Development of the first Russian genomic reference database and genomic evaluation results

D. Iakovishina, K. Maximova, D. Nemirich, A. Dekin, M. Genaev, Y. Tavirikov, M. Sumtsova, and Y. Pekov

Ksitest (LLC Ksivalue), 127051, Russian Federation, Moscow, Leninsky prospect, 30A
Corresponding Author: daria.iakovishina@ksivalue.com

Abstract

At the moment, in the Russian Federation there is no established methodology for calculating the genomic breeding value of sires and dams. Routine selection (breeding) is carried out manually based on the dam's productivity only. But to perform effective breeding, it is important to choose the worst and best cows and rank animals with higher accuracy.

Genomic breeding values are widely used all around the world, but have not been implemented in Russia due to a lack of a local reference database. Foreign reference databases have advantages in size. However, the reliability of ranking animals according to foreign bases may be inaccurate due to different conditions of keeping animals, climate and other external factors.

In 2020, we launched a project to introduce genomic selection in the Udmurt Republic, which is the fourth largest milk producing region in Russia with many farms. The first genomic reference database in Russia was developed for the Holstein breed, the most popular dairy breed. This database contains more than 8000 animals with genotypes, including cows from frontrunning breeding farms in the region and bulls from the largest suppliers of semen in the region.

In the Russian Federation not all animals have unique IDs. So, at the first stage of our work we assigned each animal its own identifier. It is necessary to illustrate pedigree reliably. For this purpose phenotypic data was analyzed and filtered. We describe the most common mistakes in accumulated phenotypic data and how to process it.

We present our experience in the development of the first genome reference database with more than 8000 genotypes, which provides more accurate breeding values estimation and information about genetic diseases most represented in the Russian Federation.

The first genomic evaluations were calculated with the single-step genomic BLUP (ss-gBLUP) method for the following parameters: productivity, fat (absolute value and percentage), protein (absolute value and percentage) and longevity. The reliability of estimated genomic breeding values is up to 66%. We show the difference in evaluation between the pedigree-based breeding values and genomic breeding values.

We developed the web service for agricultural holdings which allows accurate breeding based on the GEBV (genomic estimated breeding values) and other results of genetic tests. Our solution allows breeders to find animals with the large and low genetic potential to increase the productivity of the next generation, select animals for crossing, and choose animals for embryo transfer, as well as to identify genetic abnormalities and economically useful traits in cows at an early stage.

Keywords: Genomic evaluation, reference database, russian holsteins, ss-gblup.

Introduction

The population of dairy cattle, represented mainly by the Holstein breed, is located in many regions of Russia, where climatic, feeding, and technological conditions vary significantly. Most of the cattle are concentrated in the Central, Volga, Southern, and North Caucasus districts. According to the Dairy Intelligence Agency, the number of dairy cattle in Russia in 2019 was 10.7 million heads with 4.6 million cows among them. Of these, the herd of the largest holding is 183 thousand head, with average productivity of 8 tons of milk per cow.

There are several types of organizations in Russian dairy cattle breeding and each has several distinctive features. There are large and medium-sized agricultural organizations, small farms and individual entrepreneurs. The main difference between them is in the dairy cattle population distribution. Thus large and medium-sized agro-holdings account for up to 69.5% of total livestock in Russia, while small farms account for 30.5%. In addition, differences are observed in farming conditions and contribution to overall dairy production. For instance, the difference in cow productivity can be observed between different organization types. The average productivity of dairy cattle in Russia was 4642 kg per cow per year of farms, while the productivity of agricultural holdings was 6290 kg, and for individual entrepreneurs 3791 kg per cow per year as of 2019, the most successful farms produce 13 000 kg per cow on average.

Agricultural organizations are an important player in the dairy cattle sector accounting for 85% of all milk production in Russia. The typical customer of the Ksitest company is an agricultural organization with a breeding status. The breeding status implies some obligations such as annual reports on livestock and productivity of animals and fixation of several indicators. Breeding farms are obliged to record inseminations, calvings, disposal reasons, evaluation of the exterior, and live weight at various stages of an animal's life. Also, they are required to conduct a control milking at least once a month, whereas they record the cow's milk yield, and send samples to the laboratory, determining the minimum fat, protein, somatic cells count, and other indicators. All these indicators are recorded in a local herd management program Selex (*https://plinor.spb.ru/index.php?l=0&p=3*). In contrast, industrial farms (non-breeding farms) are generally not restricted to anything and can only record as much or as little data as they find necessary.

