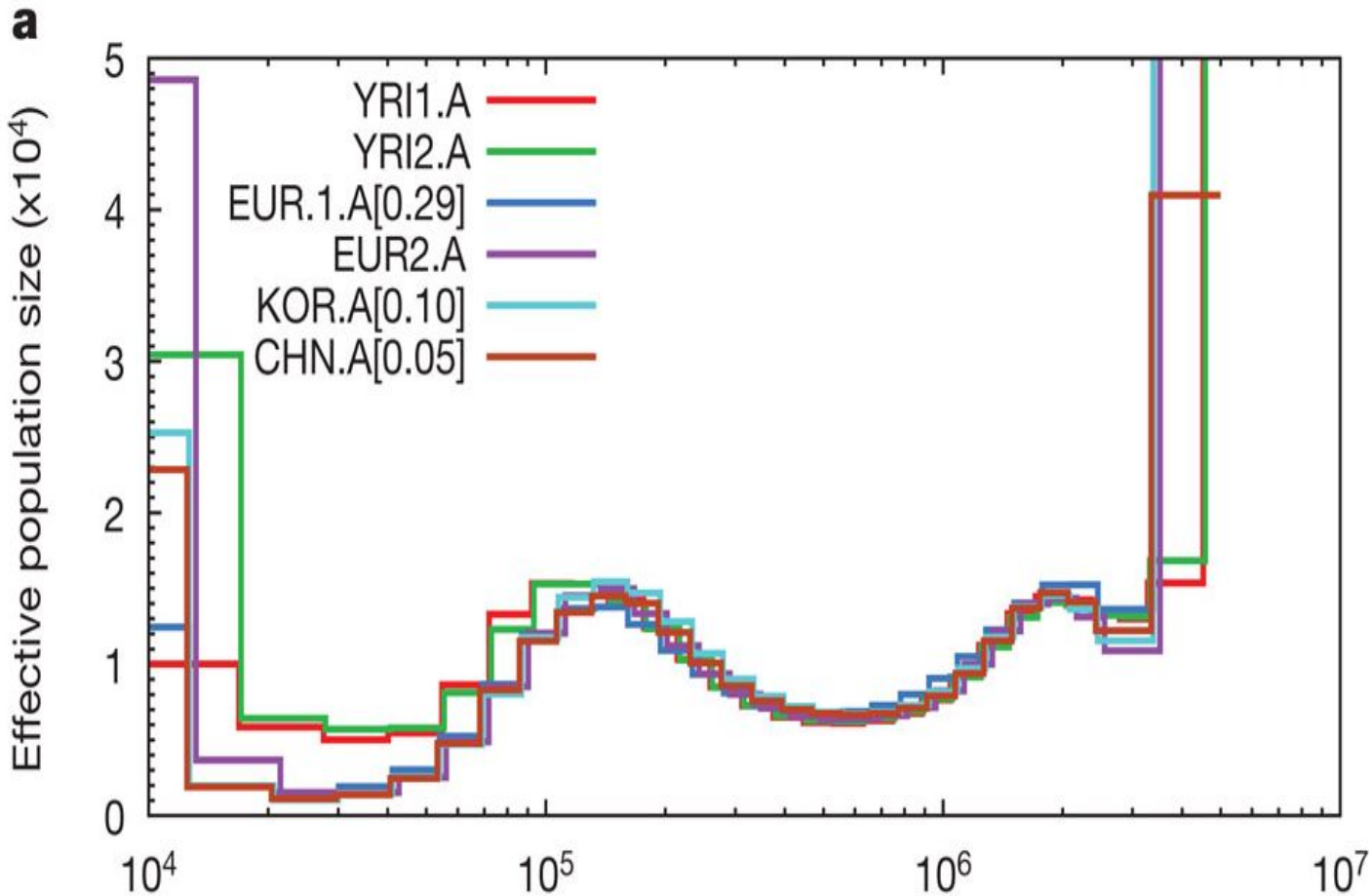
Inferring the ancestral dynamics of population size from genome wide molecular data - an ABC approach

