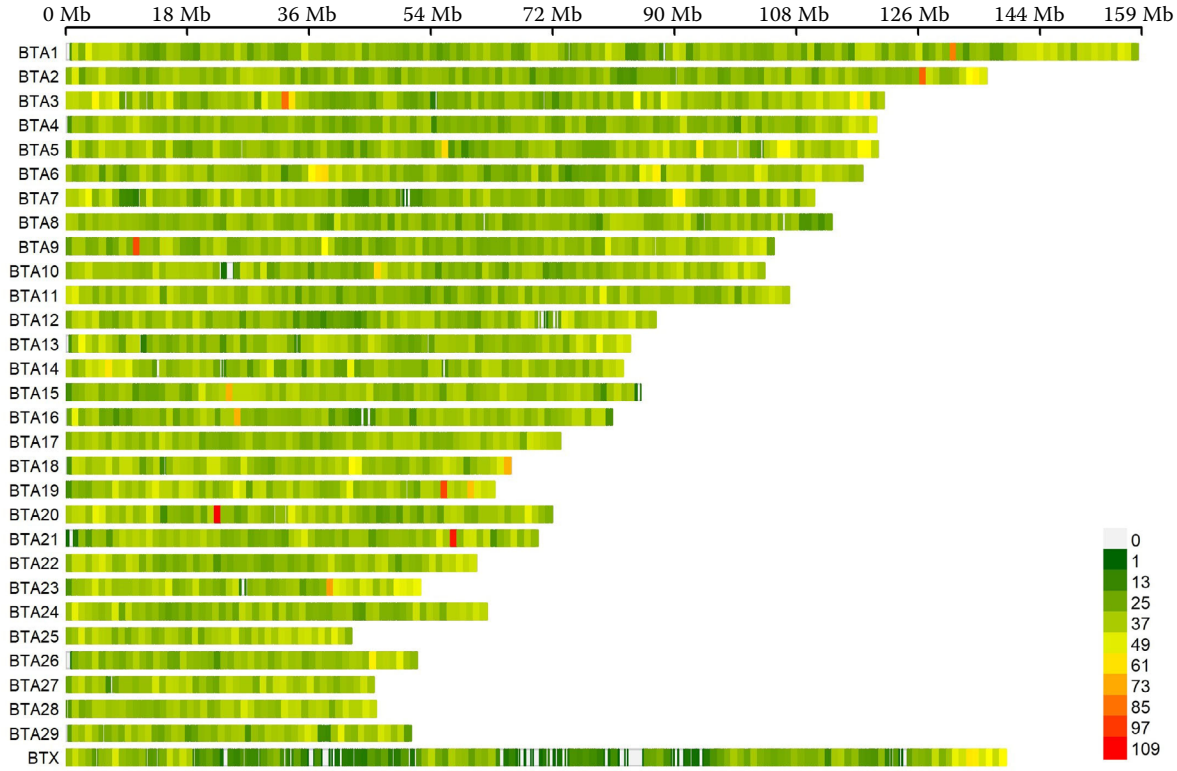

Figure 1. Distribution of SNPs on bovine chromosomes, divided into 1 Mb bins
Different bin colours indicate the number of SNPs, as shown in the legend. The names of chromosomes are indicated on the vertical axis, while the horizontal axis represents the length of chromosomes in Mb

