Co-funded by the European Union

PRIMA SCALA-MEDI

PRIMA PARTNERSHIP FOR RESEARCH AND INNOVATION IN THE MEDITERRANEAN AREA

Deliverable D8.3

On-line tools for breed and individual choice for breeding and conservation

Due date: 31/01/2026

Lead contractor: CNR-IBBA

Partners involved: UCSC, CMCC, IDELE, INRAE

Dissemination level: PU (public)

Co-funded by
the European Union

Executive Summary

This report presents the development of computational tools to support the characterisation, conservation, and sustainable management of local livestock breeds within the SCALA-MEDI project. Two complementary frameworks were developed to address key aspects of conservation genetics and adaptive breeding.

The first part of this technical report describes the tools for breed characterisation and management. The tool focuses on the estimation of the Effective Population Size (Ne), a fundamental parameter in conservation genetics and livestock breeding, using a reproducible and scalable Nextflow-based computational pipeline designated nf-neestimator. Here we illustrate the theoretical framework underlying Ne estimation, the architecture and usage of the nf-neestimator pipeline (v0.2.1), its practical implementation on High-Performance Computing (HPC) infrastructure, and the interpretation of outputs from both empirical and simulated datasets. The pipeline automates the conversion of standard genotypic data formats, the execution of NeEstimator v2.X via the linkage disequilibrium (LD) method, and the generation of tabular summary outputs, thereby ensuring analytical reproducibility and interoperability. The pipeline is publicly available at the cnr-ibba GitHub repository and constitutes a key deliverable within the SCALA-MEDI project for the genomic characterisation of Mediterranean local breeds.

The second framework presented in this report addresses genotype by environment interactions (G×E) through the development of a Reaction Norm Simulation workflow. This tool models genotype performance across environmental gradients and estimates both genetic merit and environmental sensitivity, supporting the identification of robust and climate-resilient animals. Developed using simulated sheep data and mixed-model approaches in R, the framework demonstrated clear differences among genotypes in their environmental responses and provides a reproducible methodology applicable to different species and traits.

Together, these tools provide practical and accessible resources to support conservation, breeding decisions, and the long-term sustainability of local genetic resources. They contribute to the SCALA-MEDI objectives by strengthening capacity for genomic analysis, improving the management of genetic diversity of livestock populations in the challenging environmental conditions of the Mediterranean region.

1. Tools for breed characterisation and management

1.1 Introduction

The sustainable management of local livestock breeds and the design of conservation programmes requires a robust understanding of the genetic structure and demographic history of target populations. Among the most informative parameters in this context is the Effective Population Size (Ne), which quantifies the rate of genetic drift and the accumulation of inbreeding within a population over time.

Ne can be estimated from demographic data, pedigree records, or genomic marker data. Given the limited availability of reliable pedigree information for many Mediterranean local breeds — a central focus of the SCALA-MEDI project — genomic-based approaches represent the most practical and informative estimation strategy.

Multiple methods exist for the genomic estimation of Ne, each relying on different underlying principles:

- Temporal methods, based on the rate of change in allele frequencies across generations (reflected in the inbreeding coefficient ΔF);
- Coancestry-based methods, measuring the increase in genetic relatedness among individuals over time;
- Linkage Disequilibrium (LD)-based methods, which exploit the correlation between allele frequencies at pairs of neutral loci as a proxy for historical and contemporary population size.

The growing availability of genome-wide SNP data has made LD-based estimation particularly attractive, as it requires a single sampling point and can provide estimates of Ne integrated over multiple generations. The software NeEstimator v2.X implements all three approaches within a unified framework, making it well-suited for comparative analyses across populations.

To overcome challenges related to reproducibility, scalability, and format compatibility, the IBBA-CNR team developed nf-neestimator, a Nextflow-based workflow that automates the entire Ne estimation pipeline from PLINK binary input to tabular output. This report presents the theoretical

background, implementation details, and validation results of this pipeline as an online tool that can be routinely employed for breed and individual choice in conservation and breeding strategies.

