



HAL
open science

Inferring the population size history of 4 cattle breeds using NGS data and an approximate Bayesian computation approach

Simon Boitard

► **To cite this version:**

Simon Boitard. Inferring the population size history of 4 cattle breeds using NGS data and an approximate Bayesian computation approach. 23. Plant and Animal Genome (PAG), Jan 2015, San Diego, United States. hal-01194066

HAL Id: hal-01194066

<https://hal.science/hal-01194066>

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 the population size history of 4 cattle breeds using NGS data and an approximate Bayesian computation approach



Simon Boitard^{1,2} (1) UMR 7205 Institut de Systématique Evolution et Biodiversité, Paris, France. (2) UMR 1313 Génétique Animale et Biologie Intégrative, Jouy en Josas, France.

SUMMARY

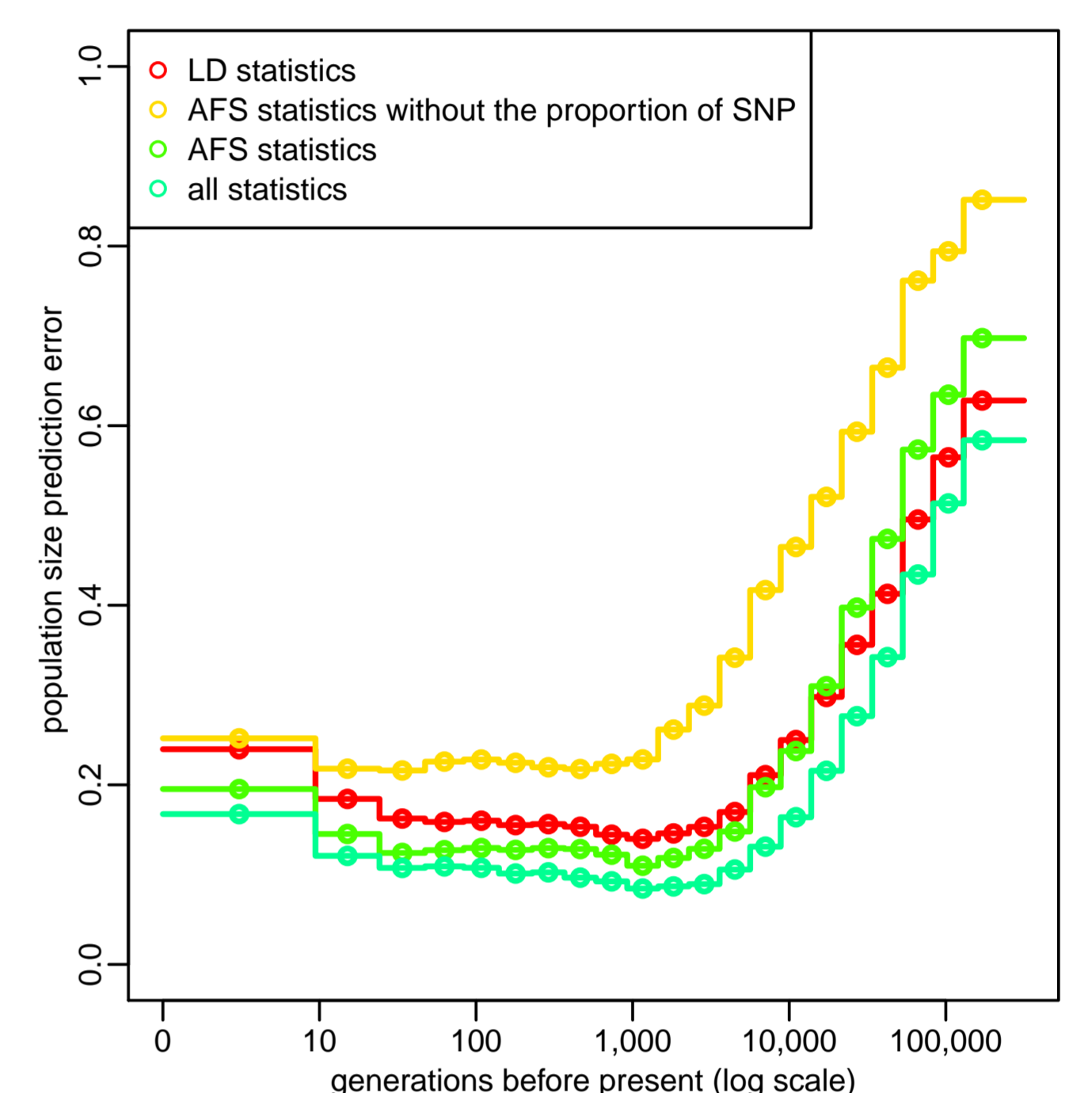
I present here a new Approximate Bayesian Computation (ABC) approach allowing to **estimate the ancestral dynamics of effective size in a population** from a **sample of unphased diploid genomes** from this population. This approach is based on summarizing the observed genomic sample using a set of statistics related to the **Allele Frequency Spectrum (AFS)** and the decay of **Linkage Disequilibrium (LD)** with physical distance. Using simulations, I show that this approach provides **accurate estimations of population sizes from present time up to at least 20,000 generations before present**. I then apply it to large genomic samples in 4 cattle breeds, which reveals interesting aspects of cattle history, from domestication to modern breed creation.