Simon Boitard

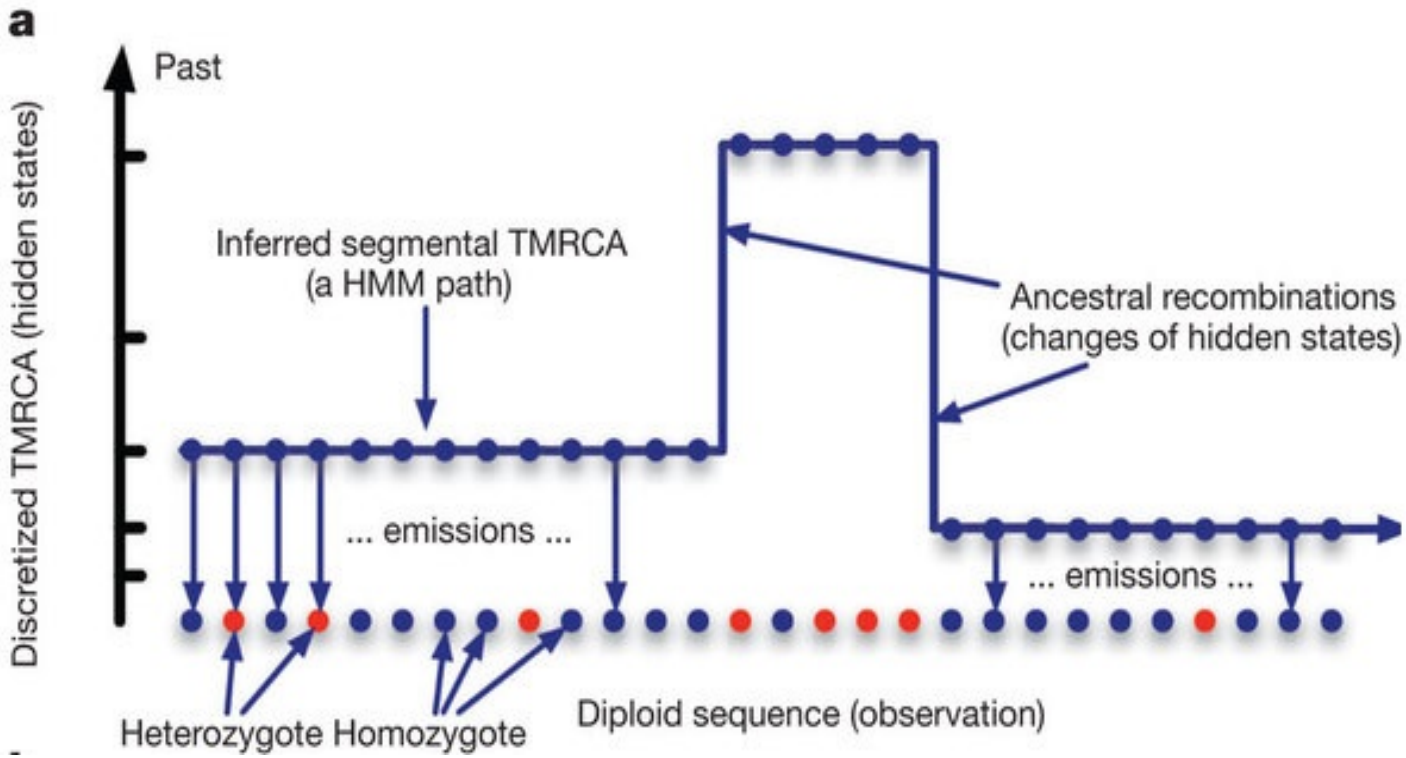
UMR 7205 OSEB (EPHE - MNHN - CNRS), Paris.
UMR 1313 GABI (INRA - AgroParisTech), Jouy en Josas

Motivation

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



Pairwise Sequentially Markovian Coalescent (PSMC)



- Markov chain for $T2$ based on the Sequentially Markovian Coalescent (SMC), transitions depend on $N(t)$.
- Estimation through an Hidden Markov Model (HMM).
- Limited to one individual ($n = 2$) $\rightarrow$ not efficient for recent times.

Development of an ABC approach

- Several estimation methods (Drummond *et al*, 2012; MacLeod *et al*, 2013; Sheehan *et al*, 2013), but limited to $n = 2$ or small genomic regions.
- ABC could take advantage of both genome wide data and large n.
- 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 25 unrelated animals ($n = 50$) from the Holstein cattle breed (www.1000bullgenomes.com)

