Inferring the ancestral dynamics of population size from genome wide molecular data - an ABC approach

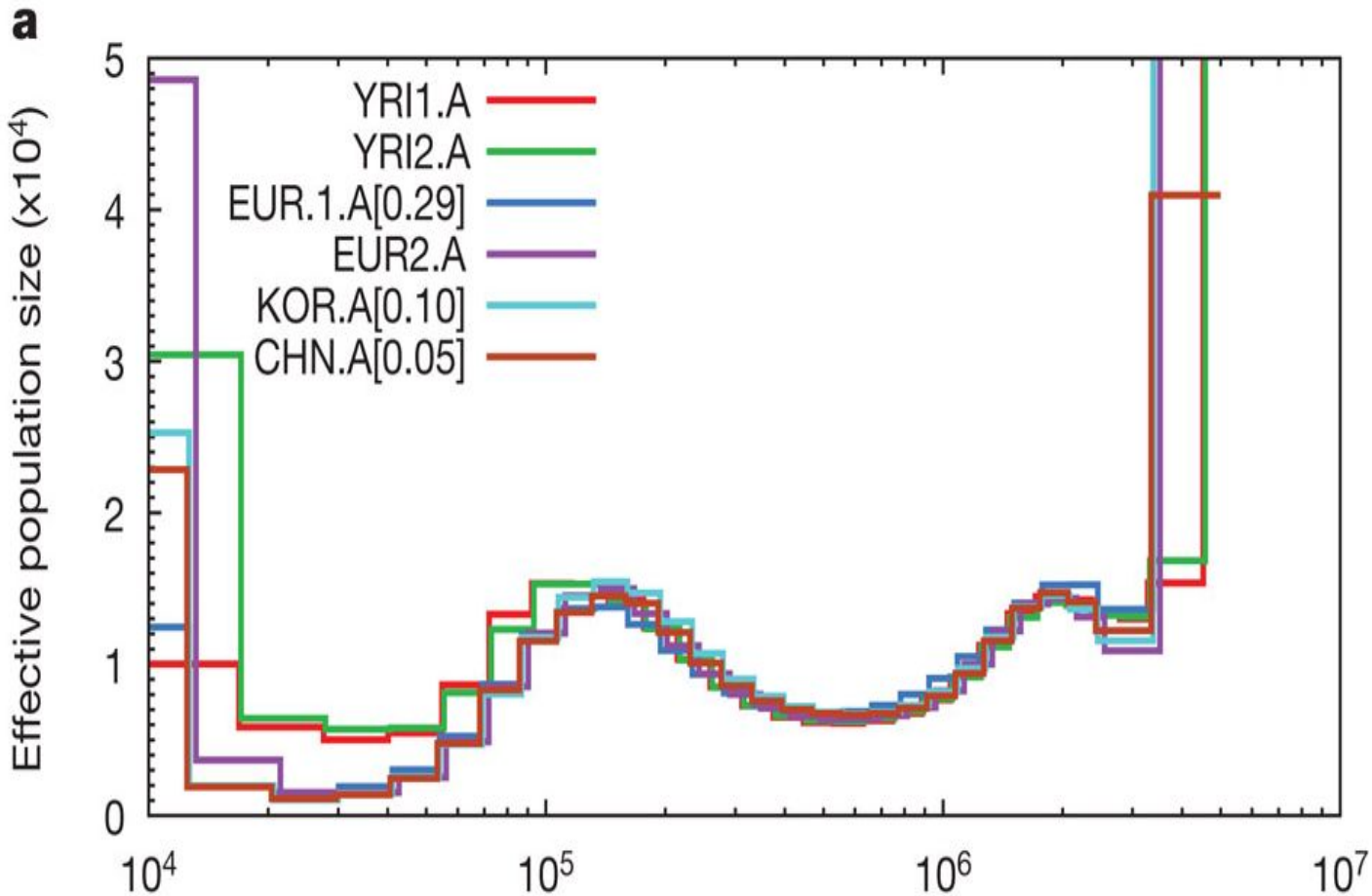
Simon Boitard

UMR 7205 ISYEB (EPHE - MNHN - CNRS - UPMC), Paris.
UMR 1313 GABI (INRA - AgroParisTech), Jouy en Josas

MCEB 2014

Motivation

Genome wide sequence data contains rich information about population size history, cf PSMC (Li and Durbin, 2011).



Development of an ABC approach

- Several estimation methods :
  - Sequentially Markovian Coalescent : PSMC (Li and Durbin, 2011), dical (Sheehan *et al*, 2013), MSMC (talk S. Schiffels).
  - MCMC : BEAST (Drummond *et al*, 2012)
  - Runs of Homozygocity : MacLeod *et al* (2013), Harris and Nielsen (2013).
  - ...
- But so far limited to small sample sizes ($n = 1$ to ≈ 5 diploid individuals) and / or genomic regions. $\rightarrow$ not appropriate for recent history estimation.
- ABC could take advantage of both genome wide data and large sample size.
- Little assumptions required concerning the underlying model.