1.2 Theoretical Background: Effective Population Size

Definition and Significance

The effective population size (Ne) is defined as the size of an idealised Wright-Fisher population that would exhibit the same rate of genetic drift or inbreeding as the observed population. It is a fundamental parameter in population genetics, with direct implications for:

- The rate of loss of genetic diversity through genetic drift;
- The accumulation of inbreeding and its associated fitness consequences;
- The efficacy of selection relative to random genetic drift;
- The design and monitoring of conservation and breeding programmes.

A low Ne signals a population at elevated risk of genetic erosion, inbreeding depression, and reduced adaptive capacity — conditions particularly prevalent in endangered or numerically small local breeds.

Genomic Estimation via Linkage Disequilibrium

The LD-based method for Ne estimation — implemented in NeEstimator as the LDNe approach — relies on the expected relationship between the squared correlation of allele frequencies at pairs of loci (r^2) and Ne:

$$E[r^2] \approx 1/(3Ne) + 1/n$$

where n is the sample size. The bias-corrected version of this estimator (Waples 2006; Waples & Do 2010) is the recommended implementation for empirical genomic data, and represents the default method employed within nf-neestimator. This method yields a contemporary Ne estimate, reflecting population size over the most recent generations.

NeEstimator v2.X

NeEstimator v2.X (Do et al., 2014) is a widely adopted software package for Ne estimation, implementing three single-sample methods and one two-sample (temporal) method:

- Bias-corrected LD method (Hill 1981; Waples 2006; Waples & Do 2010);
- Heterozygote-excess method (Pudovkin et al. 1996; Zhdanova & Pudovkin 2008);
- Molecular co-ancestry method (Nomura 2008);
- Temporal F-statistics method (Waples 1989; Nei & Tajima 1981; Pollak 1983; Jorde & Ryman 2007).

The software provides point estimates together with parametric and jackknife confidence intervals for all methods. Input data must be provided in FSTAT or GENEPOP format, necessitating pre-processing of raw SNP data generated by standard genotyping pipelines.

1.3 The nf-neestimator Pipeline

Overview and Rationale

nf-neestimator (v0.2.1) is a Nextflow-based computational workflow developed by Paolo Cozzi (IBBA-CNR) and publicly available at https://github.com/cnr-ibba/nf-neestimator. The pipeline was designed to address three key limitations of running NeEstimator directly: (i) the requirement for specific, non-standard input formats; (ii) the absence of native support for bootstrapped confidence interval estimation; and (iii) the difficulty of executing multiple independent runs in a reproducible and traceable manner. By wrapping the estimation procedure within a Nextflow workflow, nf-neestimator ensures:

- Full analytical reproducibility through containerisation (Docker and Singularity supported);
- Scalability via native integration with HPC schedulers (e.g., SLURM);
- Portability across computing environments;
- Automated bootstrapping over randomised sub-samples of individuals.

Pipeline Architecture

The pipeline is structured into four sequential processing stages:

Stage	Designation	Description	Tools
1	INPUT	Load genotype data in PLINK binary format (.bed/.bim/.fam)	PLINK v1.9

2	PREPARE	Convert PLINK to GENEPOP format for NeEstimator ingestion	PGDSpider v2.1.1.5
3	RUN	Execute NeEstimator with LDNe method; perform bootstrapped sampling	NeEstimator v2.X
4	SUMMARIZE	Parse output files; generate CSV tables with *Ne* values, CIs, and sample size	Custom scripts

Input Data Requirements

The pipeline accepts genotype data in PLINK binary format, comprising three files: the binary genotype matrix (.bed), the variant information file (.bim), and the sample information file (.fam). Internally, the pipeline invokes PGDSpider to convert these data to the GENEPOP format required by NeEstimator, using the PED2GENEPOP conversion module.

GENEPOP format is a plain-text format in which each individual is represented on a separate line with their multi-locus genotype encoded as concatenated allele codes. Missing genotypes are encoded as '000000'. The first line is a free-text title; subsequent lines list locus names (one per line or comma-separated), followed by 'Pop' delimiters separating sampling groups.