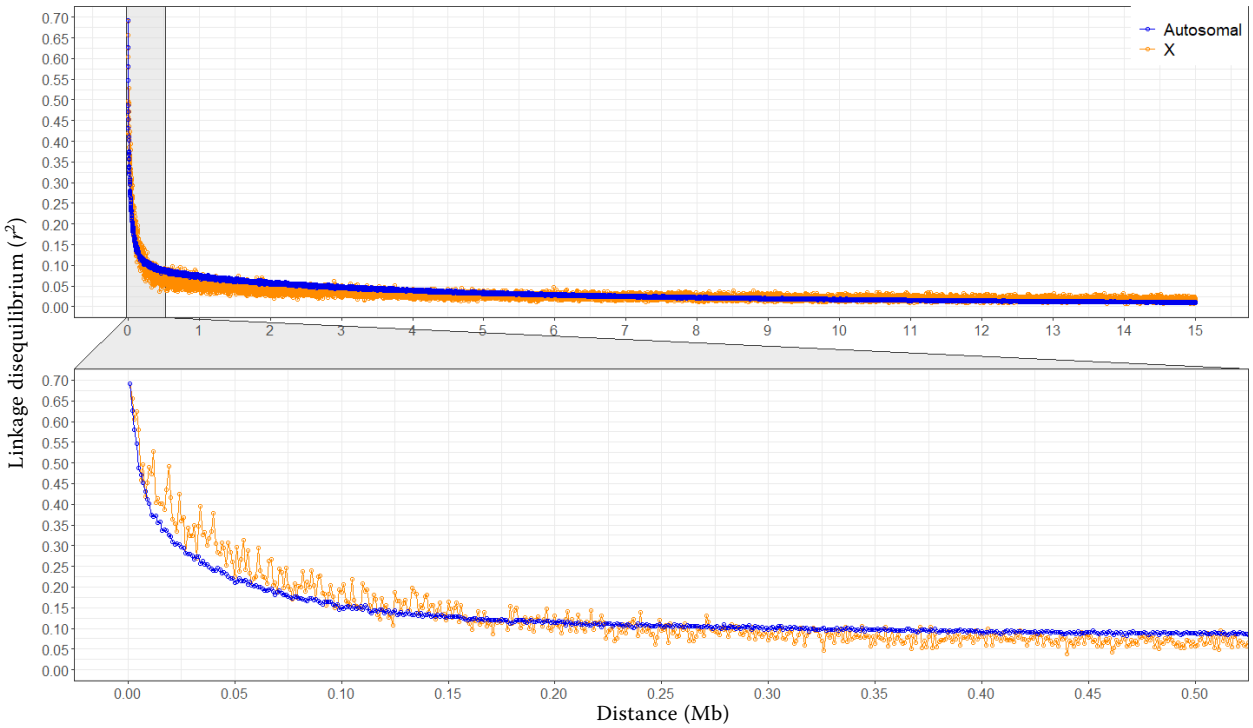

Figure 2. Linkage disequilibrium (LD) decay for distances between SNPs from 0 to 15 Mb (the upper part of the figure). Linkage disequilibrium (LD) decay for distances between SNPs from 0 to 0.5 Mb (the lower part of the figure)
The blue line represents the average r^2 between SNPs on autosomes while the orange one represents r^2 between SNPs on X chromosome

https://doi.org/10.17221/89/2023-CJAS

was in the first bin (0–10 kb) in the amount of 0.503 ± 0.367 (BTA) and 0.519 ± 0.354 (BTX) (Table 1). The highest proportion of SNP pairs that had a useful level of LD ($r^2 \geq 0.2$) was also located in the first bin 0–10 kb, in the amount of 66.99% (BTA) and 73.42% (BTX). LD on autosomes had an average value ≥ 0.2 up to a distance of 50–60 kb (r^2 = 0.211 ± 0.239), while on BTX $r^2 \geq 0.2$ was represented up to a distance of 80–90 kb (r^2 = 0.211 ± 0.256). Average LD values for the same distances differed between chromosomes and are shown in Table S3. The highest average r^2 between SNP pairs in the interval of 0–15 Mb was on chromosome BTA20 (r^2 = 0.047 ± 0.085), and the lowest on chromosome BTA27 (r^2 = 0.022 ± 0.047).