Application to farm animal species

- Many genome sequences now available (pig, cattle, sheep, chicken), and a huge amount of animals with dense genotyping data.
- Several bottlenecks expected along their history :
  - Last glaciation : -25 000 – -60 000 years
  - Domestication : -10 000 years.
  - Creation of modern breeds and intensive selection : -200 years.
- Here $n = 25$ unrelated animals from the Holstein cattle breed (www.1000bullgenomes.com)

Outline

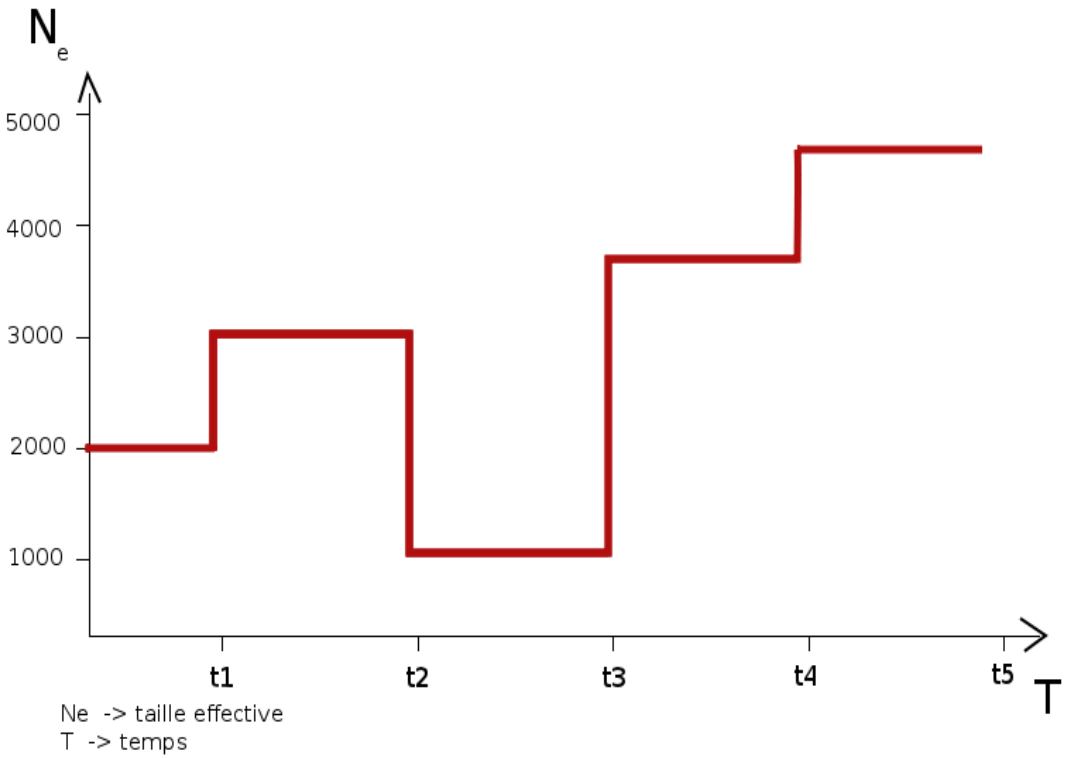
Outline

Principles of ABC (Approximate Bayesian Computation)

- To estimate the parameters θ of a model from a dataset $\mathcal{D}$, we approximate the posterior probability $\mathbb{P}(\theta|\mathcal{D})$ by the quantity $\mathbb{P}(\theta|\mathcal{S})$, for a set $\mathcal{S}$ of (meaningfull!) summary statistics.
- We estimate $\mathbb{P}(\theta|\mathcal{S})$ by simulations, with the following procedure :
  1. Compute $\mathcal{S} = f(\mathcal{D})$
  2. For i from 1 to I:
     1. Sample parameter θ_i from the prior distribution of θ.
     2. Simulate dataset $\mathcal{D}_i$ from the model with parameter θ_i.
     3. Compute $\mathcal{S}_i = f(\mathcal{D}_i)$.
     4. Select the simulation if $dist(\mathcal{S}_i, \mathcal{S}) < \epsilon$.
  3. Estimate the posterior distribution of θ from the selected θ_i values, by simple counting or other approaches (regression).

Model

- Coalescent with mutation and recombination, $2n = 50$ haplotypes.
- One single panmictic population (no structure).
- Piecewise constant effective population size, 21 fixed time windows with exponential size.

