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

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

Simon Boitard[1], Willy Rodríguez[2], Flora Jay[3], Stefano Mona[4], Frédéric Austerlitz[5]

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

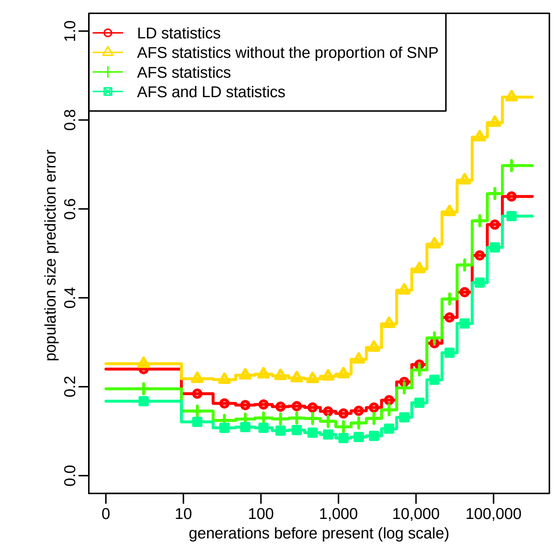
ABC Estimation approach

- **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).

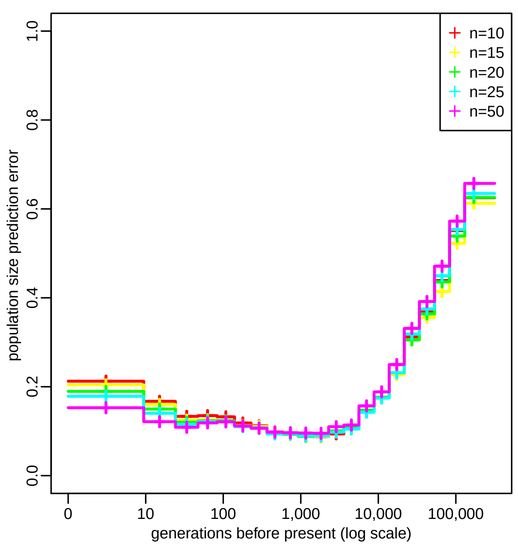
Prediction error (random histories)

Estimation of demographic history **improved by** :