Genetic diversity parameters and effective population size estimation

The average MAF value for SNPs that passed QC from autosomal chromosomes was 0.326 ± 0.121 and, depending on the chromosome, it ranged from 0.305 on BTA20 to 0.338 on BTA28, while on the BTX chromosome, it was 0.333 ± 0.113. The distribution of MAF values according to different intervals is shown in Figure 3. The largest number of SNPs was in the MAF interval of 0.45–0.50 (n = 15 767; 18.33%), and the smallest in the interval of 0.05–0.10 (n = 3 714; 4.32%). The percentage of SNPs that had a MAF value greater than 0.2 was 81.30%.

The average H_O of loci from autosomes was 0.412 ± 0.102, while for loci from the X chromosome, it was 0.422 ± 0.094. The highest average H_O was on BTA28 (0.422), and the lowest was on BTA20 (0.394) (Table S1). The average H_E of loci from autosomes was 0.410 ± 0.101, while for loci from the X chromosome, it was 0.419 ± 0.092. Depending on the chromosome, H_E ranged from 0.392 (BTA20) to 0.421 (BTA28). Ne_c value of the investigated population of HF cattle was 142.

The principal component analysis showed that cows from different farms formed overlapping groups, indicating that there is no clear stratification of the population into subpopulations (Figure 4). An exception is a certain number of cows from farm 4, but they were not significantly far from the rest of the main group either. The first two principal components (PC) explained the largest percentage of the total genetic variability in the amount of 1.51% (PC1) and 1.28% (PC2). The average value of the diagonal elements of the GRM was 0.995 ± 0.031, while the minimum and maximum values were 0.903 and 1.204, respectively. The average value of off-diagonal GRM elements was –0.001 ± 0.041, while the minimum and maximum values were –0.106 and 0.734, respectively.

Table 1. Summary statistics for r^2 values at different distances between SNP pairs on autosomal and X chromosomes

Distance (Mb)	Autosomal chromosomes			X chromosome		
	number of SNP pairs	r^2 (average) ± SD	$r^2 \geq 0.2$ (%)	number of SNP pairs	r^2 (average) ± SD	$r^2 \geq 0.2$ (%)
0–0.01	24 032	0.503 ± 0.367	67.0	790	0.519 ± 0.354	73.4
0.01–0.02	26 744	0.348 ± 0.317	56.0	978	0.436 ± 0.345	63.8
0.02–0.03	32 757	0.293 ± 0.290	48.6	1 006	0.348 ± 0.329	53.1
0.03–0.04	30 691	0.258 ± 0.273	43.1	1 015	0.335 ± 0.314	55.3
0.04–0.05	29 843	0.230 ± 0.255	39.7	999	0.281 ± 0.303	44.0
0.05–0.06	29 653	0.211 ± 0.239	36.3	977	0.254 ± 0.276	41.6
0.06–0.07	29 431	0.192 ± 0.225	33.4	961	0.234 ± 0.283	35.6
0.07–0.08	29 540	0.179 ± 0.213	31.5	998	0.205 ± 0.251	34.6
0.08–0.09	29 489	0.169 ± 0.204	29.5	1 003	0.211 ± 0.256	34.5
0.09–0.10	29 292	0.158 ± 0.194	27.5	969	0.176 ± 0.219	29.4
0.10–0.50	1 135 001	0.106 ± 0.136	17.0	37 401	0.097 ± 0.158	13.5
0.50–1.00	1 394 024	0.080 ± 0.102	11.2	44 391	0.061 ± 0.102	6.20
1.00–5.00	10 673 102	0.049 ± 0.068	4.29	322 596	0.036 ± 0.058	2.05
5.00–10.00	12 394 089	0.024 ± 0.036	0.580	332 130	0.023 ± 0.035	0.490
10.00–15.00	11 485 376	0.014 ± 0.021	0.048	288 975	0.018 ± 0.028	0.218
0–15.00	37 373 064	0.034 ± 0.065	2.69	1 035 189	0.033 ± 0.070	2.04