The program was registered in 1997 and most of the breeders started to incorporate it in their data collection process in the last 20 years. Before that, they had preferred to use handwritten cards, which needed to be transferred to the electronic form. Some of the animal's cards go up to the 1970s.

Some farms can often use a second program for herd management: DairyComp, UniForm Agri, Afifarm, and others. Such programs can produce herd reports and reduce the human factor in filling data by using chips on animals, as well as automatic data collection from milking parlors, the results of control milkings, etc.

Data collected by breeders may have several problems. While many Russian breeders have started to introduce unique identifiers for animals, the data could still have some identification issues and, therefore, the pedigree errors. Besides, most of them use microsatellites as a kinship analysis, however, there is still data with parentage verification by immunogenetics. Therefore, the first step in reference database development is to correct all factual inaccuracies and use as much accumulated data as possible. Verification of animal records is especially important to get accurate estimated breeding values based on pedigree (EBV) and estimated breeding values based on genomic information (GEBV).

Russian dairy farms differ in environmental conditions, for instance, in climatic regime: Volga and Southern districts average July temperatures are 14,7°-24,7° and 18,0°-29,4° respectively. There are also differences in feed composition, care, management, and breeding programs. Therefore, cows from different Russian districts can not be comparable by their 305d productivity. Given this diversity, ranking animals by their genomic breeding values is the only way to compare dairy cattle in Russia by their genetic worth, and genomic selection seems particularly desirable.

The proper reference database plays a crucial role in the accurate breeding values estimation and is an essential step towards genomic selection implementation. The Holstein population in Russia is huge and spread throughout the country in different environmental and management conditions, phenotypic recordings of which have accumulated over several generations. However, local breeders still use the selection based on the dam's productivity, missing the opportunity of shorter generations and higher rates of genetic gain. The objectives of this work are to describe the steps of the first reference genomic database formation and to show the main results of the genomic evaluation of the Russian Holstein population as a beginning of genomic selection in Russia.

Stages of the genomic reference database creation

Data collection and filtration

As was mentioned before all breeders in Russia have duties on data recording, so therefore as a start we use dumps from herd management programs.

The phenotypic data gathered from these programs along with pedigree information are reported in table 1.

After uploading animal data to the Ksitest database, we perform pedigree verification. Verification helps solve problems such as (1) assigning multiple IDs to the same animal, or (2) having multiple records with the same inventory number for different animals. Problem (1) can arise, for example, when animals move between farms; problem (2) occurs due to the assignment of the same inventory number to two different animals from different farms. Errors associated with incorrect assignment of identifiers can distort

Table 1. Phenotypic data collected from Selex Dairy desktop.

Milk production	Milk, kg per day (starting from 5 days after calving) sum of milking (in case of 2 or 3 milking times) 305 days productivity and throughout lactation amount of milk fat, protein, somatic cells (laboratory checked data) daily milk yield in kg/time-consuming for milking during the day, min
Service period	The value of the service period in days
Longevity	The value of productive life in the number of lactations
Insemination	Date, bull, method
Calving	Date, result, calf weight, ease
Other information	Dry period (date, method), disposal (reason, date)

the breeding index and pedigree. To solve these problems, we assign each animal its own unique identifier. This approach allows tracking the movement of animals. For example, when the animal's birth farm and current location are different. To identify and correct data errors, we compare phenotypic data, pedigree and other information contained in records about the animal. The data is analyzed for inconsistencies in total for more than 10 indicators, including: data conflicts in dates of birth; conflicts in household data of birth; conflicts in nickname data; conflicts of data on the sex and age group of the animal; conflicts in the records of the breed of the animal. Phenotypic data derived from the Selex database is also checked for outliers. Besides, we include data that is directly confirmed by farmers as reliable.

Data genotyping

The process of biological data collection and subsequent genotyping began in the summer of 2019 and is still ongoing. There are several ways of getting genotypes:

1. customers collect biological material and we send it for genotyping at the DNA Laboratory;
2. customers provide genotypes previously obtained by them at other laboratories;
3. we obtain genotypes from international databases (CDCB, GenoEx, WWS), directly or through partners.