Outline

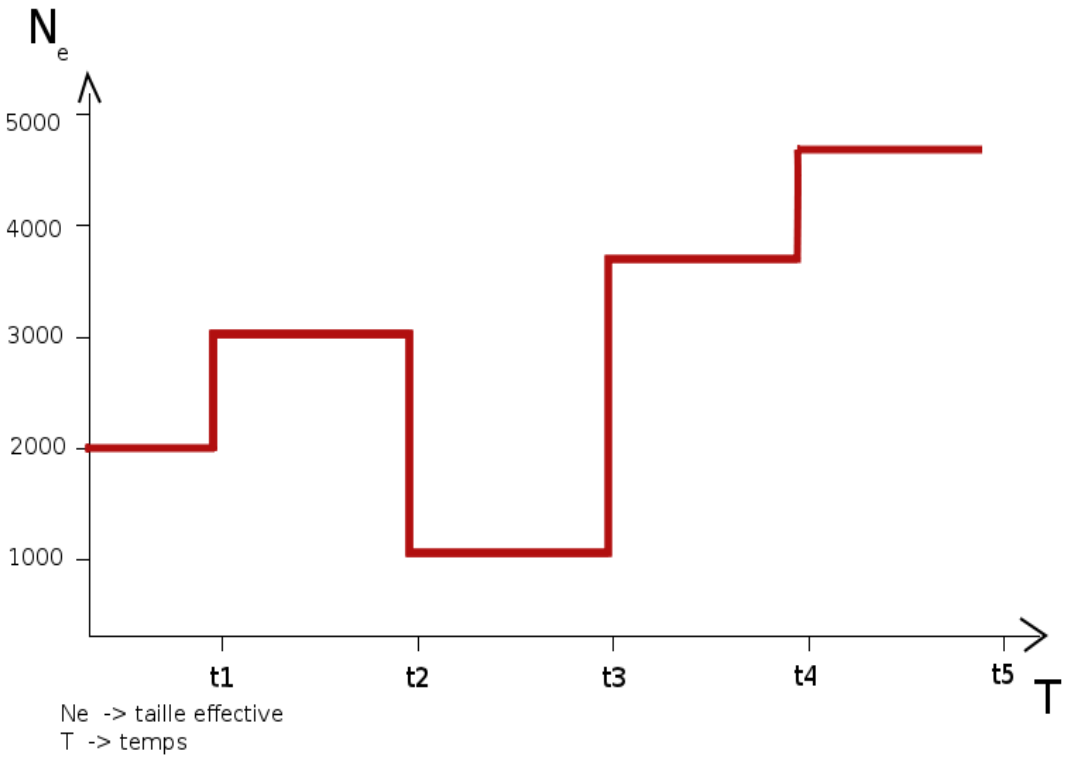
Outline

Principles of ABC (Approximate Bayesian Computation)

