Vollmer, L. Geibel, J., Kistner, L., Weigend, A., Worat, C., Weishaar, R., Petig, M., Hinrichs, D., Weigend, S., Reimer, C. (2026) Effects of sequencing depth on the reliability of heterozygosity estimation in pigs, *Proceedings 13th World Congress on Genetics Applied to Livestock Production* 24222: https://doi.org/10.31274/wcgalp.24222
©2026 *Vollmer, et al.* CC BY-NC-ND

Effects of sequencing depth on the reliability of heterozygosity estimation in pigs

L. Vollmer*[1], J. Geibel[1,2], L. Kistner[3], A. Weigend[1], C. Worat[1], R. Weishaar[4], M. Petig[4], D. Hinrichs[3], S. Weigend[1,2], C. Reimer[1,2]

[1]Friedrich-Loeffler-Institut, Neustadt, Germany; [2]University of Goettingen, Göttingen, Germany; [3]University of Kassel-Witzenhausen, Witzenhausen; [4]Züchtervereinigung Schwäbisch Hällisches Schwein e.V., Wolpertshausen, Germany; Corresponding author: lea.vollmer@fli.de

Introduction

German native pig breeds, such as the Swabian-Hall pig, are increasingly threatened by the shift towards intensive production systems, resulting in population declines inducing a small effective population size and therefore a loss of valuable genetic resources. Nevertheless, native breeds may offer unique adaptive traits, including resilience to environmental stressors and efficient utilization of non-conventional feeds, which would make them particularly suited for organic and extensive farming systems (Muñoz *et al.*, 2029, Herrero-Medrano *et al*., 2014 und FAO, 2012). Conservation and sustainable utilization of these native breeds require comprehensive genomic characterization to assess population structure, inbreeding, and adaptive potential. Genotyping arrays are the current standard for genome-wide diversity assessments but are limited to predefined variants and may miss rare or population-specific variation (Herrero-Medrano *et al*., 2014).

Advances in sequencing technologies have enabled population-scale genomic analyses beyond arrays. Low-coverage long-read sequencing provides a flexible, cost-effective alternative, as the increased read length improves alignment accuracy and reduces reference bias, which is particularly advantageous at low sequencing depth (Lou *et al.,* 2021). However, the reliability of diversity estimates, e.g. heterozygosity, at low coverage remains unclear, and systematic evaluation is necessary to determine minimal possible sequencing effort for accurate inference. This study aims to assess whether low-coverage sequencing can provide robust and consistent estimates of genetic diversity in endangered pig breeds. We compare common single nucleotide polymorphisms (SNP) array with low-coverage long-read sequencing, applying both an imputation-based and analysis of next generation sequencing data (ANGSD) approach, to evaluate the accuracy and stability of diversity estimates across methods and sequencing depths.

Materials & Methods

Samples. EDTA full-blood was collected from eight Swabian-Hall pigs originating from three farms affiliated with the "Züchtervereinigung Schwäbisch-Hällisches Schwein e.V." by approved veterinarians within the genomic diversity monitoring scheme of the German Gene Bank for Farm Animals, in accordance with national animal welfare legislation (Ref. No.: NRW LV22-2025-0014167; BW RPS35-9185-108/102). This sample served as a proof-of-concept dataset to evaluate the robustness of low-coverage sequencing-based diversity estimates as a practical tool for routine genetic diversity monitoring.

Sequencing, genotyping and raw data preparation. DNA was extracted from all samples using the Qiagen Blood and Tissue Kit, and quality and concentration were measured by spectrophotometry. Barcoded libraries were prepared with the Oxford Nanopore Technologies Native Barcoding Kit V14, multiplexing six animals per library, achieving 10–15x coverage per animal. Samples were also genotyped using the Illumina PorcineSNP80 BeadChip.

Basecalling was performed with Dorado (simplex mode), and reads were mapped to the pig reference genome Sscrofa11.1 (Warr *et al.*, 2020) using Minimap2 in the map-out mode (Li *et al.*, 2018). Raw BAM files, halotype-based imputation datasets with STICH at simulated coverage levels (see Reimer *et al.*, 2026), and array VCF genotypes were analyzed. Calculations based on sequence-based methods were applied to chromosome 1, whilst the array analysis was carried out across the entire genome. To assess sequencing depth effects on genetic diversity, BAM files were subsampled from 0.2x to 10x depth of coverage (DoC) to standardise comparisons.

