Vostry, L., Vostra-Vydrova, H., Moravcikova, N., Kasarda, R., Halvonik, A., Shihabi, M., Cubric-Curik, V., Curik, I. (2026) Evaluating the influence of sample size and SNP density on genomic estimates of effective population size in livestock, *Proceedings 13th World Congress on Genetics Applied to Livestock Production* 23934: https://doi.org/10.31274/wcgalp.23934 ©2026 Vostry, et al. CC BY-NC-ND

Evaluating the influence of sample size and SNP density on genomic estimates of effective population size in livestock

L. Vostry[1*], H. Vostra-Vydrova[1], N. Moravcikova[2], R. Kasarda[2], A. Halvonik[2], M. Shihabi[3], V. Cubric-Curik[3], and I. Curik[3,4]

[1]Czech University of Life Science Prague, Prague, Czechia; [2]Slovak University of Agriculure in Nitra, Nitra, Slovakia; [3]University of Zagreb, Zagreb, Croatia; Hungarian University of Agriculture and Life Sciences (MATE), Kaposvár, Hungary; [*]Corresponding author: vostry@af.czu.cz

Introduction

One of the key factors determining population survival is genetic diversity, which is strongly shaped by population size. Many modern breeding techniques, such as genomic selection, often lead to a severe reduction in genetic diversity within populations. This reduction genetic diversity can increase levels of inbreeding and genetic drift. As a result of the combined effects of inbreeding and genetic drift, individual fitness declines in populations with low genetic diversity, while the effectiveness of selection is reduced. An accurate estimate of effective population size (Ne) is therefore essential for understanding genetic diversity and for the sustainable management of conservation and animal breeding programs under ongoing selection. Effective population size is most commonly estimated using pedigree or molecular genetic data. The increasing availability of advanced genomic technologies has enabled Ne estimation based on single nucleotide polymorphism (SNP) markers and is particularly valuable when complete pedigree information is unavailable. One method applicable to both small and large populations is the linkage disequilibrium (LD) approach (Hill, 1981), which has been widely used to estimate both contemporary and historical effective population size. The commonly used approach is based on the theoretical relationship between linkage disequilibrium (LD) and effective population size (N_e) and is implemented within a more sophisticated modelling framework in the software GONE2 (Santiago *et al*., 2025). In addition to estimating contemporary Ne, GONE2 can infer historical Ne by applying a genetic algorithm to reconstruct the trajectory of effective population size values that best minimizes the sum of squared differences between observed D^2 values. All estimated effective population sizes based on gametic or linkage disequilibrium represent the size of an idealised, often Wright–Fisher, population exhibiting the same degree of genetic drift and changes in gametic or linkage disequilibrium as the population under study. Further details on effective population size based on linkage disequilibrium and potential sources of bias are provided by Waples (2021). The accuracy of Ne estimates derived from genomic data is primarily influenced by the number of SNPs included in the analysis and the number of individuals sampled. The present study evaluated the sensitivity of effective population size estimates to variation in both parameters by systematically testing different combinations of sample size and marker density.

Materials & Methods

Empirical and simulations data collection. This study used publicly available genome-wide high-density SNP data from several domestic species. The datasets included two cattle breeds: Holstein (2,448 individuals; 612k SNPs; Seabury *et al*., 2023) and Brown Swiss (54 individuals; 450k SNPs; Barbato *et al*., 2020); two horse breeds: Thoroughbred (370 individuals; 640k SNPs; Fewcett *et al*., 2019) and Peruvian (183 individuals; 550k SNPs; Momen *et al*., 2023); two sheep breeds: Suffolk (304 individuals, 563k SNPs; Wilson *et al*.,

2022) and Pag sheep (45 individuals, 470k SNPs; Drzaic *et al*., 2022); and two dog breeds: Slovak Cuvac (42 individuals; 220k SNPs) and Setter (226 individuals, 174 k SNPs; Dahlgren *et al*., 2016).
To assess the stability of Ne estimation, computer simulations were performed using the software SLiM (Haller and Messer, 2019), which implements an individual-based forward Wright–Fisher model of reproduction. Random-mating populations with a constant size (N = 100 individuals) and discrete generations were simulated for up to 1000 generations. The simulated genome consisted of 20 chromosomes, each 100 Mb in length. The recombination rate between nucleotides was set to 10^{-8}, corresponding to a total chromosome length of 1 Morgan. resulting in approximately 160,000 SNPs available for analysis. Quality control was performed using PLINK v1.9 (Chang *et al*., 2015). Only autosomal SNPs with known chromosomal positions were retained to avoid sex-related biases. Individuals and SNPs with call rates below 0.9 were excluded. No minor allele frequency filtering was applied.

Effective Population Size. The estimation of the contemporary effective population size based on linkage disequilibrium was used in this study. Estimates of current and historical Ne were obtained from gametic/linkage disequilibrium information implemented in the GONE2 software (Santiago *et al*., 2025).

