**JAST (Journal of Animal Science and Technology) TITLE PAGE**

**Upload this completed form to website with submission**

| ARTICLE INFORMATION | Fill in information in each box below |
|---|---|
| **Article Type** | Review article |
| **Article Title (within 20 words without abbreviations)** | Enhancing Animal Breeding through Quality Control in Genomic Data - A Review |
| **Running Title (within 10 words)** | Quality Control in Genomic Data |
| **Author** | Jungjae Lee[1], Jong Hyun Jung[2] and Sang-Hyon Oh[3] |
| **Affiliation** | [1]Jenomics Jenetics Company, Pyeongtaek 17869, Korea<br>[2]Jung P&C Institute, Yongin 16950, South Korea<br>[3]Division of Animal Science, Gyeongsang National University, Jinju 52725, South Korea |
| **ORCID (for more information, please visit https://orcid.org)** | Jungjae Lee https://orcid.org/0000-0002-6145-8862<br>Jong Hyun Jung https://orcid.org/0000-0003-3667-7710<br>Sang-Hyon Oh https://orcid.org/0000-0002-9696-9638 |
| **Competing interests** | No potential conflict of interest relevant to this article was reported. |
| **Funding sources**<br>State funding sources (grants, funding sources, equipment, and supplies). Include name and number of grant if available. | This work was carried out with the support of "Cooperative Research Program for Agriculture Science and Technology Development (Project No. RS-2023-00232087)" Rural Development Administration, Republic of Korea. |
| **Acknowledgements** | N/A |
| **Availability of data and material** | Upon reasonable request, the datasets of this study can be available from the corresponding author. |
| **Authors' contributions**<br><br>Please specify the authors' role using this form. | Conceptualization: LJJ, JJH, OSH.<br>Data curation: LJJ, JJH, OSH.<br>Formal analysis: LJJ, JJH, OSH.<br>Methodology: LJJ, JJH, OSH.<br>Software: N/A<br>Validation: LJJ, JJH, OSH.<br>Investigation: LJJ, JJH, OSH.<br>Writing - original draft: LJJ, JJH, OSH.<br>Writing - review & editing: LJJ, JJH, OSH. |
| **Ethics approval and consent to participate** | This article does not require IRB/IACUC approval because there are no human and animal participants. |

**CORRESPONDING AUTHOR CONTACT INFORMATION**

| For the corresponding author (responsible for correspondence, proofreading, and reprints) | Fill in information in each box below |
|---|---|
| First name, middle initial, last name | Prof. Sang-Hyon Oh |
| Email address – this is where your proofs will be sent | Prof. Sang-Hyon Oh at shoh@gnu.ac.kr |
| Secondary Email address | |
| Address | Division of Animal Science, Gyeongsang National University, Jinju 52725, South Korea |
| Cell phone number | |
| Office phone number | |
| Fax number | |

# Enhancing Animal Breeding through Quality Control in Genomic Data - A Review

Jungjae *Lee*[1†], Jong Hyun *Jung*[2†] and Sang-Hyon *Oh*[3*]

† Both authors contributed equally to this manuscript.

*Corresponding Author: Sang-Hyon Oh

Tel: +82-55-772-3285, Fax: +82-55-772-3689, E-mail: shoh@gnu.ac.kr

[1]Jenomics Jenetics Company, Pyeongtaek 17869, Korea

[2]Jung P&C Institute, Yongin 16950, South Korea

[3]Division of Animal Science, Gyeongsang National University, Jinju 52725, South Korea

