

Weighting the evidence: allele frequency strategies for improving LD-based N_e estimation

Battilani, D.¹, Waples, R.S.²

¹Research and Innovation Centre, Fondazione Edmund Mach, Italy
²School of Aquatic and Fishery Sciences, University of Washington, USA

CONTACTS

- Email: danielle.battilani@fmach.it
- Linkedin: Daniele Battilani
- Bluesky: danielle-battilani.bsky.social

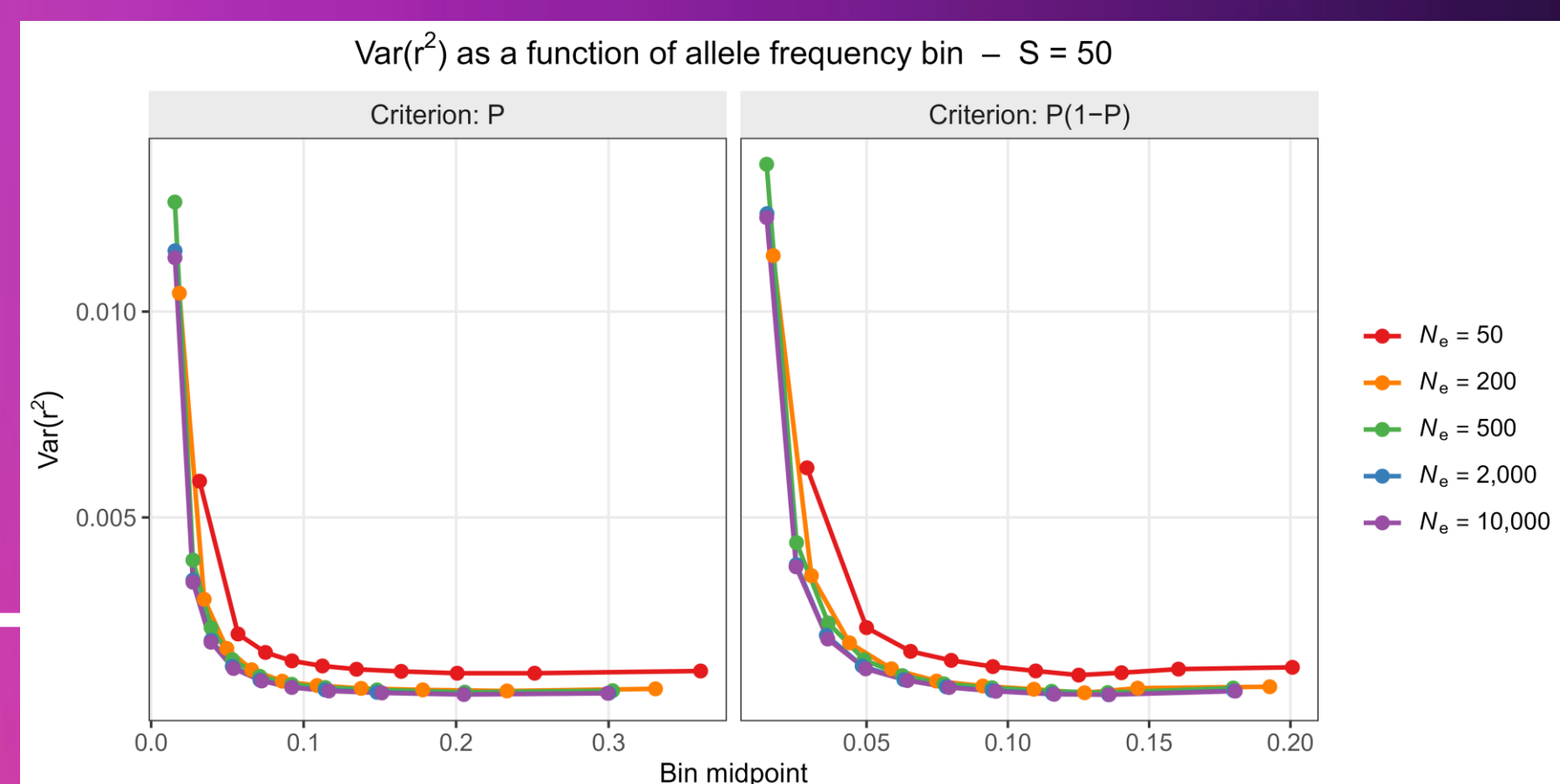
INTRODUCTION

The linkage disequilibrium (LD) method is the most widely used single-sample approach for estimating contemporary effective population size (N_e). Genomic datasets now routinely yield 10^5 – 10^6 SNPs, far more than any N_e estimator requires, raising the question of whether selecting or weighting loci by allele frequency (P) can improve estimates. Sved et al. (2013) proposed that intermediate-frequency alleles carry more reliable information about drift-generated LD, and suggested weighting r^2 by P; Waples (2024) proposed weighting by $P(1-P)$ instead. The merits of these proposals had never been systematically tested.

MATERIALS & METHODS