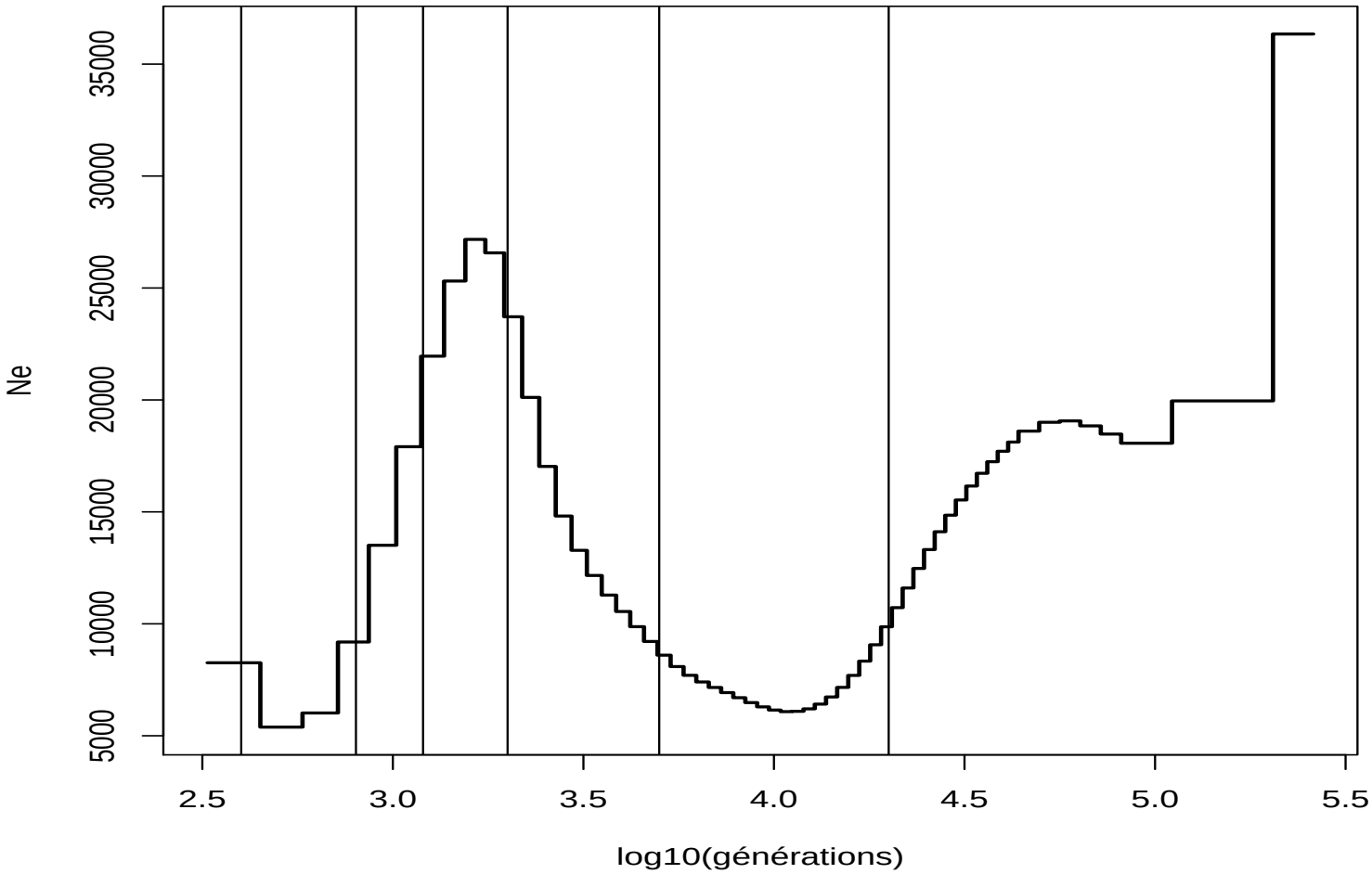
- To estimate the parameters θ of a model from a dataset $\mathcal{D}$, we approximate the posterior probability $\mathbb{P}(\theta|\mathcal{D})$ by the quantity $\mathbb{P}(\theta|\mathcal{S})$, for a set $\mathcal{S}$ of (meaningfull!) summary statistics.
- We estimate $\mathbb{P}(\theta|\mathcal{S})$ 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

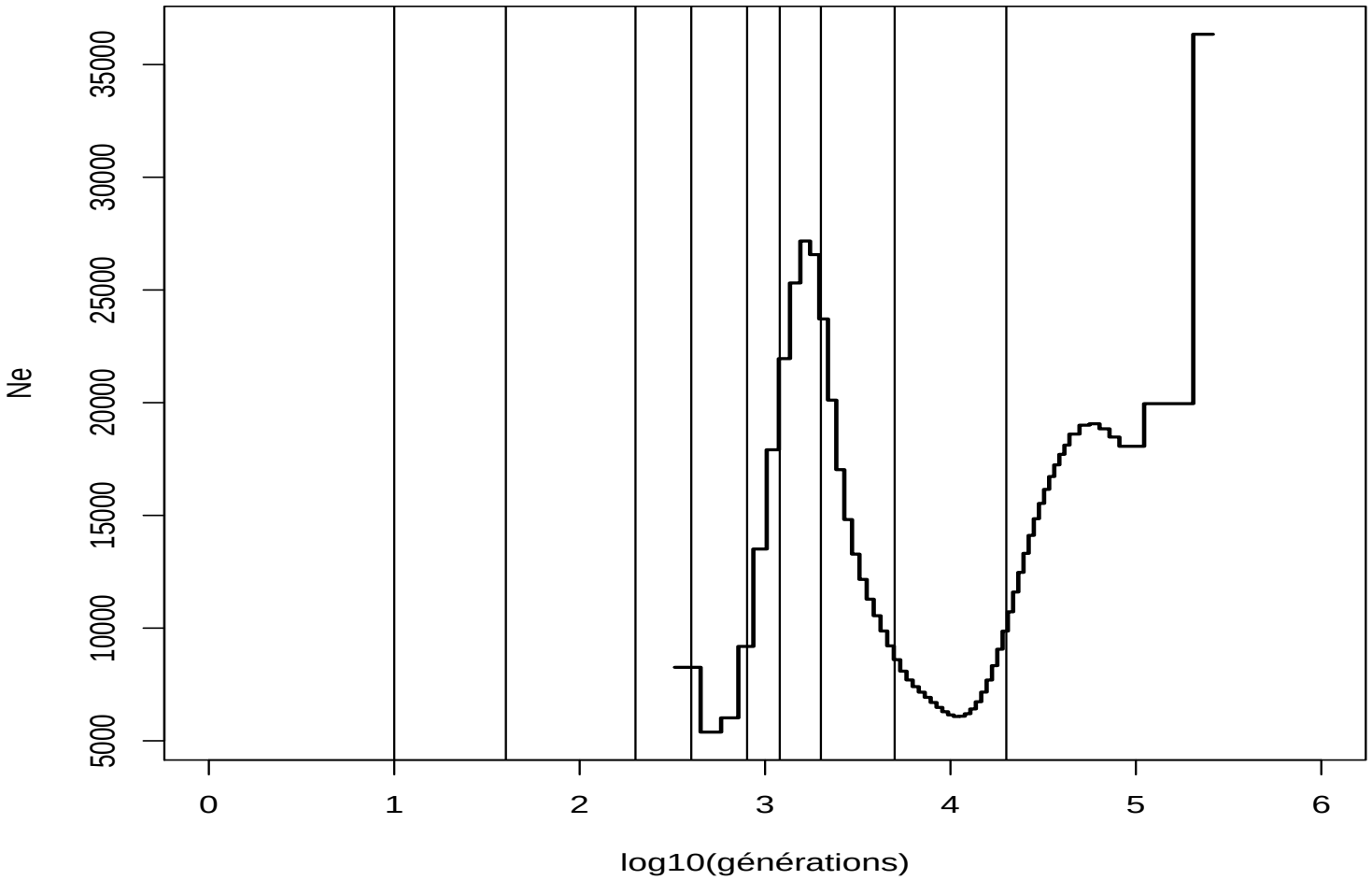
- Coalescent with mutation and recombinaison, $n = 50$ haplotypes.
- No structure.
- Piecewise constant effective population size.