**Abstract**

High-throughput genotyping and sequencing has revolutionized animal breeding by providing access to vast amounts of genomic data to facilitate precise selection for desirable traits. This shift from traditional methods to genomic selection provides dense marker information for predicting genetic variants. However, the success of genomic selection heavily depends on the accuracy and quality of the genomic data. Inaccurate or low-quality data can lead to flawed predictions, compromising breeding programs and reducing genetic gains. Therefore, stringent quality control (QC) measures are essential at every stage of data processing. Quality control in genomic data involves managing single nucleotide polymorphism (SNP) quality, assessing call rates, and filtering based on minor allele frequency (MAF) and Hardy-Weinberg equilibrium (HWE). High-quality SNP data is crucial because genotyping errors can bias the estimates of breeding values. Cost-effective low-density genotyping platforms often require imputation to deduce missing genotypes. QC is vital for genomic selection, genome-wide association studies (GWAS), and population genetics analyses because it ensures data accuracy and reliability. This paper reviews QC strategies for genomic data and emphasizes their applications in animal breeding programs. By examining various QC tools and methods, this review highlights the importance of data integrity in achieving successful outcomes in genomic selection, GWAS, and population analyses. Furthermore, this review covers the critical role of robust QC measures in enhancing the reliability of genomic predictions and advancing animal breeding practices.

**Key words:** Animal Breeding, Genomic Selection, Quality Control, Single Nucleotide Polymorphism, Genome-Wide Association Studies

## Introduction

The rapid evolution of genomic technologies has transformed the landscape of animal breeding. High-throughput genotyping and sequencing provides breeders with access to vast amounts of genomic data and enables the precise selection of desirable traits [1]. These advancements have shifted traditional breeding methods to genomic selection, which leverages dense marker information to predict the genetic variants of individuals [2]. However, the success of genomic selection depends heavily on the accuracy and quality of the genomic data. Inaccurate or low-quality data can lead to inaccurate predictions that can compromise breeding programs and reduce their genetic gains [3]. Therefore, to ensure reliable predictions and maximize the potential of genomic selection, it is essential to implement stringent quality control (QC) measures at every stage of data processing.

Genomic data quality control has several key components including the management of single nucleotide polymorphism (SNP) quality, the assessment of call rates, and filtering based on minor allele frequency (MAF) and Hardy-Weinberg equilibrium (HWE) [4]. High-quality SNP data is indispensable because errors in genotyping can lead to biased estimates of breeding values, which decreases the effectiveness of selection strategies [5]. Moreover, cost-effective low-density genotyping platforms often suffer from incomplete marker data so it is necessary to use imputation to deduce the missing genotypes [6].

Quality control processes are crucial for genomic selection, genome-wide association studies (GWAS), and population genetics analyses. These processes help ensure that the genomic data is accurate, reliable, and free from biases introduced by genotyping errors, population stratification, or other confounding factors [7, 8]. This paper reviews quality control (QC) strategies for genomic data and their applications in animal breeding programs. By examining various QC tools and methods, this paper aims to show the critical role that data

integrity plays in achieving successful outcomes in genomic selection, GWAS, and population analyses [4, 5].

## Genotyping methods

*Whole-genome Sequencing (WGS)*

Whole-genome sequencing is a comprehensive method for analyzing the entire genome. Due to the decreased cost of sequencing and the ability to produce large amounts of genomic data, whole-genome sequencing has become a powerful tool for genomic research. SNP calling from WGS genomic data involves a series of critical steps to ensure accurate identification of genetic variants. The process starts with raw data preprocessing, where tools like FastQC evaluate the read quality [9]. This step is followed by trimming to remove adapters and low-quality bases by using either Trimmomatic or Cutadapt [10, 11].

The cleaned reads are then aligned to a reference genome with BWA-MEM or Bowtie2 to generate SAM/BAM files [12, 13]. These files are subsequently sorted, indexed, and processed to mark PCR duplicates with Samtools, while the base quality scores are recalibrated using GATK [14, 15]. Variant calling is performed using tools such as GATK's HaplotypeCaller**,** FreeBayes, or Bcftools, which identify SNPs based on differences between the sequenced reads and the reference genome [15, 16, 17].

In post-calling, variants undergo filtering to remove false positives via GATK's hard filtering or Variant Quality Score Recalibration (VQSR). The filtered SNPs are then annotated with functional information using tools like ANNOVAR or SnpEff [18, 19]. Quality checks include the use of VCFtools for statistical analysis and IGV for visualization, and ensure the reliability of the called SNPs [16, 20]. Joint genotyping across multiple samples and using population-specific reference panels are recommended to enhance the accuracy of SNP calling

in WGS.