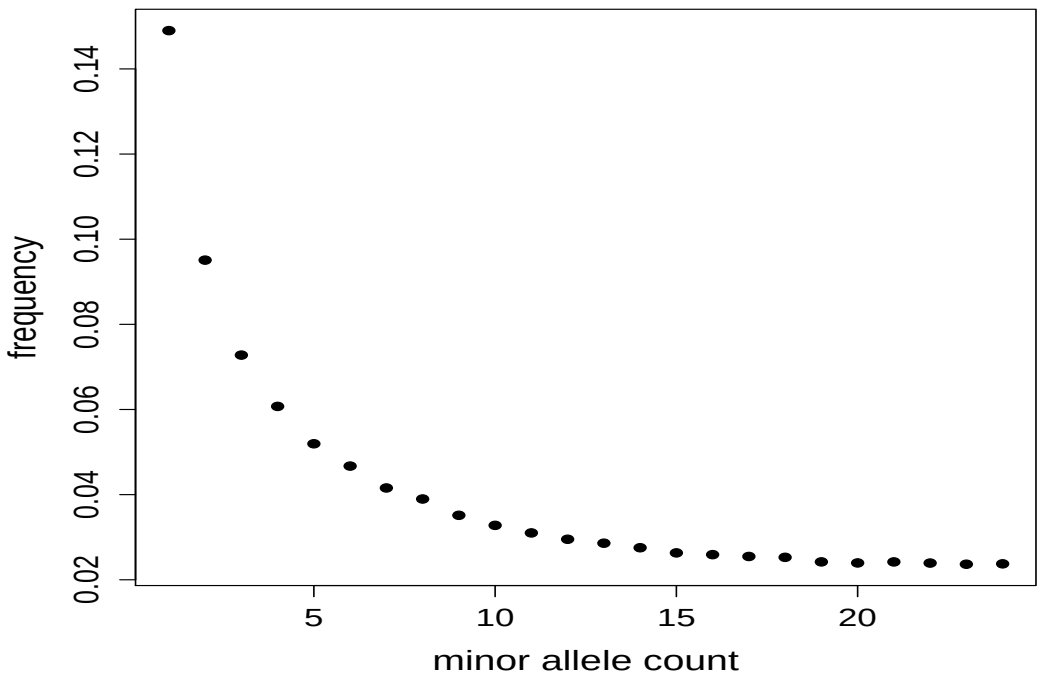


Prior distributions

- Per generation per bp mutation rate : $\mu = 1e-8$.
- Per generation per bp recombination rate : $r \sim \mathcal{U}(0.1e-8, 1e-8)$.
- Population size :
  - $\log(N_0) \sim \mathcal{U}(1,5)$.
  - $\log(N_{i+1}) = \log(N_i) + \alpha, \quad \alpha \sim \mathcal{U}(-1,1)$.
  - $1 \leq \log(N_i) \leq 5$.

Summary statistics - Allele Frequency Spectrum (AFS)

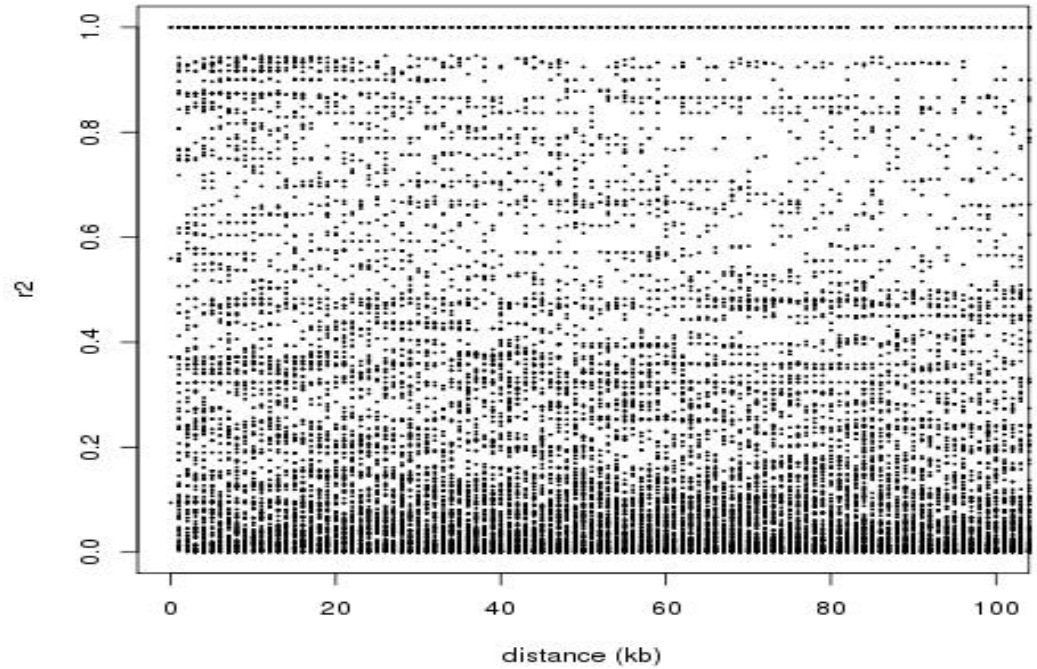
- Frequency of polymorphic sites over the genome.
- Frequency of sites with i copies of the minor allele, for i from 1 to n.



- Variance of these frequencies over the genome.

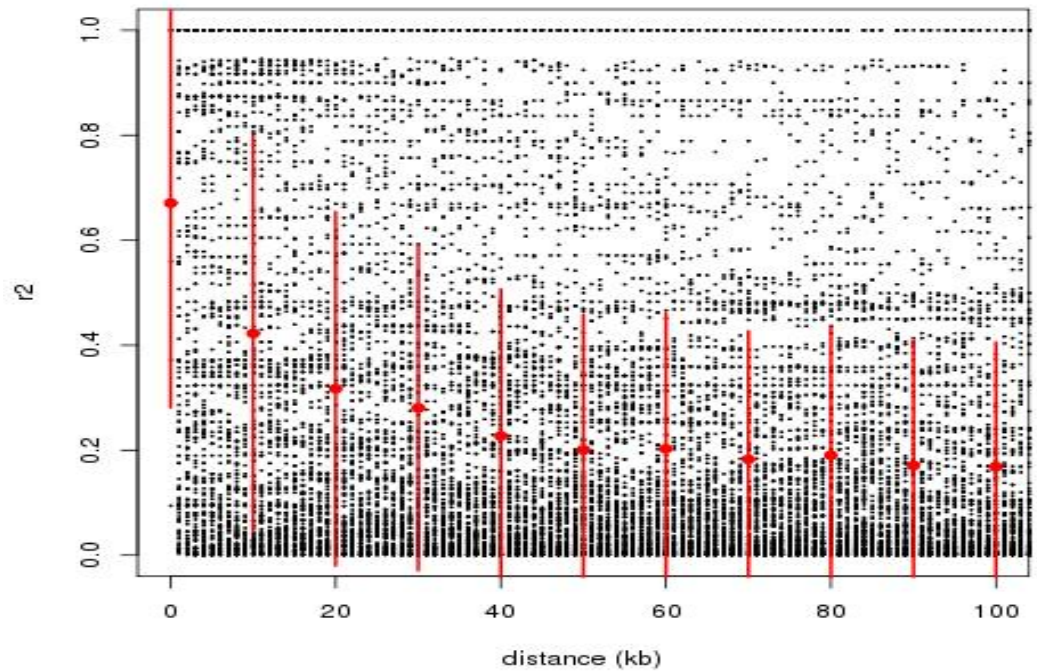
Summary statistics - Linkage Disequilibrium (LD)

- Correlation between genotype data at two polymorphic sites.