ESTIMATION APPROACH

The history of effective size in a population is estimated using the following ABC [1] approach :

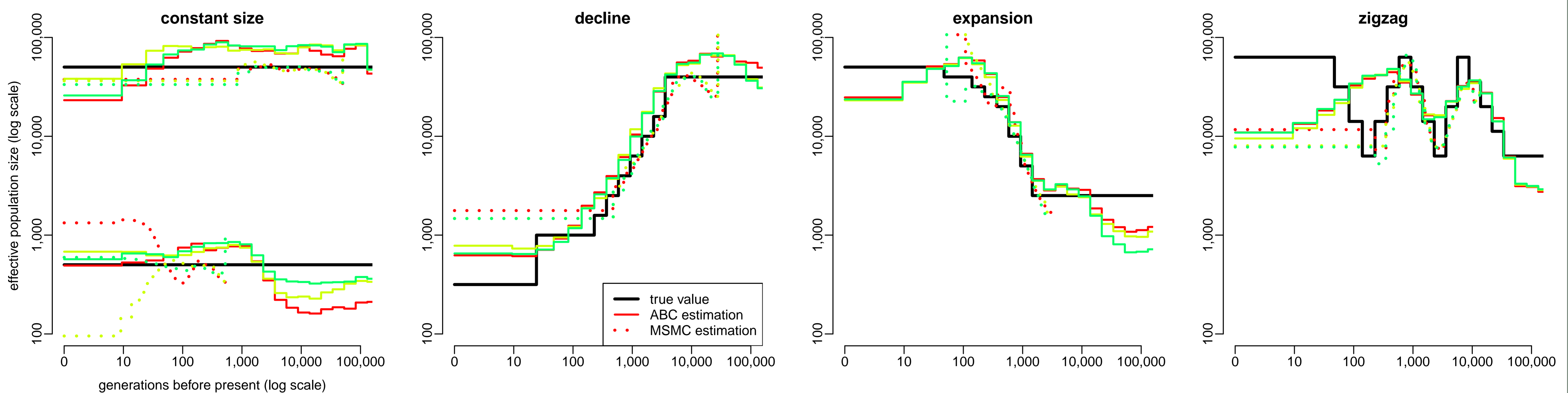
- The observed sample of n genomes is summarized using the following statistics : the overall proportion of polymorphic sites over the genome (1 statistic), the relative proportion of these polymorphic sites carrying i copies of the minor allele, for i from 1 to $n/2$ ($n/2$ statistics), and the average zygotic r^2 for 18 bins of physical distance (from 300bp to 1.4Mbp) between SNP (18 statistics).
- 450,000 genomic samples of n diploid genomes are simulated under a coalescent model with mutation and recombination, assuming a piecewise constant population size with 21 fixed time windows. For each simulated sample, the recombination rate and the 21 population sizes are sampled from prior distributions, while the mutation rate is fixed to 10^{-8} .
- Simulated samples are summarized using the same statistics as the observed sample, and the best 0.5% of population size histories (i.e. those providing statistics that are the most similar to the statistics of the observed sample) are selected. Population size history is finally estimated from these selected histories, using a neural network regression approach and the *abc* R package [2].