*SNP arrays*

SNP arrays have significantly advanced genomic research in animal science by enabling the large-scale genotyping of SNPs. The development of SNP arrays began in the early 2000s to meet the demand for efficient and cost-effective methods to genotype large numbers of SNPs across the genome [21, 22]. Early arrays marked a significant advancement by allowing simultaneous genotyping of thousands of SNPs, facilitating genome-wide association studies (GWAS) and the study of genetic variation in populations [21].

Over time, these arrays have evolved to include higher-density SNPs to improve coverage and accuracy, as seen in the Illumina BovineSNP50 array which has become a standard tool in cattle genomics [23, 24]. Today, SNP arrays are essential for selecting desirable traits, estimating genetic merit, and managing inbreeding in animal breeding [1, 2]. Quality control of SNP array data is crucial for ensuring accurate and reliable results, and involves assessing call rates, filtering based on minor allele frequency (MAF), and checking for Hardy-Weinberg equilibrium [4]. Tools such as PLINK and GenomeStudio are commonly used in these QC processes [5, 25].



**QC in animal genomics**

*Minor Allele Frequency (MAF)*

Minor allele frequency is a key metric in genetic studies. It represents the frequency at which the less common allele occurs in a given population. MAF is important for identifying rare variants which may not significantly contribute to overall genetic variation but can be crucial in specific contexts. MAF is calculated by determining the frequency of both alleles at

a locus and taking the minimum of these two values. For example, if allele A has a frequency of 0.8 and allele a has a frequency of 0.2, the MAF would be 0.2. SNPs with very low MAFs, typically below 0.01 or 0.05, are often excluded from analyses because they may represent sequencing errors or lack statistical power in association studies [5].

Tools like PLINK and VCFtools [5, 16] are widely used to calculate MAF, with PLINK's --freq command being particularly popular [4]. In animal breeding, many researchers set threshold values for MAF to balance the need for sufficient variation while minimizing noise from rare variants. Typically, MAF thresholds in animal breeding studies range from 0.01 to 0.05 depending on the study's objectives and the population structure being analyzed. For instance, a study on dairy cattle by Pryce et al. [26] and Kim et al. [27] sed a MAF threshold of 0.01 to ensure that the SNPs included were sufficiently informative for genomic predictions while also minimizing the influence of rare variants that might lead to spurious associations.

*Call rate*

Call rate is another critical quality control metric that measures the proportion of successfully genotyped samples for a specific SNP. A high call rate indicates that a SNP has been consistently detected across the sample population, while a low call rate may suggest issues with the genotyping process, such as poor quality or technical errors [7].

The call rate is calculated by dividing the number of successful genotype calls for a SNP by the total number of samples, then multiplying by 100 to express it as a percentage.

$$Call\ Rate = \frac{Number\ of\ successfully\ genotyped\ markers\ (or\ samples)}{Total\ number\ of\ markers\ (or\ samples)} \times 100$$

For instance, if 95 out of 100 samples have a successful genotype call for a SNP, the call rate would be 95% [4]. Normally, markers with a call rate less than 95% are removed, though other studies have set more stringent or lenient thresholds depending on the study design and objectives. For example, some studies have removed markers with a call rate below 99% to ensure extremely high data quality[28], while others have used a more relaxed threshold of 90% when working with larger datasets[29].

Tools like PLINK, SNP & Variation Suite (SVS), and GenomeStudio are widely used for calculating and filtering SNPs based on call rates because they offer robust functionalities for quality control in genomic studies. PLINK is particularly popular due to its comprehensive command-line interface, where the --missing command calculates call rates at both the marker and sample levels, allowing researchers to easily filter out SNPs and samples that fall below the desired threshold [5]. SNP & Variation Suite (SVS) offers a user-friendly graphical interface and integrates various statistical tools, making it ideal for complex datasets and large-scale studies [30]. GenomeStudio by Illumina is another powerful tool specifically designed for managing and analyzing genotyping data with features for calculating call rates, identifying low-quality markers, and visualizing data for further inspection [25]. These tools are essential for ensuring that only high-quality data is used in subsequent analyses to improve the reliability of genomic outcomes.