Depending on the way the data are accessed, the biomaterial used for genotyping differs. Hair follicles from the tail are the most used method because of the ease of obtaining and exporting. However, some customers prefer ear tissue sampling instead of follicles (<2% of samples). Blood sampling option is still under testing.

The selection of animals for genotyping is based on the goals and financial capabilities of farms. Generally, between 100 and 2,000 animals are collected from each farm. The main criteria are as follows: mother-daughter pairs or daughters with parents previously genotyped are collected; cows should not exceed 10 years of age; cows should have milk recordings for at least the first lactation; 305d lactation yield > 2,000 kg.

Genotyping is performed on a medium-density chip - the Weatherbys Scientific Bovine VersaSNP 50K. Usually it takes 1-1.5 months from the time the material is collected until the laboratory uploads the genotypes to the ftp-server. In addition to data from Bovine VersaSNP 50K, we also have genotypes from other chips (EuroG_MD, GGP_HD, ZMD, etc.) that differ from the a fore mentioned one in both density and content. An imputation procedure between chips is planned to add this data to the database in the nearest future.

The quality-check procedure is performed for collected genotypic data. Genotyped samples with call-rate < 0.9 and duplicates are removed. Duplicate identification is performed with PLINK v1.9 —genome option (Chang *et al.*, 2015). All animals meeting the following criteria Z0 <= 0.15 & Z1 <= 0.15 & Z2 >= 0.7 are considered to be duplicates and are being removed. The filtered genotypes are then used to search and verify the parentage, to identify statuses of farming traits and diseases, and calculate genomic breeding values.

Parentage verification

As the DNA Data Interpretation center accredited by ICAR (accreditation was obtained at 08.01.2021), we perform the parentage verification procedure directly according to the ICAR Guidelines for Parentage Verification and Parentage Discovery Based on SNP Genotypes.

Genomic evaluation

A single-step Genomic Best Linear Unbiased Prediction (ss-gBLUP) methodology is used to predict breeding values of both genotyped and non-genotyped animals in order to combine a pedigree and genomic information (Misztal *et al*., 2009). The fixed effects are as a combined effect year-season of first calving, agro-holding, farm, the age of heifer at first calving, mother's lactation number at the birth of the animal, and weather conditions. Estimated traits are 305d milk yield (kg), milk fat (absolute value, percentage), milk protein (absolute value, percentage), longevity (month).

The quality of data is determined by the coefficients of reliability, i.e. the proportion of explained variability by models with the use of all effects. Prediction accuracies of EBV and GEBV were expressed as square root of reliability, calculated from prediction error variance.

Total merit index

During the project, The Total Merit Index (KSI) was also developed (Miesenberger, J., and Fuerst, C., 2006). KSI is composed of estimated breeding values of the most valuable productive traits: 305d milk yield (kg), protein and fat content (%), and longevity (month). Each breeding value of a specific trait is weighted according to its economic importance under the Russian dairy production system. The economic importance is measured by the marginal profit per additional unit of the respective estimated breeding value in rubles assuming all other traits remain constant (Hazel, 1943). Economic data was provided by the Ministry of Agriculture of Udmurt Republic. Milk price is the largest contribution in revenue, such that milk with base protein and fat content equal to 25 rubles per 1 kg. The replacement cost is a main cost factor in the production system and amounts to 60 thousand rubles. In general, the profitability of milk production in the Russian system is about 10%. The final formula of KSI is following:

$$\begin{aligned}KSI = 2.6\ ½\ * BV_{\text{305-d milk yield}} + 2501\ ½\ * BV_{\text{protein content, \%}} + \\ + 2170\ ½\ *BV_{\text{fat content, \%}} + 98.6\ ½\ *BV_{\text{longevity, month}}, \end{aligned} \tag{1}$$

where BV corresponds to estimated breeding value derived from ss-gBLUP evaluation model. The Spearman correlation was calculated between KSI and the published Total Merit Indices of Iran, Israel and Czech Republic for auxiliary results verification (Sadeghi-Sefidmazgi *et al*., 2009, Krupová *et al*., 2018, Ezra and Weller, 2012).