Configuration

Pipeline behaviour is governed by two configuration files provided by the user prior to execution. The params.json file specifies run parameters including input file paths, output directory, sample sizes for bootstrapping, and the number of bootstrap replicates (Figure 1). The custom.config file defines HPC-specific execution parameters including scheduler directives such as SLURM queue assignments, memory allocations, CPU core counts, and wall-time limits (Figure 2).

params.json — Pipeline Parameters

```
{
    "input"        : "data/genotypes.bed",
    "outdir"       : "results/ne_estimates",
    "min_samples"  : [20, 50, 100],
    "n_replicates" : 100,
    "pcrit"        : 0.02,
    "random_seed"  : 42
}
```

Figure 1. params.json — Pipeline run parameters including input paths, output directory, sample sizes for bootstrapping, number of replicates, and critical allele frequency threshold.

custom.config — HPC/SLURM Configuration

```
// custom.config

process {

    withName: 'NEESTIMATOR' {
        executor  = 'slurm'
        queue     = 'parallel'
        cpus      = 4
        memory    = '16 GB'
        time      = '4h'
    }

    withName: 'CONVERT_TO_GENEPOP' {
        executor  = 'slurm'
        queue     = 'short'
        cpus      = 2
        memory    = '8 GB'
        time      = '1h'
    }

}
```

Figure 2. custom.config — HPC/SLURM execution directives specifying queue assignments, memory allocations, CPU cores, and wall-time limits for each pipeline process.

The NeEstimator configuration block within the pipeline activates the LD method exclusively (configuration code: 1 0 1 1), which enables the bias-corrected LDNe estimator while suppressing output from alternative methods to reduce computational overhead and output volume.

Execution on HPC Infrastructure

The pipeline is launched via a SLURM batch submission script (launch-neestimator.sh), which invokes Nextflow with the appropriate profile settings for the target HPC environment (Figure 3). The recommended invocation submits the workflow to the SLURM scheduler, which then manages resource allocation and process execution across available compute nodes.

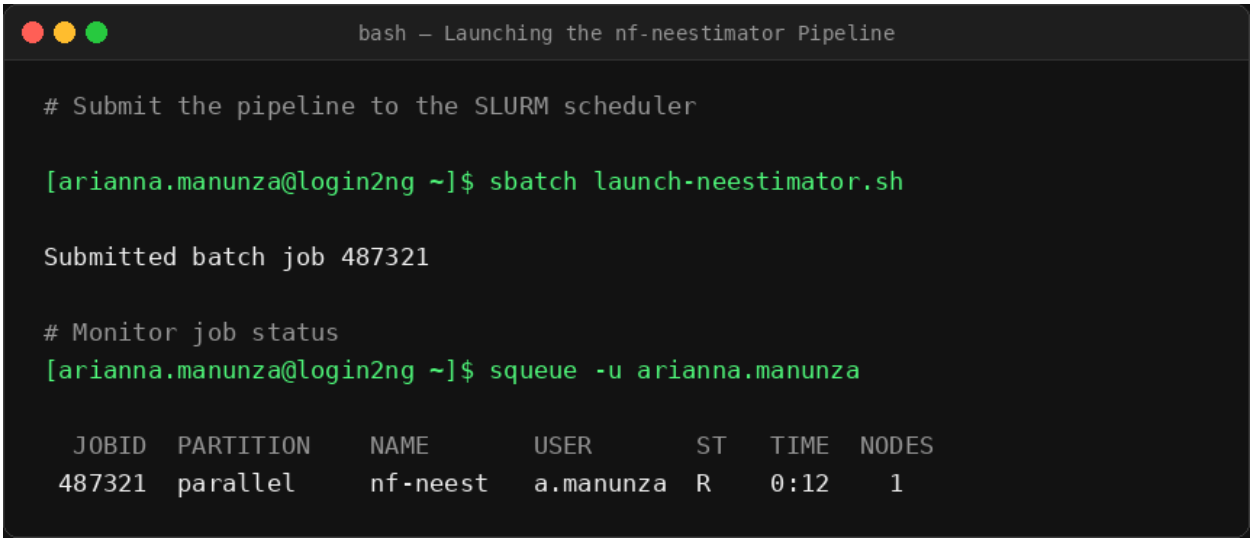


Figure 3. Terminal session illustrating the submission of the nf-neestimator pipeline to a SLURM-managed HPC cluster using sbatch, and subsequent job status monitoring with squeue.