*Hardy-Weinberg Equilibrium*

Hardy-Weinberg equilibrium (HWE) is a fundamental principle in population genetics. It states that allele and genotype frequencies in a population will remain constant from generation to generation in the absence of evolutionary influences [31]. Testing for HWE is an important quality control step because deviations from this equilibrium can indicate issues such as

genotyping errors, population stratification, or selection pressures [32]. To test for HWE, the observed genotype frequencies are compared to the expected frequencies under equilibrium conditions. For a biallelic SNP with alleles A and a, the expected genotype frequencies are $p^2$ for AA, 2pq for Aa, and $q^2$ for aa, where p and q represent the allele frequencies [33]. A chi-square test is commonly used to assess whether the differences between the observed and expected frequencies are statistically significant. Tools like PLINK and VCFtools are used to perform HWE tests [34]. SNPs that show significant deviation from HWE, typically with a p-value less than 0.001, are often excluded from analyses to prevent biases that could arise from genotyping errors or other confounding factors [4].

175 Table 1. Tool list for quality control processes

| Tools | Function | Reference |
|---|---|---|
| GEMMA | Application of linear mixed models and related models to GWAS | [4] |
| PLINK | Run association analyses and perform QC and regression steps | [5] |
| FastQC | Quality control checks on raw sequence data | [9] |
| Trimmomatic | Trim and crop FASTQ data | [10] |
| Cutadapt | finds and removes adapter sequences, primers, poly-A tails | [11] |
| BWA-MEM | produce multiple primary alignments for different part of a query sequence | [12] |
| Bowtie2 | aligning sequencing reads to long reference sequences | [13] |
| Samtools | Manipulate alignments in the SAM, BAM, and CRAM formats | [14] |
| GATK | Variant calling using sequencing data | [15] |
| VCFtools | Summarize, filter out, convert data into other file formats | [16] |
| FreeBayes | Bayesian genetic variant detector designed to fine SNPs | [17] |
| SnpEff | Annotation on genetic variants and predicts their effects on genes | [18] |
| ANNOVAR | Generate gene-based annotation | [19] |
| IGV | Visualization tool to simultaneously integrate and anlyze multiple types of genomic data | [20] |
| GenomeStudio | Normalize, cluster, and call genotypes | [25] |
| SVS | Perform analyses and visualizations on genomic and phenotypic data | [33] |
| BEAGLE | Genotype calling, phasing, and genotype imputation | [39] |
| Fimpute | Haplotype estimation or phasing and genotype imputation | [40] |
| Impute2 | Genotype imputation and haplotype phasing | [47] |
| Minimac | performs imputation with pre-phased haplotypes | [48] |

176

177 **Application**

*Population analysis*

Population analysis is invaluable for genomic studies in animal science because it enables researchers to assess the genetic structure, diversity, and evolutionary dynamics within and between populations. Accurately characterizing population structures is crucial for identifying subpopulations, measuring inbreeding levels, and understanding the genetic background of breeding populations, all of which are essential for maintaining genetic diversity and improving selection outcomes [35]. Tools such as PLINK, ADMIXTURE, and STRUCTURE are commonly employed to detect key characteristics for understanding the genetic landscape of animal populations, such as population stratification, admixture, and genetic differentiation [5, 36]. For example, ADMIXTURE provides estimates of individual ancestry proportions. These estimates allow researchers to detect mixed genetic backgrounds that could influence trait analysis [36]. Quality control measures, such as filtering based on MAF, HWE, and genotyping call rates ensure the data used for population analysis is reliable [4,37]. MAF filtering helps exclude rare alleles that may introduce noise or result from genotyping errors [5]. Similarly, HWE filtering removes SNPs that deviate from expected frequencies due to selection or population substructures in order to prevent potential biases in the analysis [37]. Proper quality control improves the accuracy of population structure analyses and mitigates the risk of confounding in subsequent analyses such as GWAS and genomic selection [4]. By accurately characterizing population structures, researchers can identify unique genetic markers and enhance their understanding of trait inheritance, and then design breeding strategies that optimize genetic gain and preserve diversity to support sustainable livestock production [35, 36].