Intervals are defined from a previous PSMC analysis ...



... as well as breeding history

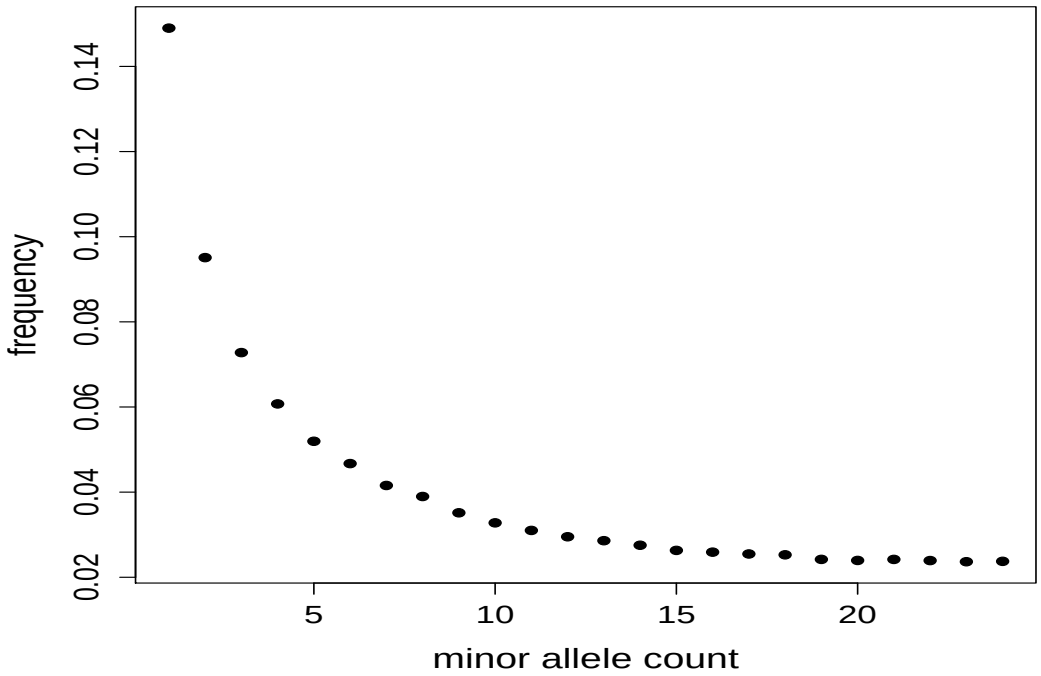


Prior distributions

- Per generation per bp mutation rate : $\mu = 2.5e-8$.
- Per generation per bp recombination rate : $r \sim \mathcal{U}(0.2e-8, 1e-8)$.
- 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)