A critical operational constraint of the current pipeline version (v0.2.1) concerns the management of temporary disk space: NeEstimator produces temporary files during LD computation, and launching multiple instances simultaneously in a shared HPC environment may exhaust the temporary storage partition. It is therefore recommended to limit concurrency or configure appropriate resource throttling within the Nextflow configuration.

Pipeline version management is handled through Nextflow's native pull mechanism, allowing users to pin execution to a specific stable release (Figure 4).

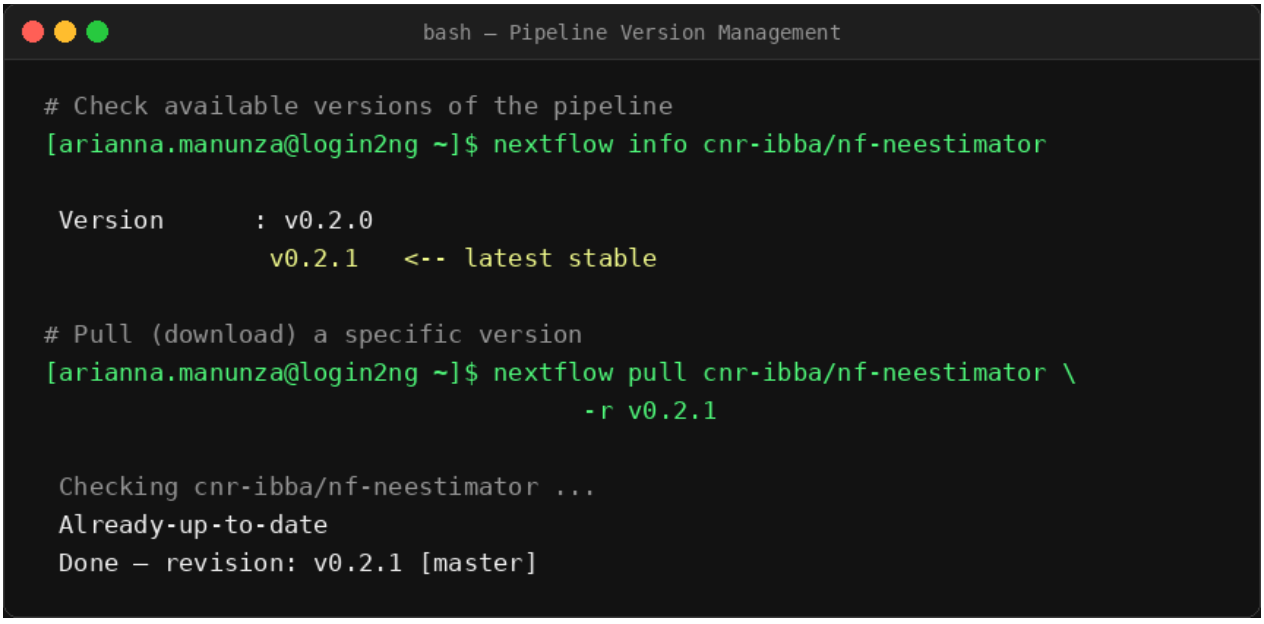


Figure 4. Pipeline version management using nextflow info and nextflow pull commands to inspect available releases and download a specific version (v0.2.1) of nf-neestimator.

Pipeline Output

Upon successful completion, the pipeline generates structured CSV output files containing, for each bootstrap replicate and sample size: *Ne* point estimate (LDNe); lower and upper bounds of the confidence interval; effective sample size; and number of loci retained after filtering. These outputs are suitable for downstream statistical analysis, visualisation, and integration into population monitoring reports.

1.4 Validation and Case Studies

Empirical Validation: Churra Sheep Breed

To assess the performance of nf-neestimator under empirical conditions, Ne estimates were generated for the Churra sheep breed, a local breed from the mediterranean area, using three sub-sample sizes: 20, 50, and 100 individuals. For each sample size, 100 independent bootstrap replicates were performed.

Results were expressed as mean Ne estimates with associated confidence intervals, benchmarked against a reference Ne value calculated from the full dataset. Estimates converged toward the reference Ne as sample size increased, confirming the expected reduction in estimation variance with larger samples. At n = 20, estimates exhibited considerable variability reflecting sampling stochasticity; at n = 100, mean estimates closely approximated the reference value with narrow confidence intervals. The workshop presentation of this case study is reproduced in Figure 5, which also shows the corresponding Frontiers in Genetics publication (Manunza et al., 2025) from which the empirical results were drawn.