https://doi.org/10.17221/89/2023-CJAS

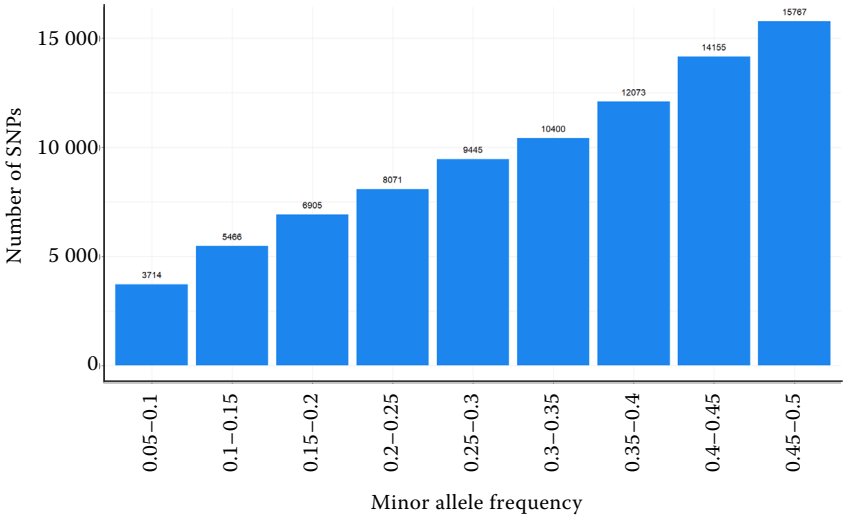

Figure 3. Distribution of minor allele frequency (MAF) for all SNPs that passed quality control

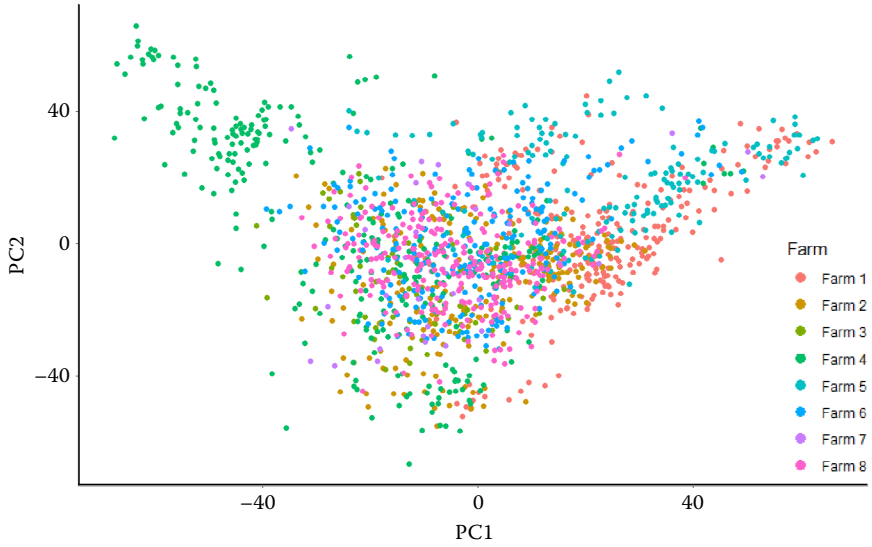

Figure 4. Principal component decomposition of the genomic relationship matrix in the Serbian Holstein-Friesian cattle population, where cows are coloured according to the farm they came from

Haplotype block structure

Descriptive statistics of the haplotype block analysis of the studied HF population are shown in Table 2. The total number of haplotype blocks in the studied population was 15 642, which covered 510 827 kb (19.49%) of the genome. The highest chromosome coverage with haplotype blocks was found on BTX (35 084 kb; 25.24%), and the lowest on BTA27 (6 374 kb; 14.02%). The average block length was 32.657 kb and it ranged from a minimum of 0.019 kb (BTA21 and BTA26) to a maximum of 999.562 kb (BTX). The largest number of blocks (n = 971) was located on BTA1 and the smallest (n = 293) on BTA25. The total number of SNPs found in haplotype blocks was 45 870 (53.34% of the total number of SNPs). On average, blocks contained 2.932 SNPs. The minimum number of SNPs in the blocks was 2 on all chromosomes, while the maximum number of SNPs was located in the haplotype block on BTA20 (36 SNPs).