Summary statistics - Linkage Disequilibrium (LD)

- Correlation between genotype data at two polymorphic sites.



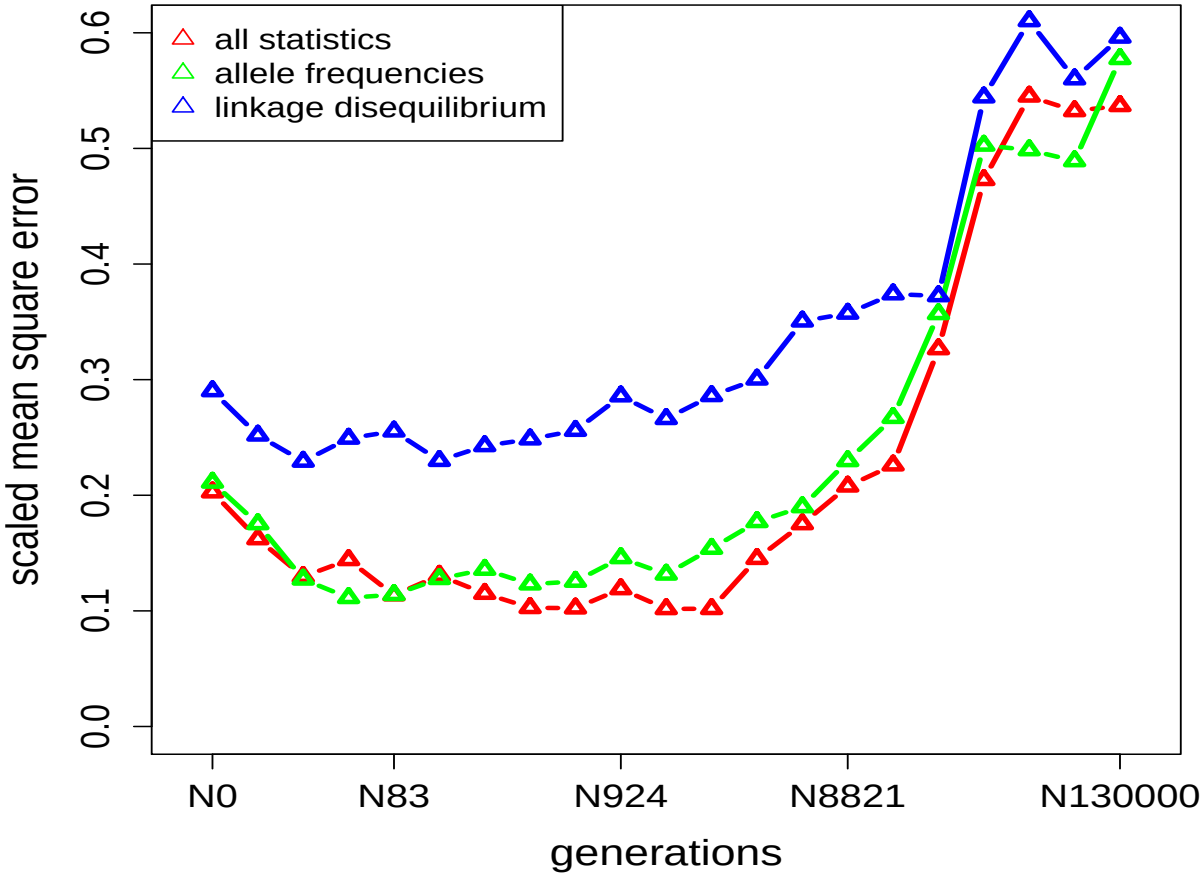
- Mean and variance of LD for several distances between sites.
- LD at distance d related to population size at time $t = \frac{1}{2c(d)}$.

Implementation

- Simulations :
  - Genotype data simulated with ms. One sample = 100 independent 2MB segments.
  - 500 000 simulated samples, $\approx$ one week on a cluster with 200 jobs in parallel.
- Holstein data :
  - Several pre-processing steps required to obtain genotype data (sequencing, alignment, genotype calling).
  - Genotype data processed with the same program.
- Final statistical analysis with the R package *abc*.

