



HAL
open science

Inferring Population Size History from Large Samples of Genome-Wide Molecular Data - An Approximate Bayesian Computation Approach.

Simon Boitard, Willy Rodríguez, Flora Jay, Stefano Mona, Frederic Austerlitz

► To cite this version:

Simon Boitard, Willy Rodríguez, Flora Jay, Stefano Mona, Frederic Austerlitz. Inferring Population Size History from Large Samples of Genome-Wide Molecular Data - An Approximate Bayesian Computation Approach.. conférence Jaques Monod "Coalescence des approches théoriques et expérimentales en génomique évolutive et biologie des systèmes", 2016, Roscoff, France. hal-01604024

HAL Id: hal-01604024

<https://hal.science/hal-01604024v1>

Submitted on 5 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Inferring Population Size History from Large Samples of Genome-Wide Molecular Data - An Approximate Bayesian Computation Approach

Simon Boitard¹, Willy Rodríguez², Flora Jay³, Stefano Mona⁴, Frédéric Austerlitz⁵

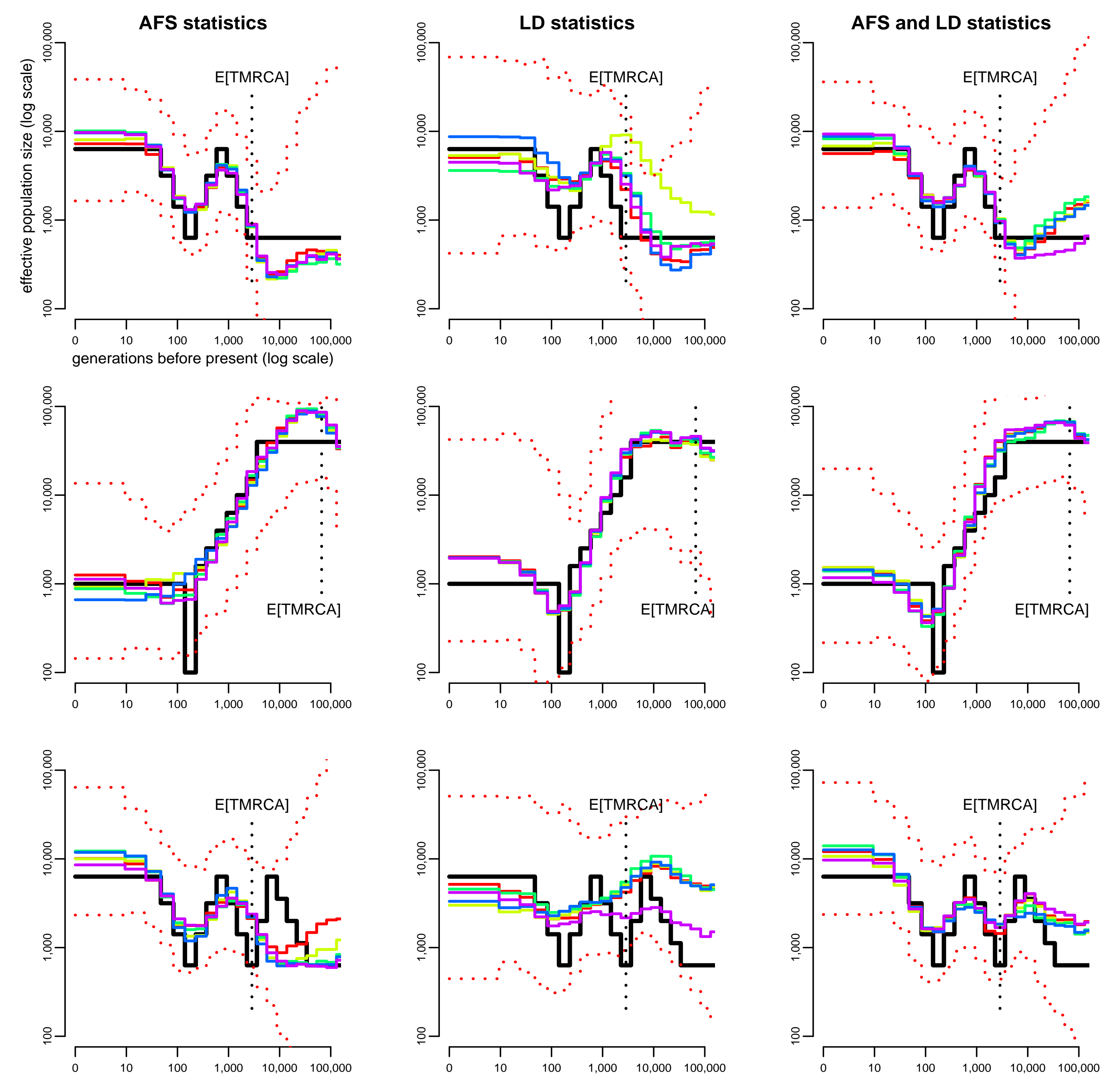
(1) GenPhySE, Castanet-Tolosan, France. (2) IMT, Toulouse, France. (3) LRI, Orsay, France. (4) ISYEB, Paris, France. (5) UMR 7206, Paris, France.