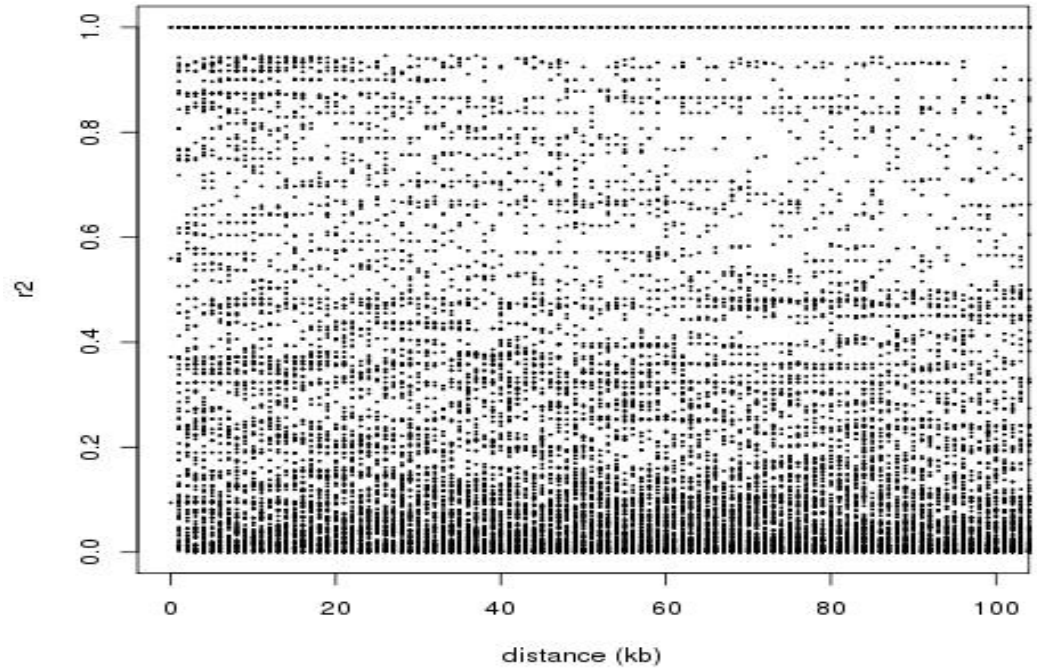
- Frequency of polymorphic sites over the genome.
- Frequency of sites with i copies of the minor allele, for i from 1 to $n/2$.



- Variance of these frequencies over the genome.

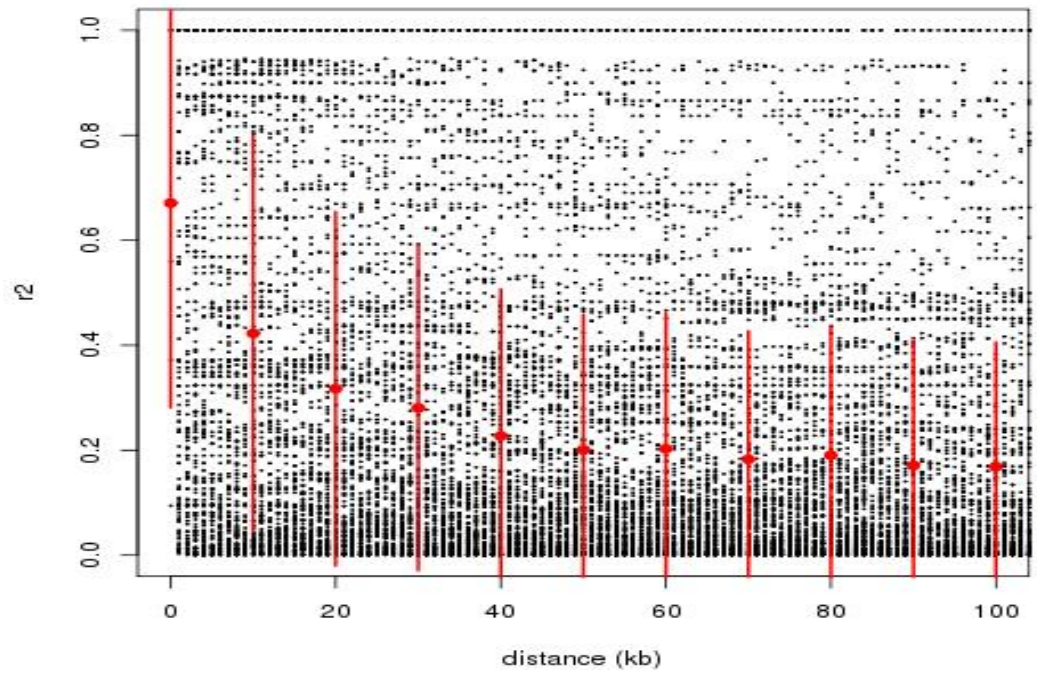
Summary statistics - Linkage Disequilibrium (LD)