(i) **combining LD and AFS** statistics

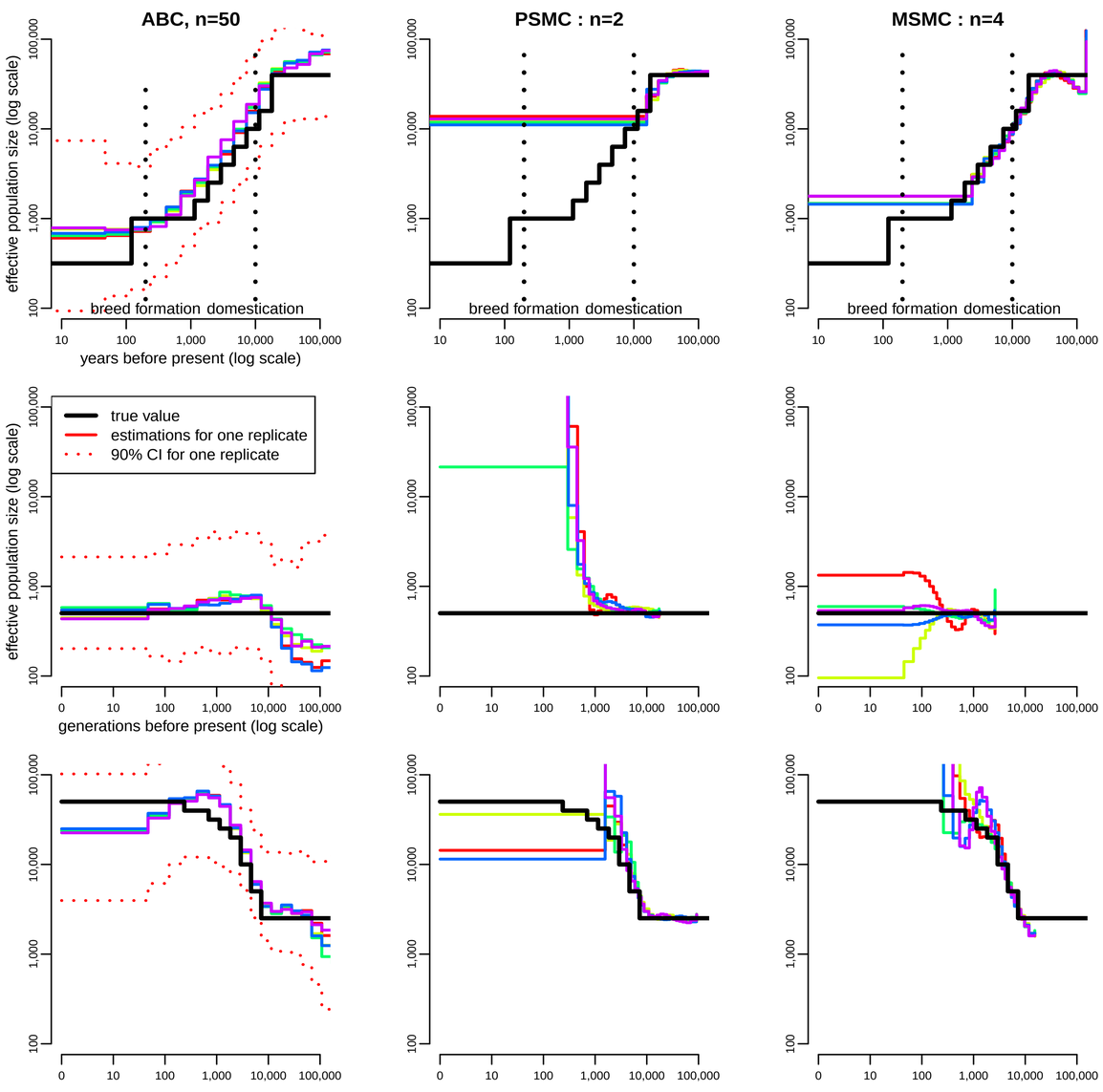


(ii) **increasing sample size**



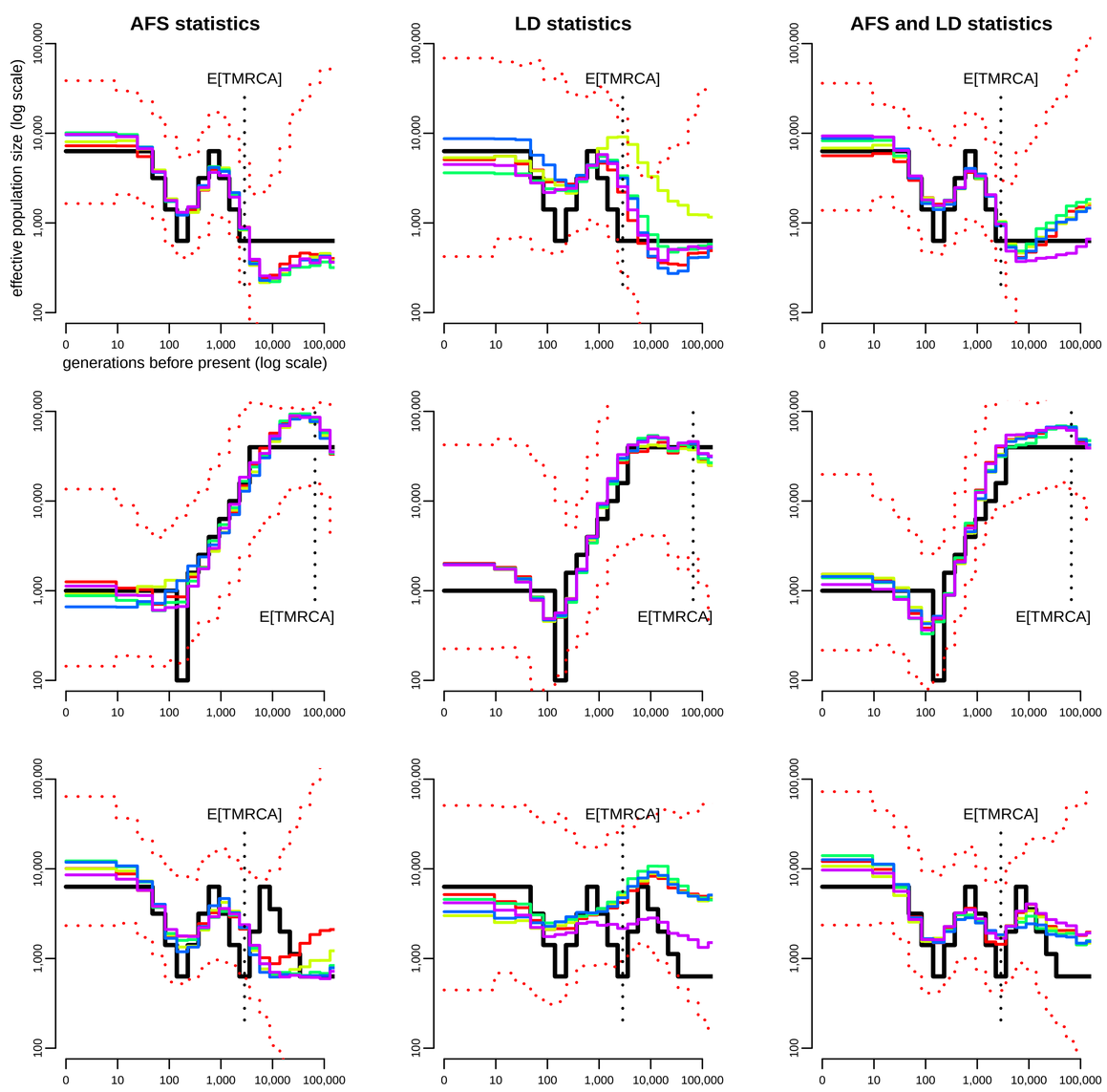
Estimation for specific histories

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



Complementarity of AFS and LD statistics

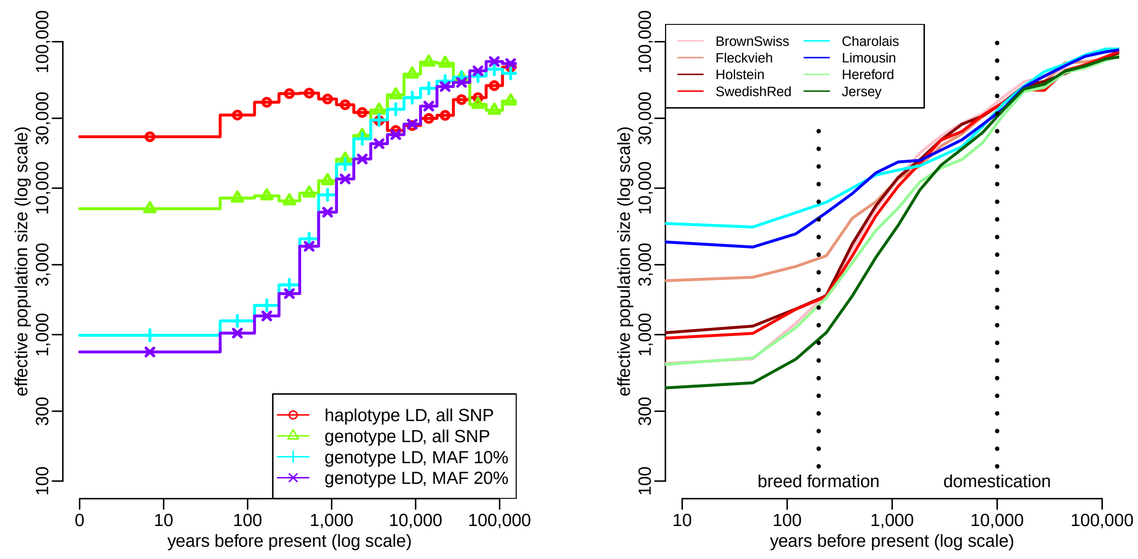
Depending on the scenario, **LD or AFS** is **more informative**.



Application to NGS data in cattle

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

After domestication, **contiuous 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/popsizeabc/
- **Perspectives :** simulate longer genomes through the **use of faster simulation** software (*msprime*, Kelleher *et al*, 2016), decoupling of observed times and change times.