*GWAS*

Genome-wide association studies (GWAS) are powerful tools for identifying genetic variants associated with complex traits in animal breeding such as growth traits, disease resistance, reproductive traits, and carcass traits [2, 4]. The reliability of GWAS findings hinges on rigorous quality control (QC) procedures that ensure high-quality data throughout the process. This begins with careful study design and population selection, where potential confounders like population stratification are addressed through methods such as Principal Component Analysis (PCA) and linear mixed models to correct for genetic structure within the population [38]. Phenotype data must be accurately collected and screened for outliers to minimize noise. Genotype data undergoes thorough QC, including filtering SNPs based on call rates, MAF, and deviations from HWE [4, 5]. For instance, SNPs with low call rates are excluded to avoid unreliable data that could lead to false-positive associations, while MAF filtering focuses the analysis on common variants that are more likely to have sufficient statistical power to detect true associations. HWE filtering is employed to remove SNPs that significantly deviate from expected allele frequencies because such deviations may indicate genotyping errors or underlying selection pressures [5]. To reduce redundancy and computational burden, linkage disequilibrium (LD) pruning is performed and missing genotypes are often imputed via reference panels using Fimpute or BEAGLE [39, 40]. Tools like PLINK and GEMMA are widely used to implement QC measures and conduct association tests because they offer a robust framework for analyzing large genomic datasets [4]. Statistical analysis in GWAS is carried out using models appropriate for the trait under study, and corrections for multiple testing to mitigate the risk of false positives and meta-analysis may be employed when integrating results from multiple studies [41]. To ensure the robustness and high accuracy of the GWAS models, a 5-fold cross-validation is often used. In this method, the datasets are divided into five subsets. The model is iteratively trained on four subsets and tested

on the remaining one to help validate the model's accuracy and mitigate overfitting [42]. The results from GWAS offer valuable genetic variants for traits which can be targeted in marker-assisted selection and genomic selection programs. Genomic selection aims to ultimately improve the genetic merit of livestock populations [2].

*Genomic Selection*

Genomic selection (GS) allows for the selection of animals based on SNP markers [43]. With the introduction of genomic selection, animal breeding has dramatically advanced by overcoming the limitations of traditional selection methods like best linear unbiased prediction (BLUP) and marker-assisted selection [43, 44]. GS relies on dense SNP data to estimate genomic breeding values, which are used to predict an individual's genetic potential for economically important traits [2]. The accuracy of genomic selection models is dependent upon the quality of the genomic data and the reliability of GS models can be enhance significantly by the inclusion of imputation methods to handle missing or low-density SNP data [45]. Imputation is beneficial in low-density platforms because it allows for the cost-effective use of genotyping while still leveraging the power of high-density SNP information. Imputation increases the accuracy of genomic predictions by inferring missing genotypes in order to improve the reliability of estimated breeding values even with fewer markers [46]. Several imputation tools, including FImpute [40], Beagle [39], Impute2 [47], and Minimac [48] are widely used in animal breeding to enhance the accuracy of genomic selection models. Therefore, strict quality control is essential [49]. Quality control methods, such as filtering SNPs based on call rates, MAF, and HWE, is critical to ensuring that the data is vigorous and reliable. High call rates are important because missing data can introduce bias and reduce the reliability of genomic estimated breeding values. Similarly, excluding SNPs with low MAF

helps to avoid the noise associated with rare variants that may have little impact on prediction accuracy. Ensuring that SNPs conform to HWE expectations also prevents the inclusion of markers affected by selection, mutation, or other factors that could bias the genomic selection models [4, 5]. Advanced computational tools, such as GBLUP (Genomic Best Linear Unbiased Prediction) and ssBLUP (single-step BLUP), and Bayesian methods (BayesA, BayesB, BayesC) integrate SNP effects across the genome to enhance the precision of breeding value predictions [50, 51]. By using high-quality genomic data, genomic selection enables breeders to make more accurate decisions that lead to faster genetic gains and the improvement of traits such as milk yield, growth rate, and carcass weight in livestock. This approach not only enhances the efficiency of breeding programs but also contributes to the long-term sustainability and productivity of animal populations [35].