- Correlation between allelic data at two polymorphic sites.



Summary statistics - Linkage Disequilibrium (LD)

- Correlation between allelic 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 $t = \frac{1}{2c(d)}$.

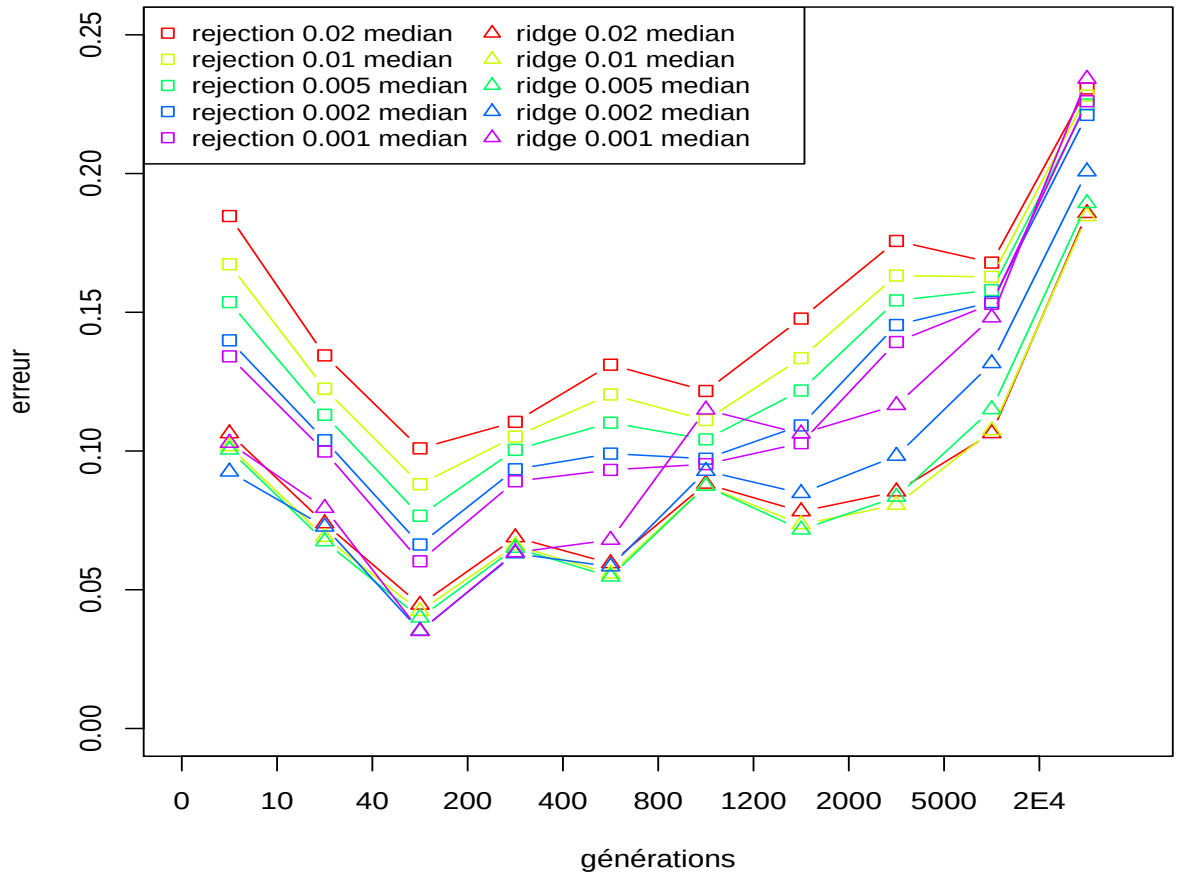
Implementation

- Simulations :
  - Haplotype data simulated with ms. One sample = 50 independent 2MB segments.
  - 500 000 simulated samples, $\approx$ 40h on a cluster with 500 jobs in parallel (4 min per sample on average).
- Holstein data :
  - Several pre-processing steps required to obtain haplotype data (sequencing, alignment, genotype calling, haplotype estimation).
  - Haplotype data processed with the same Python program.
- Final statistical analysis with the R package *abc*.