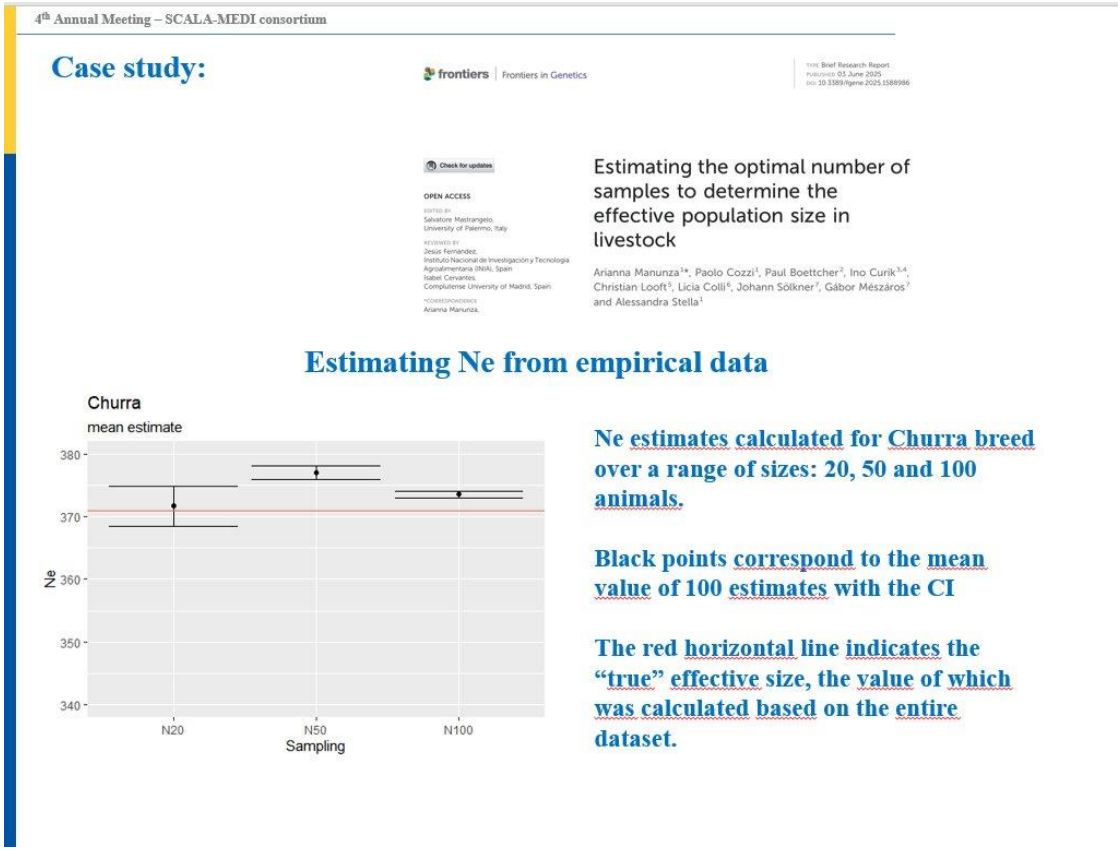


Figure 5. Workshop slide (SCALA-MEDI 4th Annual Meeting, Paris 2025) presenting the empirical case study: Ne estimates for the Churra sheep breed at sample sizes N20, N50, and N100 (100 bootstrap replicates each), alongside the corresponding Frontiers in Genetics publication (Manunza et al., 2025). Black points: mean Ne estimate with confidence intervals; red horizontal line: reference Ne calculated from the full dataset.

Simulation-Based Validation

Complementary validation was performed using simulated genotypic data generated with QMSim v2.0 (Sargolzaei & Schenkel, 2009). Two simulated populations were compared: POP1, with phenotype-based selection and constant population size; and POP3, with phenotype-based selection but incorporating a recent population bottleneck. Results confirmed accurate Ne recovery under stable demographic conditions and correct identification of reduced Ne following the simulated bottleneck. The workshop presentation of these simulation results is shown in Figure 6.