ABC ESTIMATION APPROACH

- **Model** : Kingman's coalescent with mutation and recombination, piecewise constant population size with 21 fixed time windows.
- **Simulated data** : 450,000 samples of n haploid genomes, one genome = 100 2Mb-long "contigs". Population sizes log-uniformly distributed from 10 to 100,000, with correlation between consecutive windows.
- **Summary statistics** : Folded Allele Frequency Spectrum (AFS) and average Linkage Disequilibrium (LD) for 18 bins of physical distance between SNP (from 300bp to 1.4Mbp).
- **ABC analysis** : 0.5% samples accepted, neuralnet regression (Blum and François, 2010) from *abc* R package (Csilléry *et al*, 2012).

COMPLEMENTARITY OF AFS AND LD STATISTICS

Depending on the scenario, LD or AFS is more informative.

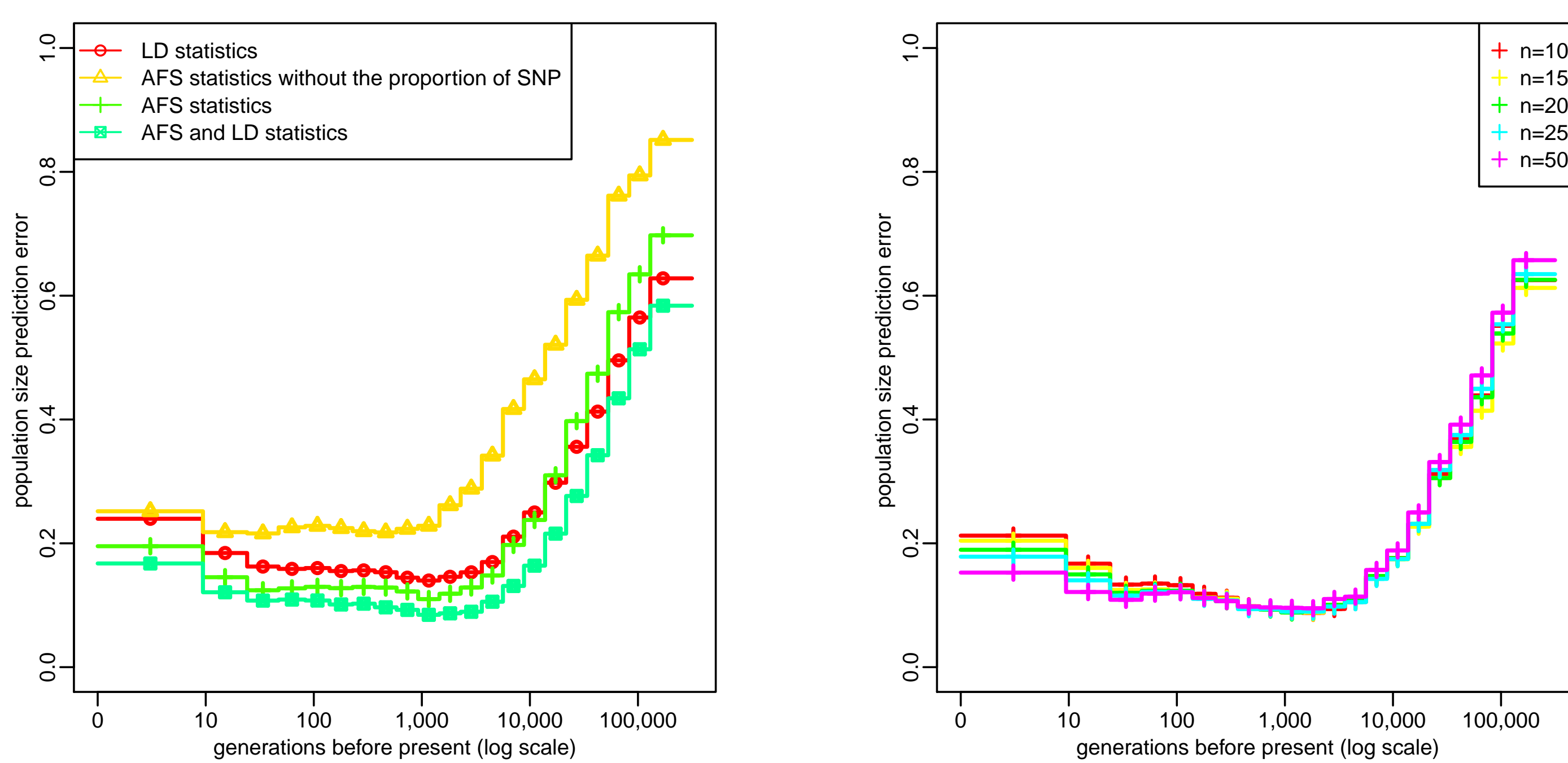


PREDICTION ERROR (RANDOM HISTORIES)