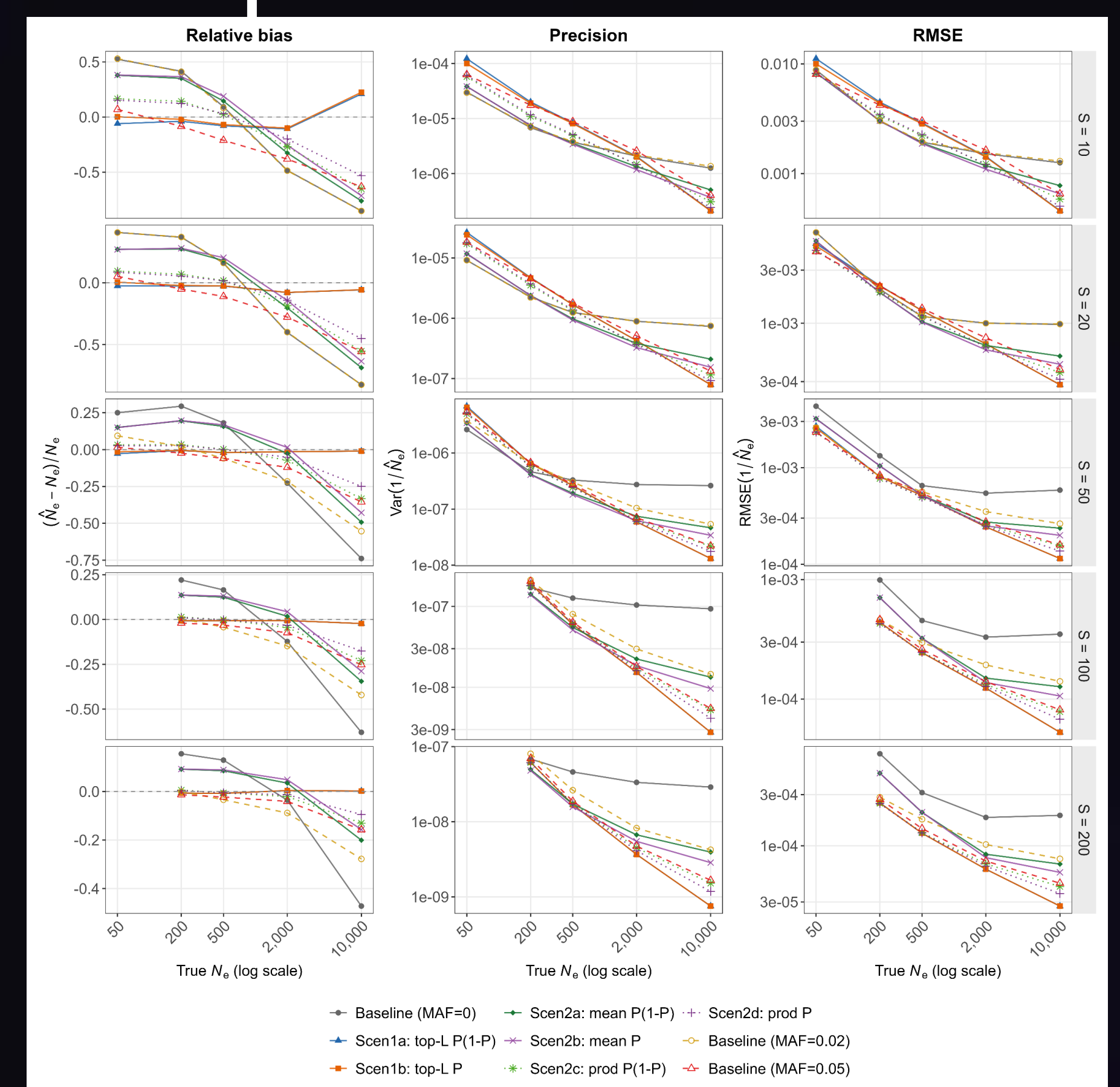
- Individual-based Wright-Fisher simulations, constant population size $N = N_e$, discrete generations, random mating, no selection/mutation/migration, diallelic and unlinked loci, initial allele frequencies from a L-shaped site-frequency spectrum.
- Full factorial grid: $N_e = \{50, 200, 500, 2,000, 10,000\} \times S = \{10, 20, 50, 100, 200\}$ ($S > N_e$ excluded); 100 replicates per cell.
- Primary design: $L = 10,000$ loci subsampled from a fixed pool of 100,000 (pool/L = 10); pool/L also varied {100, 50, 20, 10} using $L = \{1,000, 2,000, 5,000, 10,000\}$.
- Two families of allele-frequency strategies compared against unweighted baselines (MAF = 0, 0.02, 0.05):
 - Scen1 (selection): retain the L loci with highest $P(1-P)$ (Scen1a) or highest P (Scen1b);
 - Scen2 (pair weighting): mean r^2 weighted per pair by mean or product of $P(1-P)$ or P (Scen2a–d).
- Performance metrics: relative bias of N_e , precision ($\text{Var}(1/N_e)$) and RMSE($1/N_e$).
- Variance of r^2 recorded across allele-frequency bins to directly test Sved's premise.