In Figure 5, it can be seen that the largest number of haplotype blocks belonged to the 0–10 kb length interval and that there is a noticeable trend of decreasing frequency with increasing haplotype block length.

By examining the 79 longest haplotype blocks (top 0.5 percentile), they were found to overlap with 515 different genes. A total of 71 blocks contained one or more genes, while eight did not contain any genes. The largest number of genes was located on the haplotype block on BTA7 with coordinates 43 190 739–43 999 567 bp (53 genes), while 17 haplotype blocks from different chromosomes contained only one gene. Depending on the function, the identified genes belonged to the

https://doi.org/10.17221/89/2023-CJAS

Table 2. Haplotype block statistics of the Serbian Holstein-Friesian cattle population

CHR	Blocks (*n*)	Total block length (kb)	Block size (kb)			SNPs in blocks (*n*)	% of SNPs in blocks	SNPs in blocks (*n*)		
			min.	mean	max.			min.	mean	max.
BTA1	971	32 788	0.513	33.8	420	2 909	56.3	2	3.00	30
BTA2	779	26 382	0.548	33.9	535	2 314	53.2	2	2.97	16
BTA3	758	25 734	0.088	33.9	389	2 290	55.0	2	3.02	24
BTA4	676	21 702	0.027	32.1	292	1 925	51.6	2	2.85	10
BTA5	738	26 082	0.068	35.3	397	2 266	55.7	2	3.07	19
BTA6	732	24 314	0.470	33.2	463	2 198	55.6	2	3.00	22
BTA7	637	23 710	0.164	37.2	809	1 936	55.5	2	3.04	18
BTA8	631	21 356	0.061	33.8	344	1 826	52.2	2	2.89	23
BTA9	621	18 803	0.071	30.3	313	1 807	52.3	2	2.91	24
BTA10	606	18 957	0.023	31.3	337	1 753	51.9	2	2.89	15
BTA11	633	19 344	0.464	30.6	362	1 813	51.7	2	2.86	14
BTA12	498	16 486	0.465	33.1	888	1 417	50.5	2	2.85	20
BTA13	524	18 245	0.125	34.8	443	1 597	57.0	2	3.05	13
BTA14	507	19 547	0.293	38.6	354	1 556	56.0	2	3.07	16
BTA15	508	15 284	0.121	30.1	382	1 466	51.3	2	2.89	14
BTA16	445	15 961	0.299	35.9	998	1 342	52.6	2	3.02	22
BTA17	459	12 839	0.039	28.0	319	1 291	52.0	2	2.81	10
BTA18	452	12 930	0.267	28.6	337	1 301	55.1	2	2.88	10
BTA19	439	11 874	0.401	27.0	340	1 304	54.6	2	2.97	22
BTA20	451	14 531	0.509	32.2	463	1 384	55.9	2	3.07	36
BTA21	424	12 288	0.019	29.0	995	1 239	52.9	2	2.92	24
BTA22	367	10 708	0.247	29.2	451	1 015	50.0	2	2.77	12
BTA23	327	7 562	0.145	23.1	500	923	47.8	2	2.82	18
BTA24	386	11 997	0.493	31.1	459	1 105	53.3	2	2.86	14
BTA25	293	6 538	0.215	22.3	228	798	50.4	2	2.72	13
BTA26	294	7 888	0.019	26.8	440	828	48.9	2	2.82	11
BTA27	300	6 374	0.400	21.2	224	789	49.2	2	2.63	9
BTA28	297	6 882	0.475	23.2	202	788	48.6	2	2.65	6
BTA29	325	8 636	0.052	26.6	698	863	50.3	2	2.66	9
BTX	564	35 084	0.608	62.2	1 000	1 827	58.1	2	3.24	20