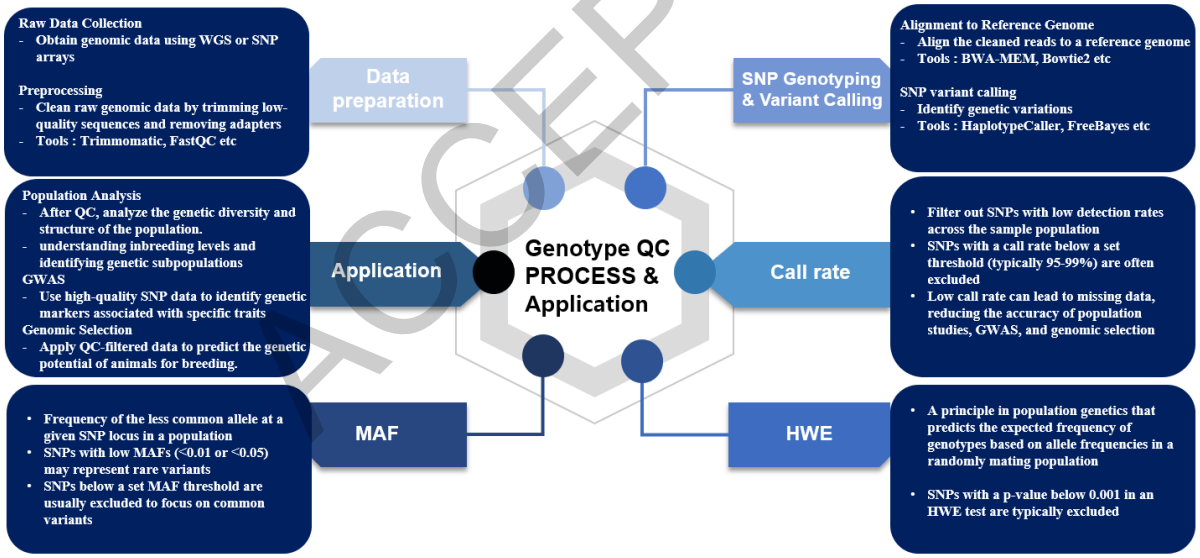


Figure 1. Overall flowchart from data preparation to Application in animal breeding

## Conclusion

High-throughput genotyping and sequencing has significantly advanced the field of animal breeding by enabling precise selection for desirable traits. However, the success of genomic selection hinges on the accuracy and quality of the genomic data used. Rigorous quality control (QC) measures are essential to ensure data integrity. These measures include SNP quality management, call rate assessment, and filtering based on minor allele frequency (MAF) and Hardy-Weinberg equilibrium (HWE). These QC processes are crucial for genomic selection, genome-wide association studies (GWAS), and population genetics analyses. Implementing stringent QC strategies enhances the reliability of genomic predictions, which improves breeding programs and genetic gains. By maintaining high standards of data quality, researchers and breeders can make informed decisions that lead to sustainable and productive advancements in animal breeding.