RESULTS



- $\text{Var}(r^2)$ declined sharply and consistently with allele frequency, under both P and $P(1-P)$ criteria, across all N_e and S.
- At $S = 50$, the lowest-frequency bin showed 5–16× higher variance than the highest bin – directly confirming Sved's premise.

- Selection scenarios (Scen1a/b) were the most stable and least biased overall, but least precise at small N_e and S.
- Pair-weighting scenarios (Scen2) were competitive at small-to-intermediate N_e and S: mean weighting (Scen2a/b) gave the best precision/RMSE there, product weighting (Scen2c/d) gave the best balance across all three metrics.
- At large N_e and $S \geq 50$, Scen1a/b dominated, with up to 20× lower variance and 40–60% lower RMSE than the unfiltered baseline.
- MAF filtering remained competitive, often optimal, only at the smallest N_e and S.



- Pair-weighting and baseline scenarios were insensitive to the pool/L ratio.
- The two selection criteria diverged sharply: Scen1b (top-P) developed strong positive bias at high pool/L (realistic for WGS datasets), while Scen1a (top- $P(1-P)$) stayed robust across all ratios, converging with Scen1b only at pool/L = 10.



- Bias: Scen1a wins across most of the $N_e \times S$ space.
- Precision: baselines win at small N_e/S , Scen2b at intermediate N_e , Scen1b at large N_e .
- RMSE: Scen1b prevails at $S \geq 50$ and $N_e \geq 200$; pair-weighting and baselines win the remaining small- N_e/S cells.

DISCUSSION

Allele-frequency weighting can meaningfully improve LD-based N_e estimation, but no single scheme is universally optimal. Selecting loci by $P(1-P)$ is the most broadly robust strategy, particularly for genome-wide datasets where pool/L ratios are large, because unlike selection by P it does not require the analyst to track or correct for that ratio. Pair-level product weighting offers the best all-round balance at small-to-intermediate N_e and S, where selection-based approaches lose precision. Conventional MAF filtering remains a simple, competitive default when both N_e and S are small. All strategies can be viewed as generalizations of MAF filtering and the practical choice ultimately reduces to which performance metric matters most.