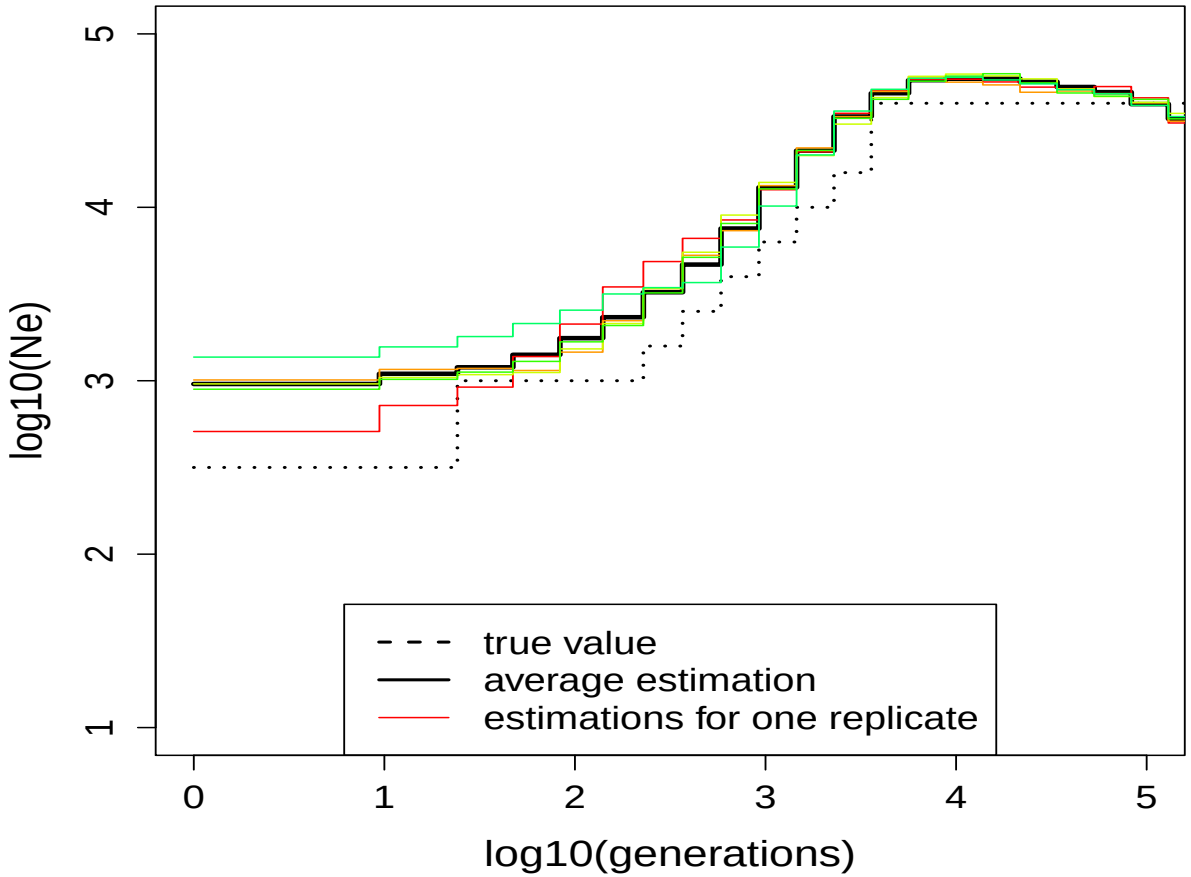
Outline

Cross validation



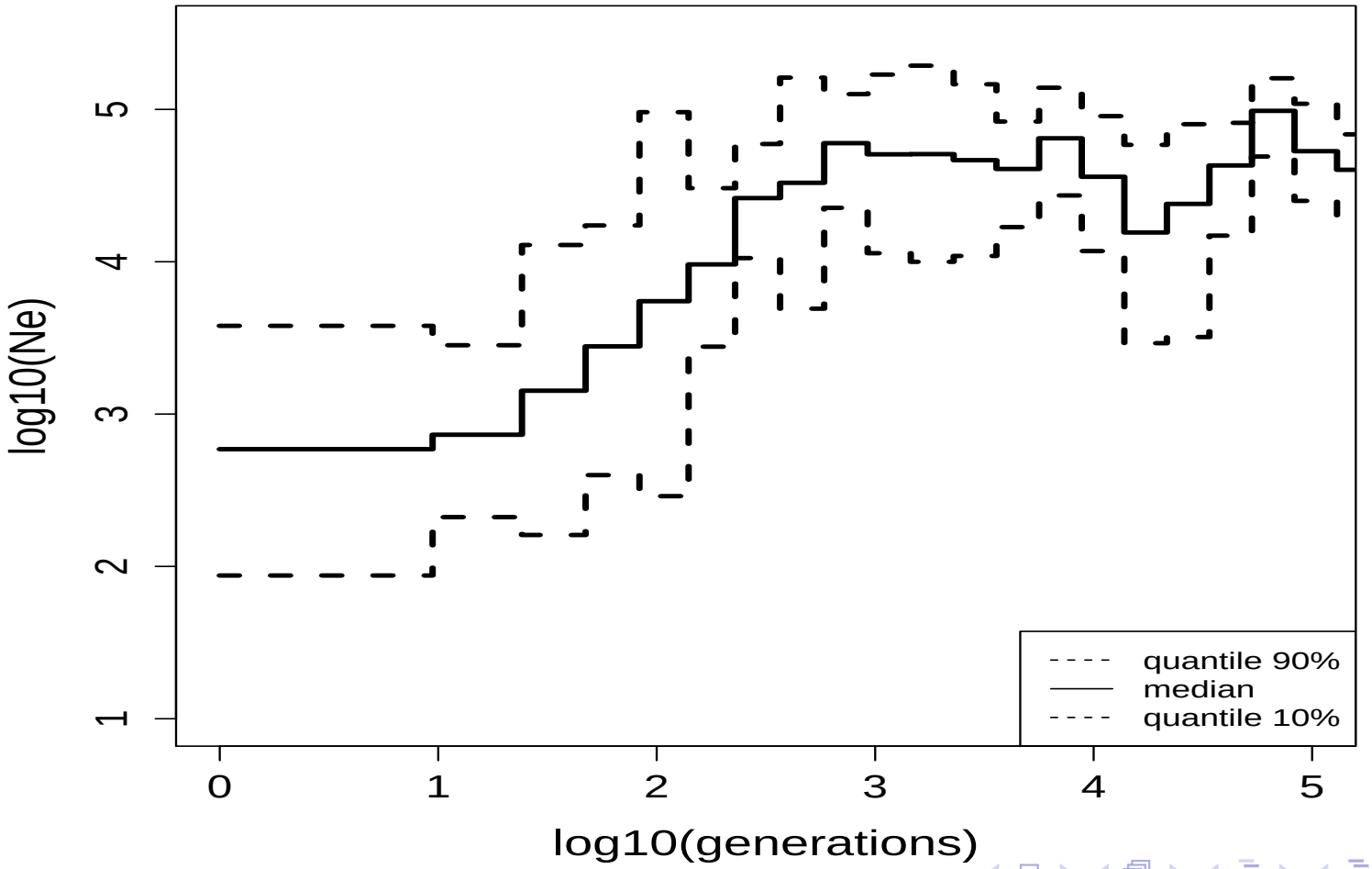
Estimation error $\frac{\sum_i (\theta_i - \hat{\theta}_i)^2}{I * Var(\theta_i)}$ based on 200 CV replicates.