## Reference


1. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME. Invited review: Genomic selection in dairy cattle: Progress and challenges. Journal of dairy science. 2009;92(2):433-43. https://doi.org/10.3168/jds.2008-1646

2. Meuwissen TH, Hayes BJ, Goddard M. Prediction of total genetic value using genome-wide dense marker maps. genetics. 2001;157(4):1819-29. https://doi.org/10.1093/genetics/157.4.1819

3. Wiggans G, VanRaden P, Cooper T. The genomic evaluation system in the United States: Past, present, future. Journal of dairy science. 2011;94(6):3202-11. https://doi.org/10.3168/jds.2010-3866

4. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. Nature protocols. 2010;5(9):1564-73. https://doi.org/10.1038/nprot.2010.116

5. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American journal of human genetics. 2007;81(3):559-75. https://doi.org/10.1086/519795

6. García-Ruiz A, Cole JB, VanRaden PM, Wiggans GR, Ruiz-López FJ, Van Tassell CP. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. Proceedings of the National Academy of Sciences. 2016;113(28):E3995-E4004. https://doi.org/10.1073/pnas.1519061113

7. Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, Bhangale T, et al. Quality control and quality assurance in genotypic data for genome-wide association studies. Genetic epidemiology. 2010;34(6):591-602. https://doi.org/10.1002/gepi.20516

8. Marees AT, De Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. International journal of methods in psychiatric research. 2018;27(2):e1608. https://doi.org/10.1002/mpr.1608

9. Andrews S. FastQC: a quality control tool for high throughput sequence data. Cambridge, United Kingdom; 2010.

10. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20. https://doi.org/10.1093/bioinformatics/btu170

11. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet journal. 2011;17(1):10-2. https://doi.org/10.14806/ej.17.1.200

12. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. bioinformatics. 2009;25(14):1754-60. https://doi.org/10.1093/bioinformatics/btp324

13. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods. 2012;9(4):357-9. https://doi.org/10.1038/nmeth.1923.

14. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. bioinformatics. 2009;25(16):2078-9. https://doi.org/10.1093/bioinformatics/btp352

15. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research. 2010;20(9):1297-303. https://doi.org/10.1101/gr.107524.110

16. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156-8. https://doi.org/10.1093/bioinformatics/btr330

17. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:12073907. 2012. https://doi.org/10.48550/arXiv.1207.3907

18. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. fly. 2012;6(2):80-92. https://doi.org/10.4161/fly.19695

19. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research. 2010;38(16):e164-e. https://doi.org/10.1093/nar/gkq603

20. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al.

Integrative genomics viewer. Nature biotechnology. 2011;29(1):24-6. https://doi.org/10.1038/nbt.1754

21. Kennedy GC, Matsuzaki H, Dong S, Liu W-m, Huang J, Liu G, et al. Large-scale genotyping of complex DNA. Nature biotechnology. 2003;21(10):1233-7. https://doi.org/10.1038/nbt869

22. Matsuzaki H, Dong S, Loi H, Di X, Liu G, Hubbell E, et al. Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. Nature Methods. 2004;1(2):109-11. https://doi.org/10.1038/nmeth718

23. Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, et al. Development and characterization of a high density SNP genotyping assay for cattle. PloS one. 2009;4(4):e5350. https://doi.org/10.1371/journal.pone.0005350

24. llumina, Inc. BovineSNP50 Genotyping BeadChip. Technical Note; 2009.

25. Illumina, Inc. GenomeStudio Software. Technical Note; 2010.

26. Pryce JE, Haile-Mariam M, Goddard ME, Hayes BJ. Identification of genomic regions associated with inbreeding depression in Holstein and Jersey dairy cattle. Genetics Selection Evolution. 2014;46:1-14. https://doi.org/10.1186/s12711-014-0071-7

27. Kim S, Lim B, Cho J, Lee S, Dang C-G, Jeon J-H, et al. Genome-wide identification of candidate genes for milk production traits in Korean Holstein cattle. Animals. 2021;11(5):1392. . https://doi.org/10.3390/ani11051392

28. Verma SS, de Andrade M, Tromp G, Kuivaniemi H, Pugh E, Namjou-Khales B, et al. Imputation and quality control steps for combining multiple genome-wide datasets. Frontiers in genetics. 2014;5:370. . https://doi.org/10.3389/fgene.2014.00370

29. Lee J, Kim Y, Cho E, Cho K, Sa S, Kim Y, et al. Genomic analysis using Bayesian methods under different genotyping platforms in Korean Duroc pigs. Animals. 2020;10(5):752. https://doi.org/10.3390/ani10050752

30. *SNP & Variation Suite ™ (Version 8.x) [Software]. Bozeman, MT: Golden Helix, Inc. Available from http://www.goldenhelix.com.*

31. Edwards A. GH Hardy (1908) and hardy–weinberg equilibrium. Genetics.

2008;179(3):1143-50. https://doi.org/10.1534/genetics.104.92940

32. Nielsen R, Slatkin M. An introduction to population genetics: theory and applications: Sinauer Associates Sunderland, MA; 2013.