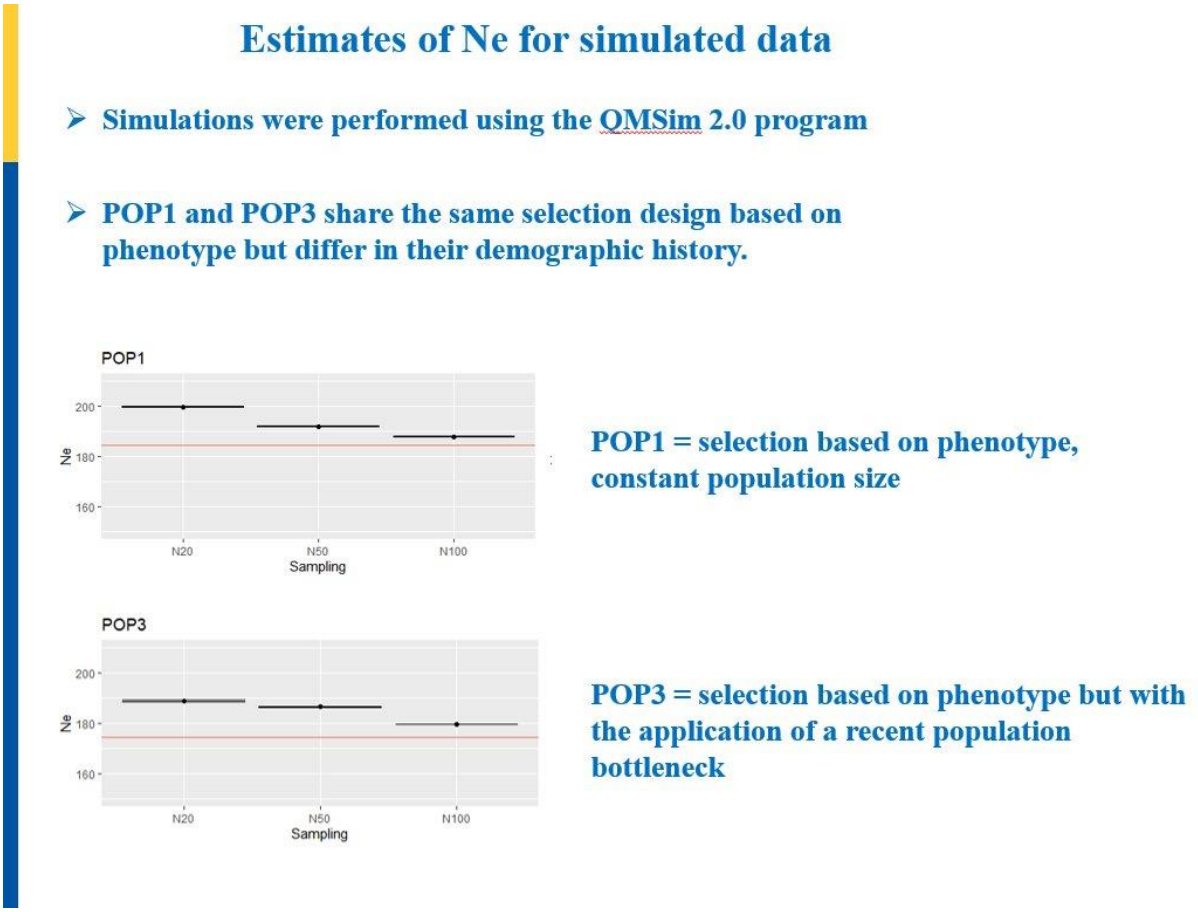


Figure 6. Workshop slide (SCALA-MEDI 4th Annual Meeting, Paris 2025) presenting Ne estimates from QMSim v2.0 simulations. POP1 (top): phenotype-based selection with constant population size. POP3 (bottom): phenotype-based selection with a recent population bottleneck. Black points: mean Ne estimate with confidence intervals; red horizontal line: true simulated Ne.