Simulations under a bottleneck scenario (MacLeod *et al*, 2013)

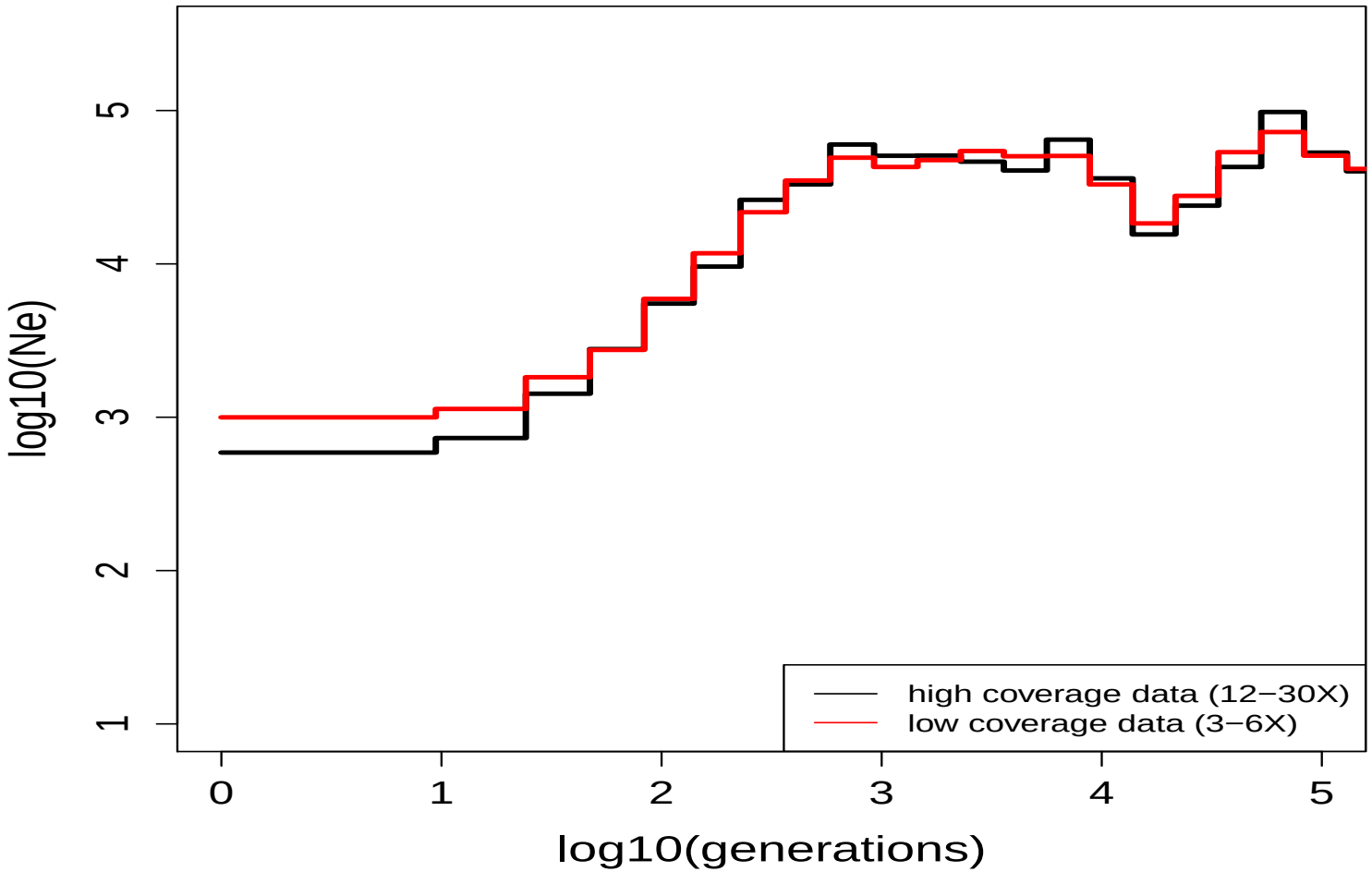


Outline

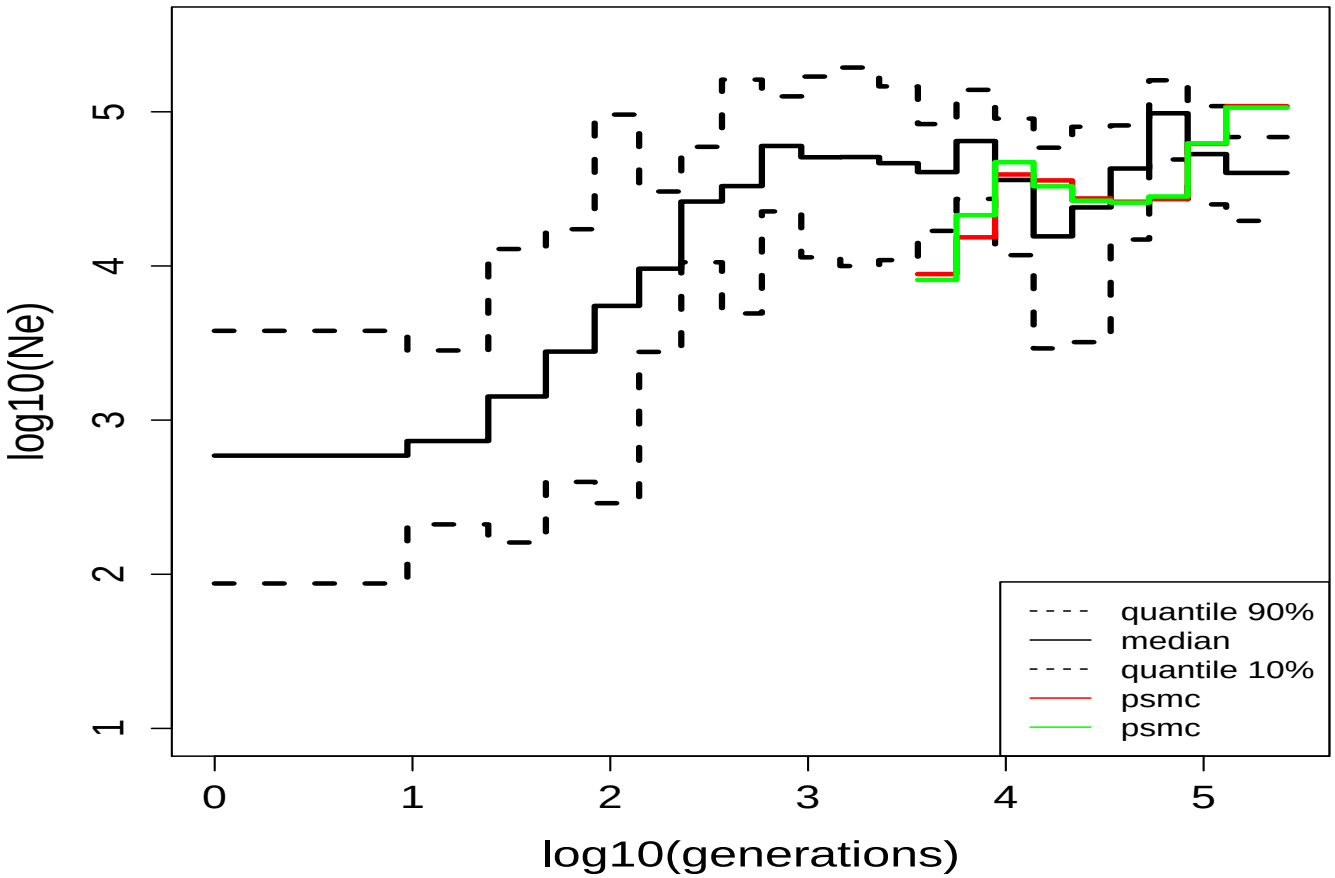
Estimated dynamics : strong bottleneck from $\approx$ 5 000 years b.p.



Little influence of sequencing errors



Comparison with PSMC



PSMC analysis by Willy Rodriguez.