33. Mayo O. A century of Hardy–Weinberg equilibrium. Twin Research and Human Genetics. 2008;11(3):249-56. https://doi.org/10.1375/twin.11.3.249

34. Graffelman J, Weir B. Testing for Hardy–Weinberg equilibrium at biallelic genetic markers on the X chromosome. Heredity. 2016;116(6):558-68. https://doi.org/10.1038/hdy.2016.20

35. Hill WG. Understanding and using quantitative genetic variation. Philosophical Transactions of the Royal Society B: Biological Sciences. 2010;365(1537):73-85. https://doi.org/10.1098/rstb.2009.0203

36. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome research. 2009;19(9):1655-64. https://doi.org/10.1101/gr.094052.109

37. Minelli C, Thompson JR, Abrams KR, Thakkinstian A, Attia J. How should we use information about HWE in the meta-analyses of genetic association studies?. International journal of epidemiology. 2008;37(1):136-46. https://doi.org/10.1093/ije/dym234

38. McVean G. A genealogical interpretation of principal components analysis. PLoS genetics. 2009;5(10):e1000686. https://doi.org/10.1371/journal.pgen.1000686

39. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. The American Journal of Human Genetics. 2009;84(2):210-23. https://doi.org/10.1016/j.ajhg.2009.01.005

40. Sargolzaei M, Chesnais JP, Schenkel FS. A new approach for efficient genotype imputation using information from relatives. BMC genomics. 2014;15:1-12. https://doi.org/10.1186/1471-2164-15-478

41. Pearson TA, Manolio TA. How to interpret a genome-wide association study. Jama. 2008;299(11):1335-44. https://doi.org/10.1001/jama.299.11.1335

42. Kohavi R, editor A study of cross-validation and bootstrap for accuracy estimation and model selection. Ijcai; 1995: Montreal, Canada.

43. Meuwissen T, Hayes B, Goddard M. Genomic selection: a paradigm shift in animal breeding. Animal frontiers. 2016;6(1):6-14. https://doi.org/10.2527/af.2016-0002

44. Zhang Z, Zhang Q, Ding X. Advances in genomic selection in domestic animals. Chinese science bulletin. 2011;56:2655-63. https://doi.org/10.1007/s11434-011-4632-

45. Berry D, Kearney J. Imputation of genotypes from low-to high-density genotyping platforms and implications for genomic selection. Animal. 2011;5(8):1162-9. https://doi.org/10.1017/S1751731111000309

46. García-Ruiz A, Cole JB, VanRaden PM, Wiggans GR, Ruiz-López FJ, Van Tassell CP. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. Proceedings of the National Academy of Sciences. 2016;113(28):E3995-E4004. https://doi.org/10.1073/pnas.1519061113

47. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS genetics. 2009;5(6):e1000529. https://doi.org/10.1371/journal.pgen.1000529

48. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nature genetics. 2012;44(8):955-9. https://doi.org/10.1038/ng.2354

49. VanRaden P. Symposium review: How to implement genomic selection. Journal of Dairy Science. 2020;103(6):5291-301. https://doi.org/10.3168/jds.2019-17684

50. Habier D, Fernando RL, Kizilkaya K, Garrick DJ. Extension of the Bayesian alphabet for genomic selection. BMC bioinformatics. 2011;12:1-12. https://doi.org/ 10.1186/1471-2105-12-186

51. Misztal I, Legarra A, Aguilar I. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. Journal of dairy science. 2009;92(9):4648-55. https://doi.org/10.3168/jds.2009-2064