Genotype-based analysis and distance calculation. Heterozygosity per sample was calculated as the proportion of heterozygous genotypes among all SNPs. Minor allele frequencies at selected SNPs were estimated using ANGSD with likelihood-based methods to account for uncertainty in low-coverage sequencing, analyzing all sites across chromosome 1. Spearman rank correlations assessed relationships between samples. Genetic distances were derived from genotype matrices by computing pairwise covariance and correlation, then converting to distance matrices using 1−r. Hierarchical clustering with average linkage produced unrooted phylogenetic trees, exported in Newick format. Tree similarity was quantified using Robinson-Foulds distance. All analyses were automated with Bash scripts integrating R workflows, ensuring reproducibility and efficient processing of low-coverage genomic data (Yu et al., 2017).

Results & Discussion

To evaluate how different analytical paradigms influence diversity estimates at low sequencing depth, we compared read-based inference using ANGSD and imputation with STITCH against a commonly used SNP array. Traditional genotype calling approaches, such as pileup or GATK-based calling, are known to perform poorly at low sequencing depth due to increased uncertainty in genotype assignment. Heterozygosity was used as the primary diversity metric because it is commonly applied in conservation genetics to quantify individual-level genetic variation. Importantly, heterozygosity can be estimated using SNP arrays, imputed genotypes, and probability-based sequencing data, rendering it well-suited for comparing analytical approaches at different sequencing depths. It is important to note that heterozygosity values differ conceptually between methods. SNP arrays and STITCH-imputed genotypes report heterozygosity at pre-selected polymorphic sites, where variability is assumed, resulting in a high degree of heterozygosity. On the other hand, ANGSD estimates heterozygosity per base pair across all sequenced positions, including invariant sites, and thus typically yields lower values. Heterozygosity estimates from the array data ranged between 28.6% and 31.9%, reflecting the fact that the SNPs on the array were pre-selected for common polymorphism. Consequently, the array overrepresents variable sites, resulting in higher observed heterozygosity compared with sequencing-based estimates.

For STITCH, heterozygosity ranged from 8.0 to 21.5% at 10x, 10.8 to 21.2% at 5x, and 9.0 to 19.8% at 1x coverage, showing minor to no changes across coverage levels. Lower values at 10x coverage likely reflect the conservative calling behaviour of the imputation model, which tends to adjust uncertain heterozygotes toward homozygosity, as well as the limited coverage and statistical uncertainty in genotype inference. Overall, STITCH heterozygosity shows minor variation across coverage levels, suggesting relative robustness of imputation-based estimates.

For read-based inference using ANGSD, heterozygosity at 10x coverage ranged from 0.13% to 0.23% (Fig. 1, Left) and down-sampling led to a nearly linear decrease in observed heterozygosity of 20% to 45% at 1x coverage (Fig. 1, Right). At lower coverage, estimates were more variable, occasionally exceeding higher-coverage values, reflecting the increased uncertainty in genotype likelihoods when few reads support a site. Overall, the trend of

decreasing heterozygosity with lower coverage was consistent across all samples, highlighting the impact of sequencing depth on the accuracy of heterozygosity estimation.
Imputed genotypes consistently exhibited higher heterozygosity than ANGSD estimates at low coverage, likely due to the influence of reference haplotypes and linkage disequilibrium, where nearby DNA variants are correlated. In such cases, the imputation method can anticipate missing genotypes based on these correlations, which can inflate heterozygosity when coverage is very low (≤1x). The markedly lower heterozygosity values obtained with ANGSD compared with array-based and imputed estimates reflect fundamental differences in calculation. While array-based and STITCH-imputed heterozygosity is derived from called genotypes at a predefined set of SNPs, ANGSD estimates heterozygosity from genotype likelihoods across all analyzed sites, including many invariant positions. Consequently, ANGSD values represent the fraction of heterozygous sites relative to all genomic positions considered, rather than a curated SNP panel. Although the absolute values are much lower, they remain informative for relative comparisons across samples and sequencing depths.