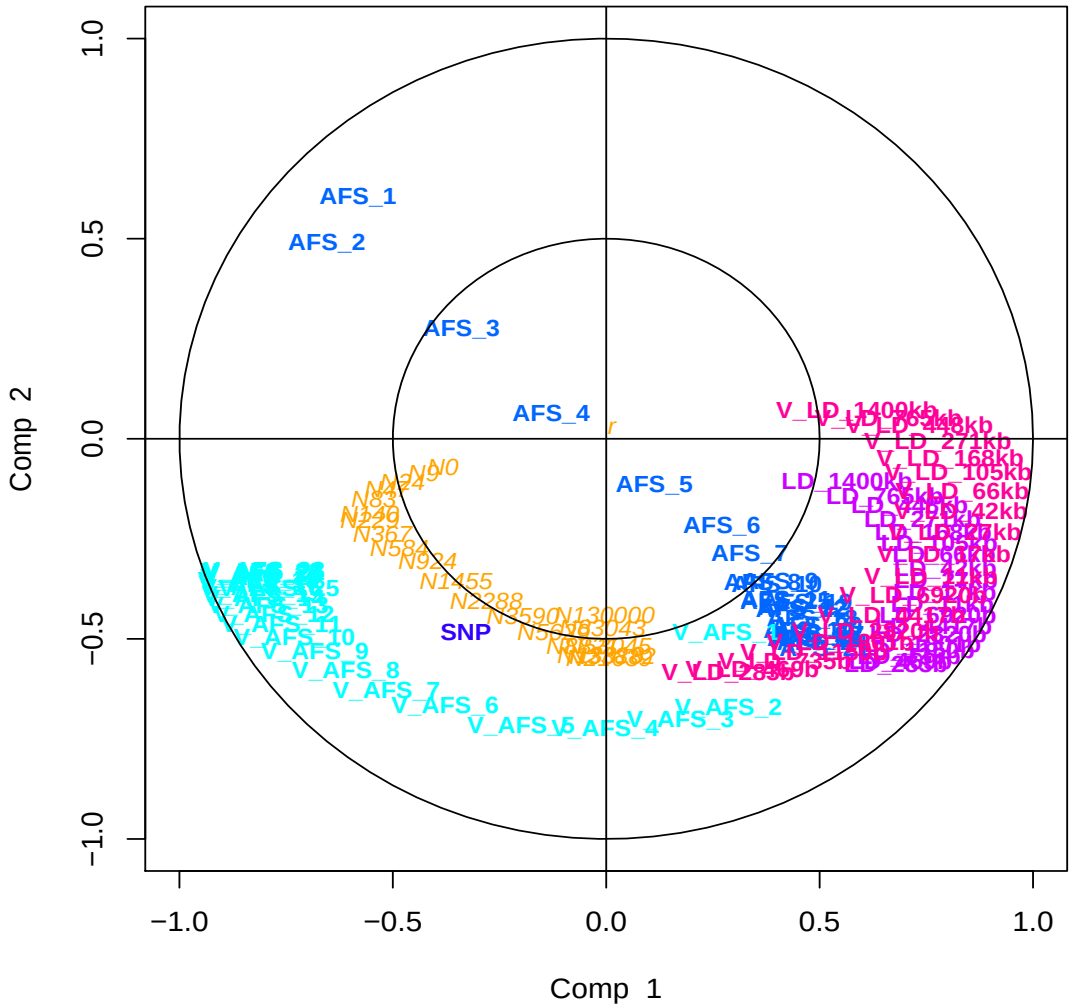
Conclusions and perspectives

- Methodology :
  - ABC provides accurate estimation of population size dynamics from 0 up to at least 20 000 generations b.p..
  - Combining AFS and LD is useful.
  - ABC can be applied to a wide range of data types : large sample size, unphased data, sparse sequencing, RAD sequencing ...
- Cattle history :
  - Population size dynamics estimated by ABC and other approaches are quite consistent, but population size started to strongly decrease more recently according to ABC.