Analysis Design. The influence of sample size and SNP density on effective population size estimates from genomic data was evaluated in two independent steps. (i) Effect of sample size: Only breeds with a high number of genotyped individuals (Holstein, Suffolk, Thoroughbred, Peruvian and Setter) were selected from each analysed species. From each dataset, subsets of individuals were randomly sampled, starting with a minimum of 10 individuals and increasing in increments of 10 up to 100, then in increments of 50 until the full dataset was reached. (ii) Effect of SNP density: All available datasets were included. Analyses began with 5,000 randomly selected SNPs, progressively increasing by 5,000 up to 100,000, and then by 50,000 until reaching the full SNP density. Each scenario, for both sample size and SNP density analyses, was replicated ten times to assess variability and reproducibility. A threshold for Ne stability was determined using piecewise linear regression applied to the mean and standard deviation (SD) of Ne estimates across 10 replicate analyses. This method fits separate linear trends to different portions of the data and estimates the breakpoint using the segmented regression framework (Toms & Lesperance, 2003; Muggeo, 2008). The stability threshold indicates where increases in dataset size, either more loci or individuals, no longer substantially improve the precision of Ne estimates. An example of this procedure for the Thoroughbred breed is shown in Figure 1.

Results & Discussion

The first analyses focused on the relationship between the stability of Ne estimates and sample size. Only populations with a higher number of genotyped individuals were included in this analysis. For each subset of individuals randomly sampled from the analysed populations, descriptive statistics (e.g., mean, SD) of Ne estimates were calculated after ten replications. Thresholds for Ne stability, determined using piecewise linear regression applied to the mean or SD of Ne estimates across the 10 replications, are presented in Table 1. The estimated thresholds depended on the mean of 10 replications were higher than those estimated using the SD of 10 replications. This difference may be influenced by the fact that the analysed populations can represent subdivided populations. Random sampling can result in different proportions individuals from different subpopulations, thereby increasing the estimated Ne. This pattern corresponds to the observed threshold values for populations expected to have a low degree of subdividing (e.g., Peruvian) achieved lower threshold of

based on mean and SD values than populations expected to have a high degree of subdivision (e.g., Holstein). The lowest thresholds values were recorded in simulated data, where population subdivisions were not considered. These lowest thresholds correspond to the minimum recommended number of genotyped individuals (n = 20) for the use of the GONE2 program (Santiago *et al*., 2020). Our empirical results indicate that sampling 20 individuals is insufficient to guarantee stable Ne estimates across all populations, whereas a minimum of 40 individuals provides more reliable estimates in most cases. Further increases in the number of individuals beyond this threshold contributed only negligible improvements.

Table 1. Recommended minimum thresholds for the number of individuals and SNPs required to obtain stable estimates of effective population size, based on mean and SD of 10 replications.

				Animals		SNPs	
Species	Breed	Individuals	SNPs	Mean	SD	Mean	SD
Cattle	Holstein	2448	612k	65	40	15 k	20 k
	Braun Swiss	54	450k	-	-	25 k	25 k
Sheep	Suffolk	304	563k	30	40	25 k	25 k
	Pag Island Sheep	45	470k	-	-	20 k	20 k
Horse	Thoroughbred	370	640k	40	21	25 k	25 k
	Peruvian	183	550k	30	21	25 k	25 k
Dog	Setter	226	174k	30	40	25 k	25 k
	Slovak Cuvac	42	220k	-	-	25 k	20 k
Theoretical	Simulations	100	160k	20	20	25 k	25 k