1.5 Known Issues and Recommendations

Issue	Recommendation
Temporary file exhaustion on HPC	Limit the number of concurrent Nextflow jobs; configure SLURM resource constraints to avoid saturation of the /tmp partition.
NeEstimator temporary file conflicts	Tracked on GitHub (cnr-ibba/nf-neestimator Issue #1). Avoid running multiple pipeline instances in the same working directory simultaneously.
Pipeline version compatibility	Always specify the target version explicitly (e.g., -r v0.2.1) to ensure reproducibility and avoid unintended use of development builds.
Input format compatibility	Only PLINK binary format is natively supported. Users with VCF or other formats must perform prior conversion using PLINK v1.9.

1.6 Integration with the SCALA-MEDI Project

Within the broader SCALA-MEDI project framework, the nf-neestimator pipeline contributes to the genomic characterisation workstream, specifically addressing: phylogenetic relationships and genomic characterisation of African small ruminant and chicken breeds; assessment of domestication processes and selection footprints; and comparative analysis of selection signatures in livestock breeds. Ne estimates will be integrated with complementary population genomic analyses to provide a comprehensive genomic profile of SCALA-MEDI target breeds, informing both conservation status assessments and the design of sustainable breeding programmes.

1.7 Online Resources

nf-neestimator GitHub repository: https://github.com/cnr-ibba/nf-neestimator

NeEstimator v2.X source code: https://github.com/bunop/NeEstimator2.X

PLINK v1.9 documentation: https://www.cog-genomics.org/plink/1.9/

PGDSpider — Population Genetics Data Conversion: http://www.cmpg.unibe.ch/software/PGDSpider/

Pipeline walkthrough video tutorial: https://youtu.be/xdLljTKFad0

Co-funded by the European Union

2. Report on the use of Reaction Norms for the Study of G×E in Animal Breeding

2.1 Background

SCALA-MEDI project aims at optimizing the sustainable use and conservation of local genetic resources (Chicken, Sheep) from the Mediterranean region. Phenotypic plasticity refers to the ability of a genotype to produce different phenotypes under varying environmental conditions. Reaction norms provide a formal representation of this concept by describing the functional relationship between phenotypic expression and an environmental gradient.

2.2 Objectives

The objectives of this work were to design and implement a computational Reaction Norm Simulation tool capable of modelling genotype performance across continuous environmental gradients. The tool was developed to quantify genotype-specific environmental sensitivity through the estimation of slope parameters and to provide graphical outputs that facilitate the interpretation of phenotypic plasticity patterns. In addition, the framework was designed to ensure methodological reproducibility and to align with the research priorities of the PRIMA programme, supporting the study of environmental adaptation in livestock populations.

2.3 Methods

- **Environmental Gradient Definition**: An environmental gradient was defined as a continuous covariate representing systematically varying environmental conditions (e.g., temperature, water availability, management intensity). The gradient is standardized to facilitate slope interpretation.
- **Genotype Parameterization**: Each genotype was assigned a unique intercept (baseline performance) and slope (environmental sensitivity coefficient). These parameters were sampled to represent stable, moderately plastic, and highly plastic genotypes.
- **Simulation Framework:** Phenotypic observations were generated according to the linear reaction norm model. Random noise was introduced to reflect biological and measurement variability. The simulation was implemented in a reproducible computational environment.

- **Visualization and Estimation**: Reaction norms were visualized through genotype-specific regression lines across the environmental gradient. Slope estimates were extracted and compared to evaluate relative plasticity and performance stability.

2.4 Results

The simulation successfully generated distinct reaction norm patterns corresponding to predefined slope parameters. Genotypes with low slope coefficients exhibited environmental robustness, whereas genotypes with steeper slopes demonstrated higher environmental sensitivity. The graphical outputs clearly illustrate crossover interactions under selected parameter configurations, confirming the presence of genotype re-ranking across environments.

The Reaction Norm Simulation framework aligns with PRIMA objectives related to climate resilience, sustainable agriculture, and environmental adaptation in the Mediterranean region. Modeling genotype performance across environmental gradients is essential for identifying climate-resilient varieties and adaptive management strategies.

The developed tool supports capacity building, methodological harmonization, and evidence-based decision-making within multi-environment research platforms.