Results and discussion

As part of the project for genomic reference database development, we collected more than 7,300 genotypes of the Udmurt Holstein dairy population in 2 month. In order to account for all available data, we also combined phenotypes and genotypes data from other regions of various previous local projects. Overall, by the end of 2020 we had 161 048 animals in the database with 17 924 milking cows and 1299 evaluated bulls from 5 regions of Russia.

During the data collection, Selex databases from dairy farms were received and processed. Mistakes, typos, duplicates, and deviations from expected values have been filtered from the data. Every animal has been provided with a unique identifier. There were 9 farms associated with the Udmurt region; overall 15.7% of the animal records were filtered. For each farm, on average 2% of animal records did not pass the data validation. For other 4 regions about 1% were filtered.

All the animal-parent pairs for which the genotypes were obtained were checked for consistency of parentage in the pedigree with parentage by genotype. In the case of a mismatch between the parents of an animal by pedigree and genotype, the search for the true parents was performed.

The obtained results are presented in table 2 and 3.

With the formed reference population of five regions, EBV and GEBV were estimated for both cows and bulls. The distribution of calculated reliability of 305d milk yield EBV and GEBV is shown in figure 1.

Table 2. The number of animals with accepted and excluded dams.

Genotypes in total	Dam is accepted	Dam is excluded	True dams found
8426	2086 (75.9%)	187 (7.9%)	114

Table 3. The number of animals with accepted and excluded sires.

Genotypes in total	Bulls genotypes	Sire is accepted	Sire is excluded	True sires found
8426	23	922 (86.5%)	92 (8.6%)	91

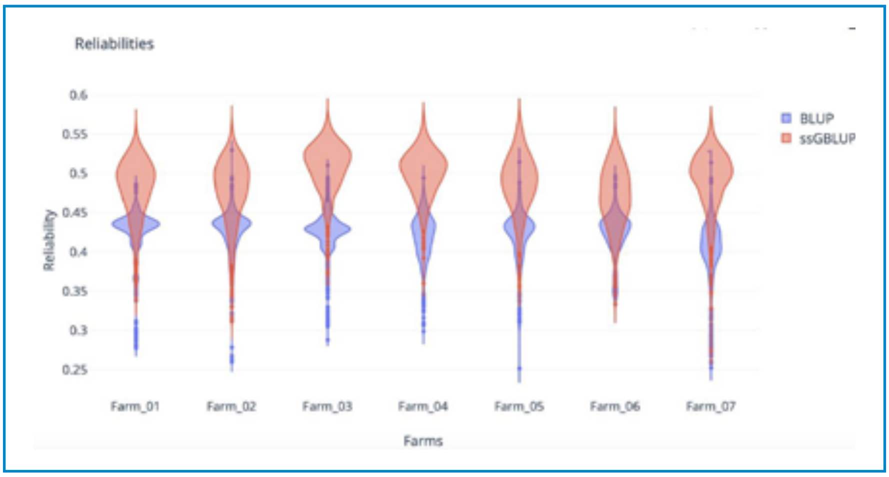

Figure 1. Reliability from BLUP and ss-gBLUP models of Udmurt region data.

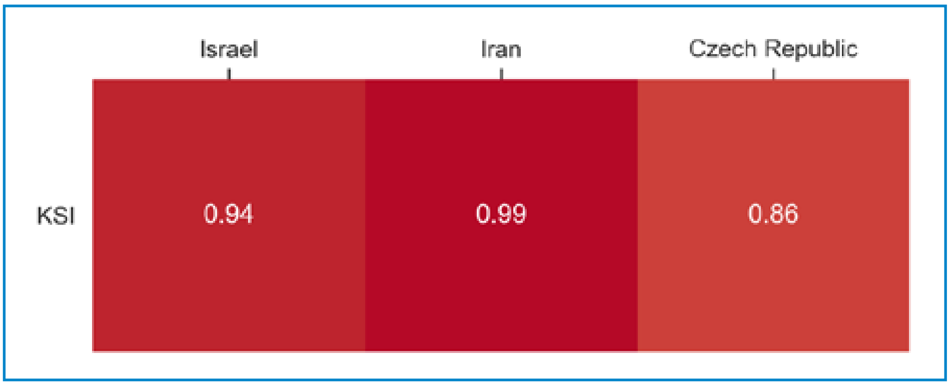

Figure 2. Spearman's correlation coefficients with foreign total merit indexes.