  - More breeds will be analyzed.

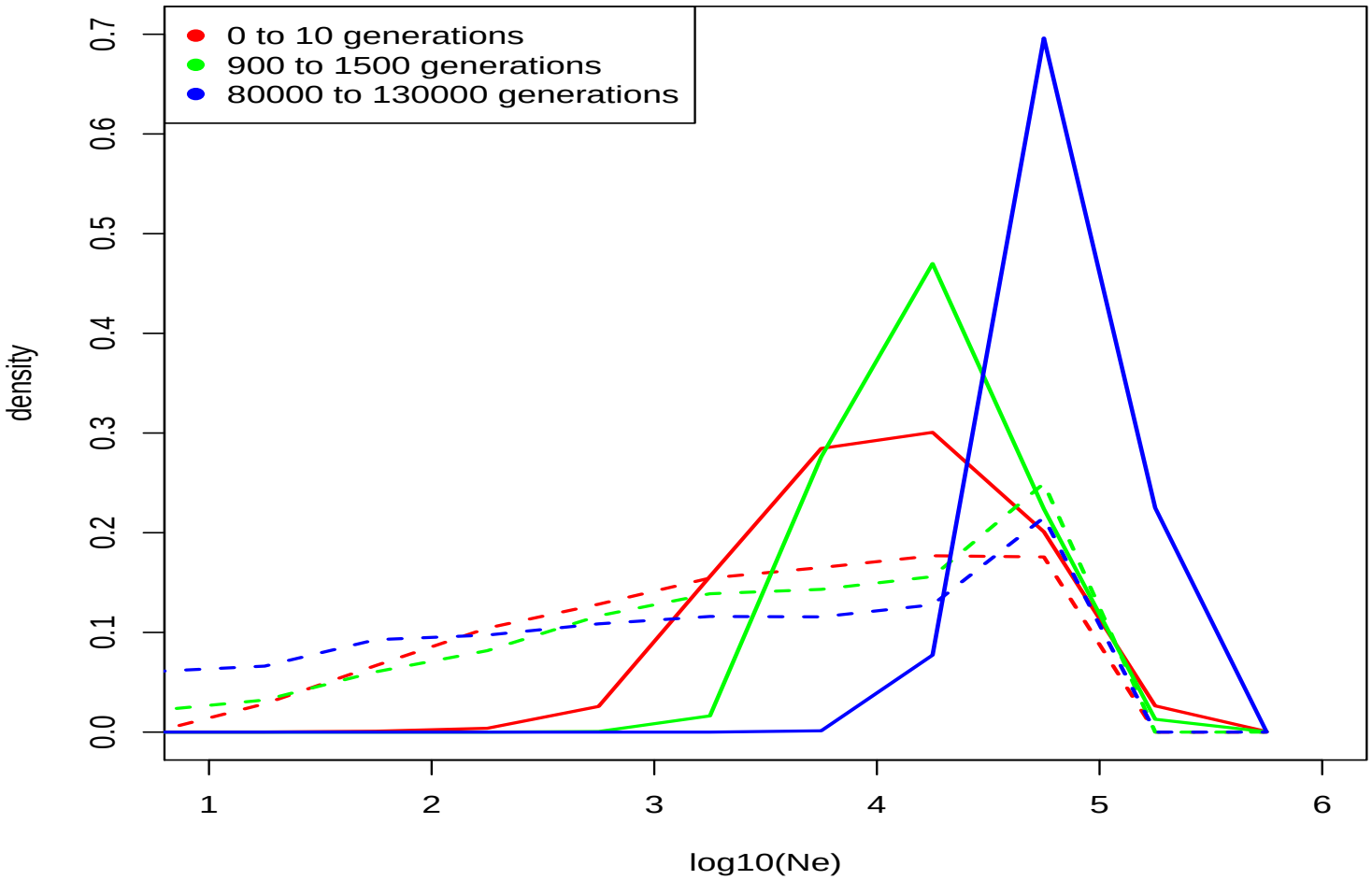
Acknowledgements

- Stanislas Sochacki (Ecole Polytechnique), Willy Rodriguez (INSA Toulouse).
- Lounes Chikhi (University Toulouse III), Olivier Mazet, Simona Grusea (INSA Toulouse).
- Bertrand Servin (INRA, Toulouse).
- 1000 bull genomes project.
- Genotoul Bioinformatics Platform
- ANR Demochips : Fréderic Austerlitz, Stefano Mona ...

Influence of AFS and LD statistics - PLS regression



Data is informative



Prior check