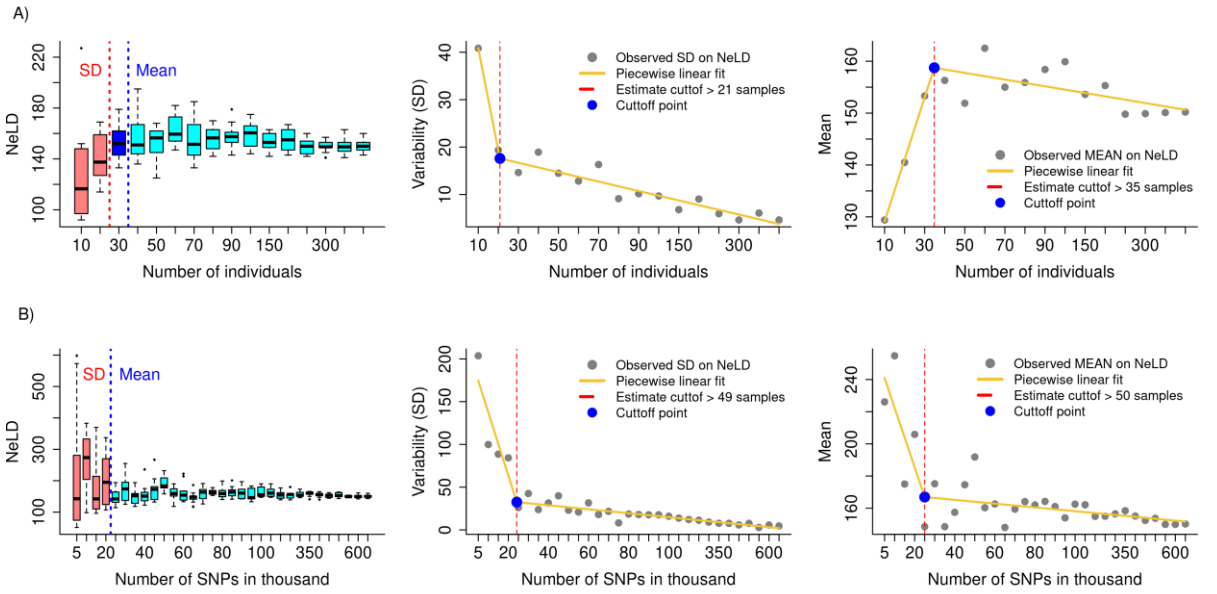

Figure 1. Illustration of the methodological approach used to determine the minimum individuals (A) and SNP (B) threshold for stable N_e estimates using the SD and mean of 10 replications.

Based on the previous results regarding the optimal number of individuals, only populations with more than 40 genotyped individuals were included in the subsequent analysis aimed at estimating the threshold for the minimum number of SNPs required for stable Ne estimates (Table 1). The results of the analysis of the minimum SNP number necessary for stable Ne estimation showed that both applied statistics, the mean and the standard deviation (SD)

across 10 replicates, showed very similar threshold values. The lowest threshold was observed in the Holstein breed when using the mean across 10 replicates, at approximately 15,000 SNPs. The lower threshold values observed in the Holstein population are attributable to the larger number of genotyped individuals (n = 2,448). Such large numbers of genotyped individuals are not always available. For the remaining populations, a threshold of 25,000 SNPs was identified for obtaining stable Ne estimates. The only exception was the Pag Island Sheep population. The Pag Island Sheep represents a closed population with very low genetic diversity, which affects Ne estimates. Based on these results, we recommend a minimum marker density of approximately 25,000 SNPs for obtaining stable estimates of effective population size. The largest deviations from the expected Ne values were observed when both the number of individuals and the number of SNPs were low.

Conclusions

Using empirical and simulated genomic data analyzed with GONE2, we show that the stability of linkage disequilibrium–based Ne estimates depend strongly on sample size and SNP density, with at least 40 individuals and approximately 25,000 SNPs required for stable estimates in most populations. These findings provide practical guidance for the cost-effective design of genomic studies, while future work will extend this analysis to additional livestock populations and species, further refining recommendations for cost-effective genomic applications.

Acknowledgements

This research was supported by GACR 24-14325L, QL25020022, APVV-20-0161, HRZZ IP-2022-10-6914 and the NextGenerationEU project (NPOO.C3.2.R3-I1.04.0141).

References

Barbato M., Hailer F., Upadhyay M., *et al*. (2020) Scientific Report 10, 1279 (2020). https://doi.org/10.1038/s41598-020-57880-4

Dahlgren S., Ziener M.L. and Lingaas F. (2016) Animal Genetics. 47. https://doi.org/10.3389/fgene.2023.1201628

Drzaic I., Curik I., Lukic B., *et al*. (2022) Front. Genet. 13:940736. https://doi: 10.3389/fgene.2022.940736

Fawcett J.A., Sato F., Sakamoto T., *et al*. (2019) PLoS One. 14:e0218407. https://doi.org/10.1371/journal.pone.0218407

Haller B.C., and Messer P.W. (2019) Molecular Biology and Evolution 36:632–637. https://doi.org/10.1093/molbev/msy228

Hill W.G. (1981) Advances in Applied Probability 13, 4–4. https://doi.org/10.2307/1426455

Momen M., Brauer K., Patterson M.M., *et al*. (2023) Frontiers in Genetics. 14:1201628. https://doi.org/10.3389/fgene.2023.1201628

Muggeo V.M.R. (2008) R News 8:20-25

Santiago E., Köpke C. and Caballero A. (2025) Nature Communications 16, 6054. https://doi.org/10.1038/s41467-025-61378-w

Seabury C.M., Smith J.L., Wilson M.L., *et al*. (2023) G3. 3. https://doi.org/10.1093/g3journal/jkad043

Santiago E., Novo I., Pardiñas A.F., *et al*. (2020) Molecular Biology and Evolution. 37, 3642–3653. https://doi.org/10.1093/molbev/msaa169

Waples, R.S. (2021) Journal of Heredity. 112, 535–539. https://doi.org/10.1093/jhered/esab042

Wilson C.S., Petersen J.L., Blackburn H.D. *et al*. (2022) Journal of Heredity, 2022, 113, 431–443. https://doi.org/10.1093/jhered/esac026

Toms J.D. and Lesperance M.L. (2003) Ecology 84:2034-2041 https://doi.org/10.1890/02-0472