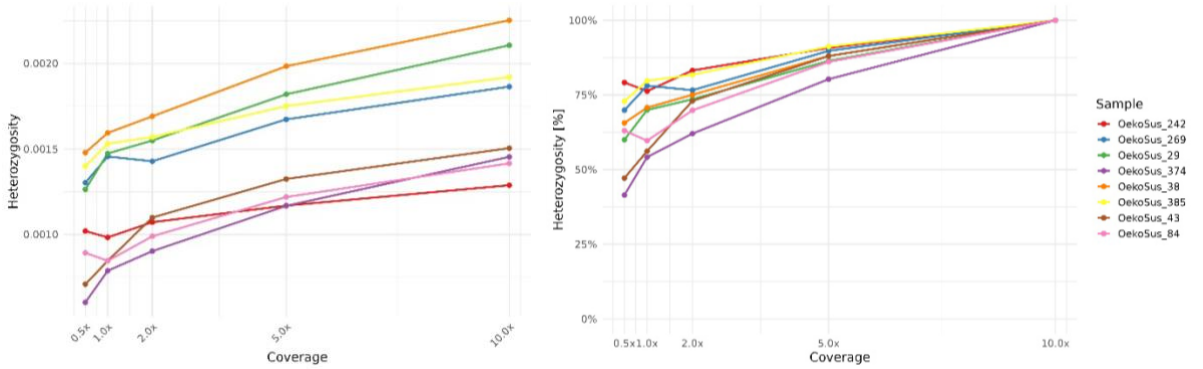

Figure 1. Heterozygosity per sample across coverage (ANGSD) in absolute values (Left) and linear decrease of heterozygosity over coverage in percent (Right).

While absolute heterozygosity values provide information on genome-wide diversity, it is also important to assess whether the relative ranking of individuals remains consistent across sequencing depths and analytical methods. To this end, we evaluated the stability of heterozygosity-based rankings with decreasing coverage. Spearman rank correlations were used to quantify the similarity in ordering of individuals across datasets, with higher values indicating more consistent rankings.
Spearman rank correlations among ANGSD-derived heterozygosity estimates indicated that rankings were unstable at 0.2x coverage. Moderate to high correlations (0.81–0.98) were observed at 0.5x–5x (Fig. 2 Left). For STITCH-imputed genotypes, rank correlations were generally higher, even at 0.2x, reflecting the stabilizing effect of the reference haplotype panel and linkage disequilibrium. Correlations with array data ranged 0.40–0.69 for ANGSD and 0.38–0.76 for STITCH. Within the imputation dataset, rank correlations between adjacent coverages were high (0.76–1.00) from 0.5x upwards. Overall, these results indicate that both imputation and ANGSD provide more consistent heterozygosity rankings with increasing coverage, whereas DoC of 0.2x produces unstable rankings in both methods.
To further investigate whether differences in coverage affect downstream analyses, we assessed phylogenetic stability using Robinson-Foulds (RF) distances (Fig. 2 Right). RF distance quantifies how different two phylogenetic trees are, with larger values indicating more divergent branching patterns. Trees from very low coverage (e.g., ANGSD 0.5x and imputed 0.2x) exhibit the largest RF distances when compared to higher-coverage trees, indicating unstable phylogenetic reconstructions at low DoC. In contrast, trees from moderate to high coverage (≥0.5x for ANGSD and ≥0.5x for imputed data) show more consistent topologies.

Across coverage levels, imputation generally produces lower RF distances than ANGSD. Notably, for ANGSD, coverage levels of ≥1× show minimal differences in tree topology, indicating that individual animals can be reliably analysed at this depth.