PREDICTION ERROR



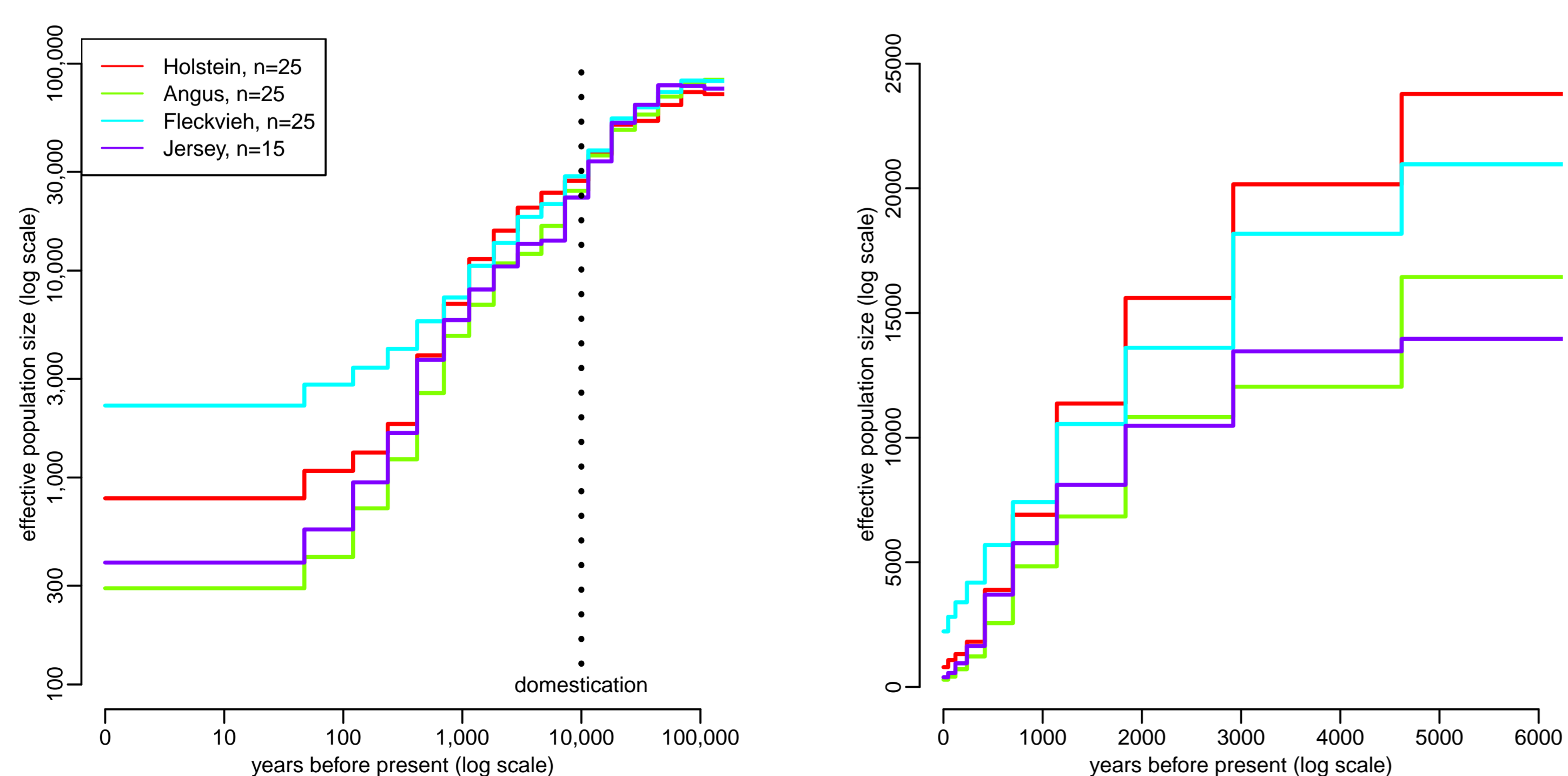
Prediction errors computed from 2,000 simulated samples ($n = 25$). Random histories sampled from the prior result in an error of 1.

ESTIMATED HISTORY FOR 5 DIFFERENT SIMULATED SCENARIOS



For each scenario, 3 simulated samples ($n = 25$) were analyzed with ABC or MSMC [3]. **Reconstruction of recent history is more accurate with ABC.**

ESTIMATED HISTORY IN 4 CATTLE BREEDS



Long term history (log scale)

Short term history (natural scale)

ACKNOWLEDGMENTS

Cattle genomes were obtained from the 1000 bull genomes project (www.1000bullgenomes.com). Data analyzes were performed on the computer cluster of the Genotoul bioinformatics platform Toulouse Midi-Pyrénées. I would like to thank Willy Rodriguez, Flora Jay, Stefano Mona and Frédéric Austerlitz for their help on this study.

REFERENCES

- [1] Mark A. Beaumont, Wenyang Zhang, and David J. Balding. Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.
- [2] Katalin Csilléry, Olivier François, and Michael G. B. Blum. abc: an r package for approximate bayesian computation (abc). *Methods in Ecology and Evolution*, 3(3):475–479, 2012.
- [3] S. Schiffels and R. Durbin. Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 46:919–925, 2014.