As expected, the result showed the larger reliability values with genomic information added. Considering the results of 305d milk yield prediction, the maximum reliability value of GEBV is 0.66 for cows and 0.93 for bulls, medium reliability value is 0.53 for cows and 0.58 for bulls. The lowest reliability performance both for GEBV and EBV is associated with low-quality phenotypic data from farm 4. Average gain in reliability values of ss-gBLUP are about 8%.

After the genomic evaluation process the total merit index, KSI, was implemented. Spearman's correlation coefficients between KSI and Israel, Iran, Czech Republic Indexes are 0.94, 0.99, 0.86, respectively (Figure 2). The 305d milk yield trait has the biggest contribution (76%) in KSI variance, since the milk price has the greatest impact on Russian agro-holding economics.

Moreover, a web service was implemented (https://app.ksitest.ru/) for local breeders. Ksitest web-service combined almost all features described above for each animal obtained database: their own information (name, farm, birth date), breeding values, genomic passports, total merit index, parentage verification, pedigree information. Genotyped animals also have results from genetic tests, such as monogenic disease (e.g. Bovine Leukocyte Adhesion Deficiency) or selection trait (e.g. polledness) statuses. As a result, web-service allows breeders:

- to rank animals by their genetic potential;
- to explore analytics (e.g. mean milk yield on the farm, dynamics of EBVs and GEBVs in years);
- to chose animals for embryo transfer or sale;
- analyze bull's performance (e.g. descendants production)
- receive documents (e.g. genomic passports and genotyping results in Illumina Final Report format).

In the next year 200 bull's genotypes and 22 500 cow's genotypes will be produced. The reference database will be expanded to 3 more Russian regions. Furthermore, to make a more efficient KSI total merit index for the dairy farmers, calving interval, fertility and conformation traits will be implemented and more economic parameters will be obtained by the end of 2021.

Conclusion

We aimed to create the first genomic reference database in Russia and perform genomic evaluation collaboratively with the Ministry of Agriculture of Udmurt Republic. We collected and validated more than 160 thousand animals data to the ksitest database. Besides, as of July 2021, the database contains records from 18 farms of 8 Russian regions. Currently we have data on 234,582 animals in total with 23,811 milking cows and 1561 evaluated bulls with 8974 genotypes and 328,917 phenotypic records. Genomic evaluation of obtained data outperformed the pedigree based evaluation. In addition, we incorporated all developments to our web service. Our results imply that genotyping information tends to reach higher reliability of prediction and higher selection accuracy. Genomic selection could replace currently common selection by dam's productivity in Russia and can improve accuracy for young animals without phenotypic information. Evaluations of new traits and more genotypes are expected in 2021. We expect that the accuracy of genomic evaluation will continue to improve with more data and these results pave the way for genomic selection implementation in Russia.

List of references

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M. and Lee, J.J., 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience, 4(1), pp 13742-015.

Ezra, E., and J. I. Weller. 2012. The Israeli Selection index. Page 28 in The Dairy Industry in Israel 2012. D. Hojman, Y. Malul, and T. Avrech, ed. Israel Dairy Board and Israel Cattle Breeders Association, Caesaria Industrial Park, Israel.

Hazel, L.N., 1943. The genetic basis for constructing selection indexes. Genetics, 28(6), pp.476-490.

Krupová, Z., Wolfová, M., Krupa, E., Zavadilová, L. and Pøibyl, J., Economic weights of traits in the breeding objective for Czech Holstein cattle.

Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M. and Chen, W.M., 2010. Robust relationship inference in genome-wide association studies. Bioinformatics, 26(22), pp.2867-2873.

Miesenberger, J. and Fuerst, C., 2006. Experiences in selecting on a total merit index in the Austrian Fleckvieh breed. Biotechnology in Animal Husbandry, 22(1-2), pp.17-27.

Misztal, I., Legarra, A. and Aguilar, I., 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. Journal of dairy science, 92(9), pp.4648-4655.

Sadeghi-Sefidmazgi, A., Moradi-Shahrbabak, M., Nejati-Javaremi, A. and Shadparvar, A., 2009. Estimation of economic values in three breeding perspectives for longevity and milk production traits in Holstein dairy cattle in Iran. Italian Journal of Animal Science, 8(3), pp.359-375.