Below a detailed description of the workflow and an example of its applicability, as presented during a dedicated SCALA-MEDI training:

Data, context and software

The workflow operates on a simulated sheep dataset for post weaning growth rate (PWGR) pedigree data and 568 records collected within one year across 10 farms are used. The analysis is implemented in R, loading tidyverse/explore/knitr for data handling and reporting, and breedR/MCMCglmm/postMCMCglmm for mixed model and (conceptual) reaction norm steps.

Estimation of the environmental gradient

First, a pedigree based animal model is fitted with breedR::remlf90 (REML, AI): fixed effect = AGE; random effect = CG (Contemporary Groups); genetic = additive animal using the pedigree.

The CG solutions (BLUE/BLUPs as provided by the fit) are extracted and standardized (mean = 0, SD = 1) to define the environmental gradient (EG) later used in the reaction norm interpretation.

Reaction norm formulation and interpretation

The reaction norm is framed as a linear random regression on the standardized environmental gradient, yielding two EBVs per animal: an intercept EBV (genetic merit at the average environment) and a slope EBV (genetic sensitivity/robustness across the gradient). Although the HTML emphasizes the formulation and its interpretation rather than a full code listing for this step, the RNM is presented as the target model for G×E assessment.

Visualization and reproducibility

The workflow includes graphical exploration (e.g., PWGR density and CG boxplots) and visualization of the standardized CG based gradient (EG) using ggplot2, with code snippets provided to reproduce every step—from data loading through model fitting and gradient construction—so that the same pipeline can be transferred to other datasets and settings.

Note. CG = Contemporary Groups

Training

A dedicated training session on reaction-norm analysis and genotype by environment (G×E) modelling was organized to provide participants with a hands-on, reproducible workflow (data, code, and step by step guidance). The session focused on: (i) constructing an environmental gradient from Contemporary Groups (CG), (ii) fitting the animal model to obtain CG solutions, and (iii) interpreting the reaction norm framework (intercept and slope EBVs) within applied breeding scenarios.

All training materials, including R scripts, datasets, slides, and the step-by-step tutorial, are openly available:

- GitHub repository (source code and assets):

https://github.com/stebif68/ScalaMedi_ReactionNorm

- Web tutorial (rendered HTML guide):

https://stebif68.github.io/ScalaMedi_ReactionNorm/ReactionNorm_Simulation.html

These resources are intended for reuse in real world applications and can be directly adapted by substituting local datasets or redefining the environmental descriptor.

2.5 Conclusions

The Reaction Norm Simulation practical demonstrates how to build an environmental gradient and fit a reaction norm model to obtain intercept and slope EBVs. The same workflow can be reused for other traits and species by adapting the definition of the environmental gradient and the model specification.

References

Chang, C.C., Chow, C.C., Tellier, L.C.A.M., Vattikuti, S., Purcell, S.M., & Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience, 4(1), 7. doi:10.1186/s13742-015-0047-8

Di Tommaso, P., Chatzou, M., Floden, E.W., Barja, P.P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. Nature Biotechnology, 35(4), 316–319. doi:10.1038/nbt.3820

Do, C., Waples, R.S., Peel, D., Macbeth, G.M., Tillett, B.J., & Ovenden, J.R. (2014). NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. Molecular Ecology Resources, 14(1), 209–214. doi:10.1111/1755-0998.12157

Hill, W.G. (1981). Estimation of effective population size from data on linkage disequilibrium. Genetical Research, 38(3), 209–216.

Lischer, H.E.L., & Excoffier, L. (2012). PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. Bioinformatics, 28(2), 298–299. doi:10.1093/bioinformatics/btr642

Sargolzaei, M., & Schenkel, F.S. (2009). QMSim: a large-scale genome simulator for livestock. Bioinformatics, 25(5), 680–681. doi:10.1093/bioinformatics/btp045

Waples, R.S. (2006). A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. Conservation Genetics, 7(2), 167–184.

Waples, R.S., & Do, C. (2010). Linkage disequilibrium estimates of contemporary *Ne* using highly variable genetic markers. Evolutionary Applications, 3(3), 244–262.

Manunza, A., Cozzi, P., Boettcher, P., Curik, I., Looft, C., Colli, L., Sölkner, J., Mészáros, G., & Stella, A. (2025). Estimating the optimal number of samples to determine the effective population size in livestock. Frontiers in Genetics, 16. doi:10.3389/fgene.2025.1588986