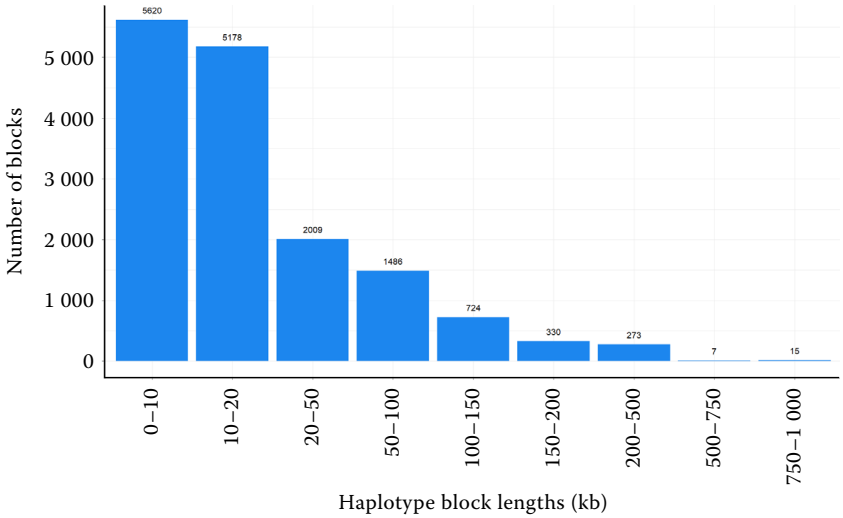

Figure 5. Distribution of haplotype blocks based on their length

https://doi.org/10.17221/89/2023-CJAS

following categories: protein-coding (80.97%), small nucleolar RNA (7.57%), small nuclear RNA (3.50%), long non-coding RNA (3.30%), microRNA (2.33%), pseudogenes (0.97%), ribosomal RNA (0.58%), miscellaneous RNA (0.39%) and processed pseudogenes (0.39%). A list of identified genes that overlapped with the longest haplotype blocks is presented in Table S4.

DISCUSSION

The use of genomic data has enabled new and more accurate ways to characterize the diversity of domestic animal populations. In this research, analyses were performed to determine the genetic diversity within the Serbian Holstein-Friesian population using the GGP Bovine 100K chip. Analysis of the structure of linkage disequilibrium in the population is a prerequisite for the successful application of GWAS and genomic selection. In our research, useful LD values ($r^2 \geq 0.2$) on autosomes were present at distances smaller than 60 kb. The average distance between SNPs on autosomes was ~30 kb, which means that there is a satisfactory level of LD between adjacent SNPs. Average r^2 values in the German HF population at distances < 25 kb were $r^2 = 0.30 \pm 0.32$ (Qanbari et al. 2010), which is almost identical to r^2 at distances < 25 kb in the examined Serbian HF population. In the North American Holstein population, Sargolzaei et al. (2008) found an average $r^2 = 0.58$ at distances up to 0.1 Mb, which was higher than in our population, comparing the same distance. Interchromosomal LD heterogeneity was observed in the studied population. The highest average r^2 was found at BTA20 (0.047), and the lowest at BTA27 (0.022). Interchromosomal LD heterogeneity has also been observed in the North American HF population (Sargolzaei et al. 2008). According to Sargolzaei et al. (2008), the potential cause for this phenomenon is selection, to what the Holstein breed was intensively subjected in the past.

The average MAF value on autosomes in this population (0.326 ± 0.121) was higher than the values found in the German Holstein population (0.28 ± 0.15) (Qanbari et al. 2010). High MAF values may indicate that the population has high genetic diversity. PCA results show that there is a significant genotype variability in the population, but that it is not possible to identify clearly separated subpopulations. The average H_O and H_E on the autosomes were high with values of 0.412 and 0.410, respectively, which was higher than the values found in the Irish HF cattle population ($H_O = 0.370$; $H_E = 0.372$) (Kelleher et al. 2017).