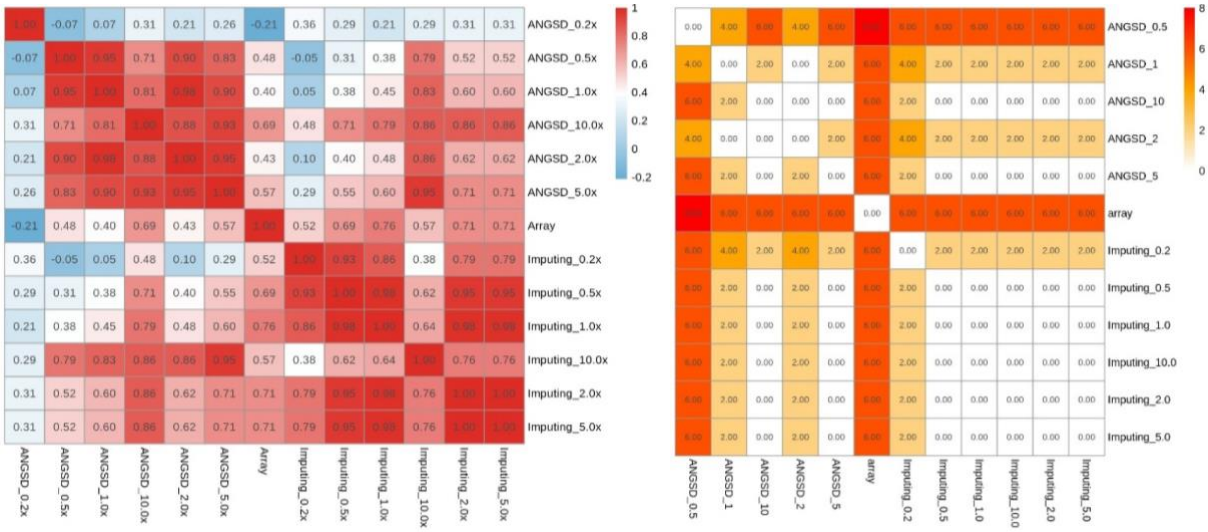

Figure 2. Heatmap showing **Left** rho-values to indicate spearman rank correlations between heterozygosity and **Right** pairwise RF distances between trees derived from ANGSD, STITCH-imputed genotypes, and array-based data.

Conclusions

Building on previous work demonstrating that imputation can recover reliable genotypes from low-coverage sequencing (Reimer *et al.*, 2026), our results show that, within the studied Swabian-Hall pig population, even single-animal analyses yield robust heterozygosity-based rankings. Importantly, this indicates that for applications such as gene bank candidate selection, accurate prioritization of individuals can be achieved at sequencing depths as low as 1x, providing a cost-effective strategy without compromising reliability.

Acknowledgements

The project was funded by the Federal Ministry of Agriculture, Food and Regional Identity (BMLEH) via the Federal Office for Agriculture and Food (BLE) under the Federal Organic Farming Scheme (FKZ 2823OE030A). We thank the breeding associations and the Baden-Württemberg Pig Health Service for their support in obtaining blood samples, and the Association for Bioeconomy Research (Förderverein Bioökonomieforschung e.V.) for providing access to their custom array.

References

FAO: Intergovernmental Technical Working Group on Animal Genetic Resources for Food and Agriculture. 7th Session. Rome: (2012) Available at https://www.fao.org/4/mf227e/mf227e.pdf , visited on: 28-01-2026

Herrero-Medrano, J., Megens, H., Groenen, M., Bosse, M., Pérez-Enciso, M.,(2014) BMC Genomics. https://doi.org/10.1186/1471-2164-15-601

Li, Heng (2018) Bioinformatics (Oxford, England) 34 (18), https://doi.org/10.1093/bioinformatics/bty191

Lou, R., Jacobs, A., Wilder, A., Overgaard Therkildsen, N. (2021) Molecular ecology 30 (23), S. 5966–5993. https://doi.org/10.1111/mec.16077

Muñoz, M.; Bozzi, R., García-Casco, J., Núñez, Y., et al. (2019) Scientific reports 9 (1), S. 13546. https://doi.org/10.1038/s41598-019-49830-6

Reimer, C., Vollmer, L., Geibel, J., Kistner, L., et al. (2026). Submitted for publication at WCGALP.

Warr, A., Affara, N., Aken, B., et al. (2020) GigaScience 9. https://doi.org/10.1093/gigascience/giaa051

Yu, G., Smith, D., Zhu, H., Guan, Y., Tsan-Yuk Lam, T. (2017) Methods Ecology Evolution 8(1):28-36. https://doi.org/10.1111/2041-210X.12628