Estimation of demographic history improved by :

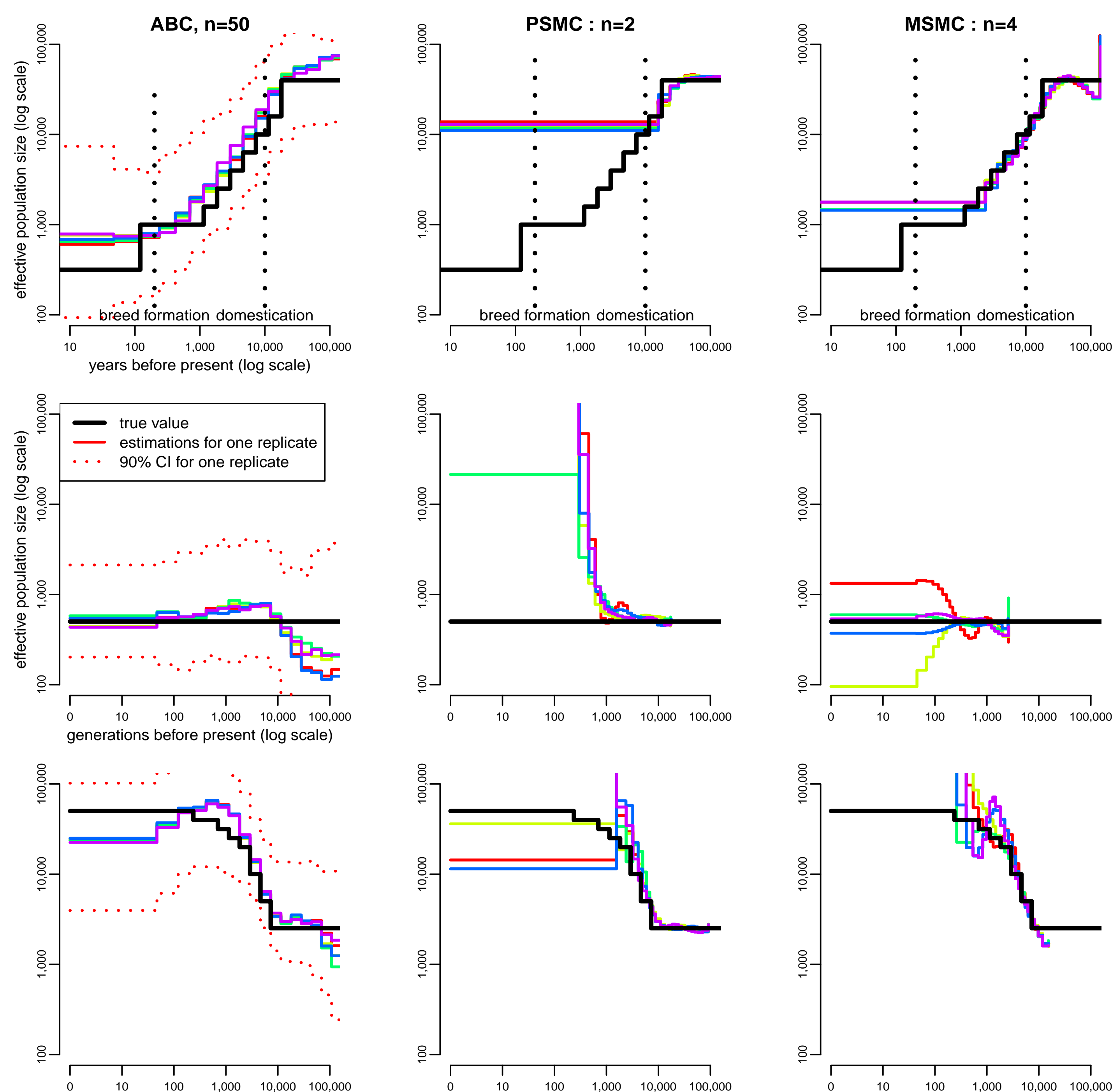
(i) combining LD and AFS statistics

(ii) increasing sample size



ESTIMATION FOR SPECIFIC HISTORIES

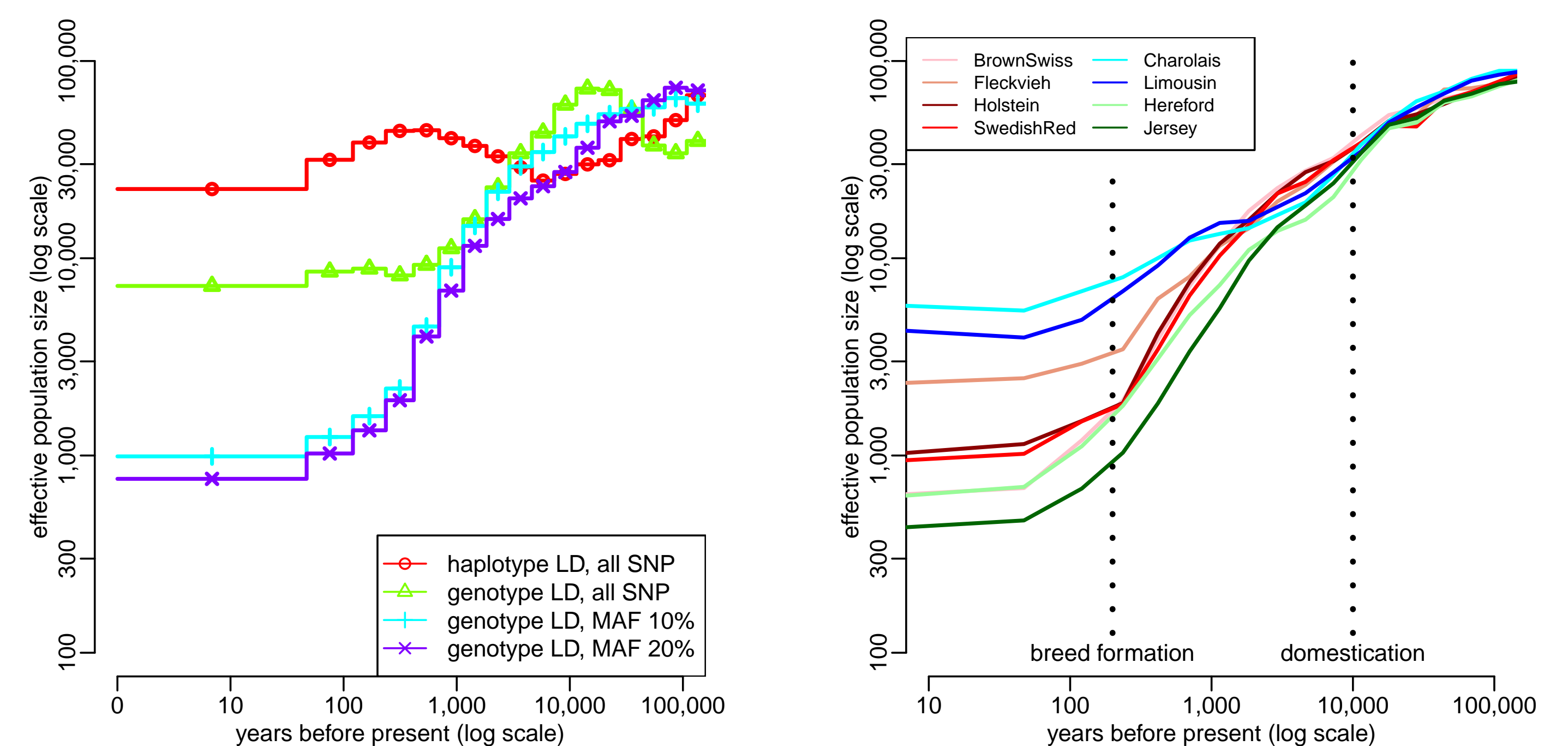
Recent demography recovered and credible intervals, in contrast to PSMC (Li and Durbin, 2011) and MSMC (Schiffels and Durbin, 2014).



APPLICATION TO NGS DATA IN CATTLE

Use genotype LD because of phasing errors and common alleles because of sequencing errors.

After domestication, continuous decline and clustering of breed histories by area of origin.



CONCLUSIONS

- Accurate estimation of population size history, including the recent past, through the use of large samples of genomes and the combination of AFS and LD information.
- Based on unphased data and robust to sequencing errors.
- Ref PLoS Genet 12(3): e1005877, source code at <https://forge-dga.jouy.inra.fr/projects/popsizabc/>
- Perspectives : simulate longer genomes through the use of faster simulation software (*msprime*, Kelleher *et al*, 2016), decoupling of observed times and change times.