Ne is a valuable indicator of the viability of the population and it is important to monitor it in order to prevent negative trends and the extinction of the population. In order to prevent extinction caused by inbreeding depression, the effective size of populations must be at least 100 individuals (Frankham et al. 2014). Based on this value, it can be concluded that the studied population is not critically endangered, because the Ne_c value was 142. The estimated value of Ne_c was higher than the estimated *Ne* of the most recent generation in the German (Qanbari et al. 2010) and North American (Sargolzaei et al. 2008) Holstein cattle populations. The low *Ne* values found in different Holstein populations in recent generations were caused by the use of a small number of bulls and their intensive selection (Sargolzaei et al. 2008), resulting in an increased rate of inbreeding.

For certain genes that were found on the longest haplotype blocks, which were potentially created due to selection pressure, in earlier associative studies in cattle, it was determined that they affect the expression of various economically important traits. Among others, Teng et al. (2023) found that the *GHR* gene from BTA20 and the genes *SLC39A4*, *CPSF1*, *ADCK5*, *SLC52A2*, *SCRT1*, *FBXL6*, *TMEM249*, *DGAT1*, *ENSBTAG00000053637* from the haplotype block on BTA14 were significantly associated with milk production traits.

CONCLUSION

In this research, the genetic diversity and main genetic parameters within the Serbian Holstein-Friesian cattle population were examined. Principal component analysis, MAF, and heterozygosity analyses revealed significant genetic diversity within the population, which reflects the history of the formation of this population. The same conclusion about variability can be reached based on LD and *Ne* data. The LD results show that the used chip has a sufficient number of SNPs for genomic analyses, considering that useful LD ($r^2 \geq 0.2$) on autosomes extended up to 60 kb on average. The es-

https://doi.org/10.17221/89/2023-CJAS

timated value of the contemporary effective population size was at a satisfactory level and had a value of Ne_c = 142. A significant number of haplotype blocks were found, where some of the longest ones contained genes with previously proven influence on the expression of economically important traits. The data obtained in this study will support the application of genomic selection and GWAS in the Serbian HF population, but also the conservation of genetic diversity and the development of breeding programs. Since this was the first analysis of this kind in the Serbian HF population, it is necessary to continue the research with more markers and animals.

Acknowledgement

The SNP data used in this research were collected during the realization of PROMIS project: "BioITGenoSelect", No. 6066512, funded by the Science Fund of the Republic of Serbia.

Conflict of interest

The authors declare no conflict of interest

REFERENCES

Alhaddad H, Khan R, Grahn RA, Gandolfi B, Mullikin JC, Cole SA, Gruffydd-Jones TJ, Haggstrom J, Lohi H, Longeri M, Lyons LA. Extent of linkage disequilibrium in the domestic cat, Felis silvestris catus, and its breeds. PLoS One. 2013 Jan 7;8(1): 11 p.

Allendorf FW, Funk WC, Aitken SN, Byrne M, Luikart G. Effective population size. In: Allendorf FW, Luikart G, Aitken S, editors. Conservation and the genomics of populations. Oxford: Oxford University Press; 2022. p. 133-50.

Browning BL, Zhou Y, Browning SR. A one-penny imputed genome from next-generation reference panels. Am J Hum Genet. 2018 Sep;103(3):338-48.

Calus MPL. Genomic breeding value prediction: Methods and procedures. Animal. 2010;4(2):157-64.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: Rising to the challenge of larger and richer datasets. GigaSci. 2015 Dec;4(1): 16 p.

Charlesworth B. Effective population size and patterns of molecular evolution and variation. Nat Rev Genet. 2009 Mar;10(3):195-205.

Djedovic R, Vukasinovic N, Stanojevic D, Bogdanovic V, Ismael H, Jankovic D, Gligovic N, Brka M, Strbac L. Genetic parameters for functional longevity, type traits, and production in the Serbian Holstein. Animals. 2023 Feb 2; 13(3): 21 p.

Fonseca PAS, Suarez-Vega A, Marras G, Canovas A. GALLO: An R package for genomic annotation and integration of multiple data sources in livestock for positional candidate loci. GigaScience. 2020 Dec 30;9(12): 9 p.

Frankham R, Bradshaw CJA, Brook BW. Genetics in conservation management: Revised recommendations for the 50/500 rules, Red List criteria and population viability analyses. Biol Conserv. 2014 Feb;170:56-63.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. The structure of haplotype blocks in the human genome. Science. 2002 Jun 21;296(5576):2225-9.

Hayes BJ, Visscher PM, McPartlan HC, Goddard ME. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Res. 2003 Apr 1; 13(4):635-43.

Hill WG, Robertson A. Linkage disequilibrium in finite populations. Theoret Appl Genetics. 1968 Jun;38(6):226-31.

Kelleher MM, Berry DP, Kearney JF, McParland S, Buckley F, Purfield DC. Inference of population structure of purebred dairy and beef cattle using high-density genotype data. Animal. 2017;11(1):15-23.

Mitchell M. An introduction to genetic algorithms. Cambridge: The MIT press; 1998. 221 p.

Pedersen TL. ggforce [Internet]. 2022 [cited 2023 Jun 21]. Available from: https://github.com/thomasp85/ggforce

Qanbari S. On the extent of linkage disequilibrium in the genome of farm animals. Front Genet. 2020 Jan 17;10: 11 p.

Qanbari S, Pimentel ECG, Tetens J, Thaller G, Lichtner P, Sharifi AR, Simianer H. The pattern of linkage disequilibrium in German Holstein cattle. Anim Genet. 2010 Aug;41(4):346-56.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2022. Available from: https://www.R-project.org/

Santiago E, Novo I, Pardinas AF, Saura M, Wang J, Caballero A. Recent demographic history inferred by high-resolution analysis of linkage disequilibrium. Mol Biol Evol. 2020 Dec 16;37(12):3642-53.

Sargolzaei M, Schenkel FS, Jansen GB, Schaeffer LR. Extent of linkage disequilibrium in Holstein Cattle in North America. J Dairy Sci. 2008 May;91(5):2106-17.

Shin D, Lee C, Park KD, Kim H, Cho KH. Genome-association analysis of Korean Holstein milk traits using

https://doi.org/10.17221/89/2023-CJAS

genomic estimated breeding value. Asian-Australas J Anim Sci. 2017;30(3):309-19.

Teng J, Wang D, Zhao C, Zhang X, Chen Z, Liu J, Sun D, Tang H, Wang W, Li J, Mei C, Yang Z, Ning C, Zhang Q. Longitudinal genome-wide association studies of milk production traits in Holstein cattle using whole-genome sequence data imputed from medium-density chip data. J Dairy Sci. 2023 Apr;106(4):2535-50.

VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008 Nov;91(11):4414-23.

Veerkamp RF, Mulder HA, Thompson R, Calus MPL. Genomic and pedigree-based genetic parameters for scarcely recorded traits when some animals are genotyped. J Dairy Sci. 2011 Aug;94(8):4189-97.

Wang J, Zhang Z. GAPIT Version 3: Boosting power and accuracy for genomic association and prediction. Genom, Proteom Bioinform. 2021 Aug;19(4):629-40.

Wickham H. ggplot2 [Internet]. Cham: Springer International Publishing; 2016 [cited 2022 Dec 15]. Available from: http://link.springer.com/10.1007/978-3-319-24277-4

Yin L, Zhang H, Tang Z, Xu J, Yin D, Zhang Z, Yuan X, Zhu M, Zhao S, Li X, Liu X. rMVP: A memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. Genom Proteom Bioinform. 2021 Aug;19(4):619-28.

Received: June 27, 2023
Accepted: November 26, 2023
Published online: December 15, 2023