Outline

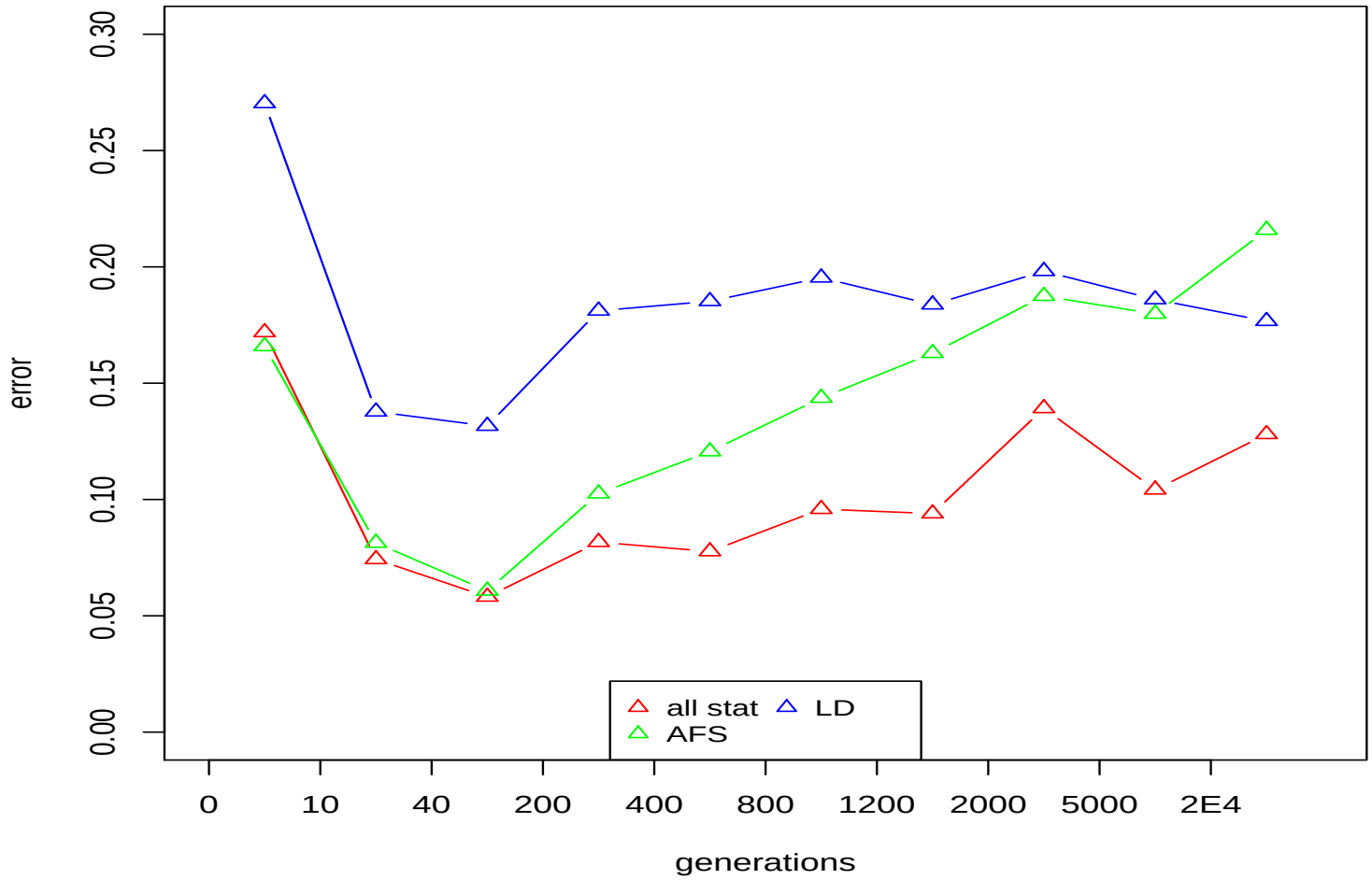
Outline

Cross validation

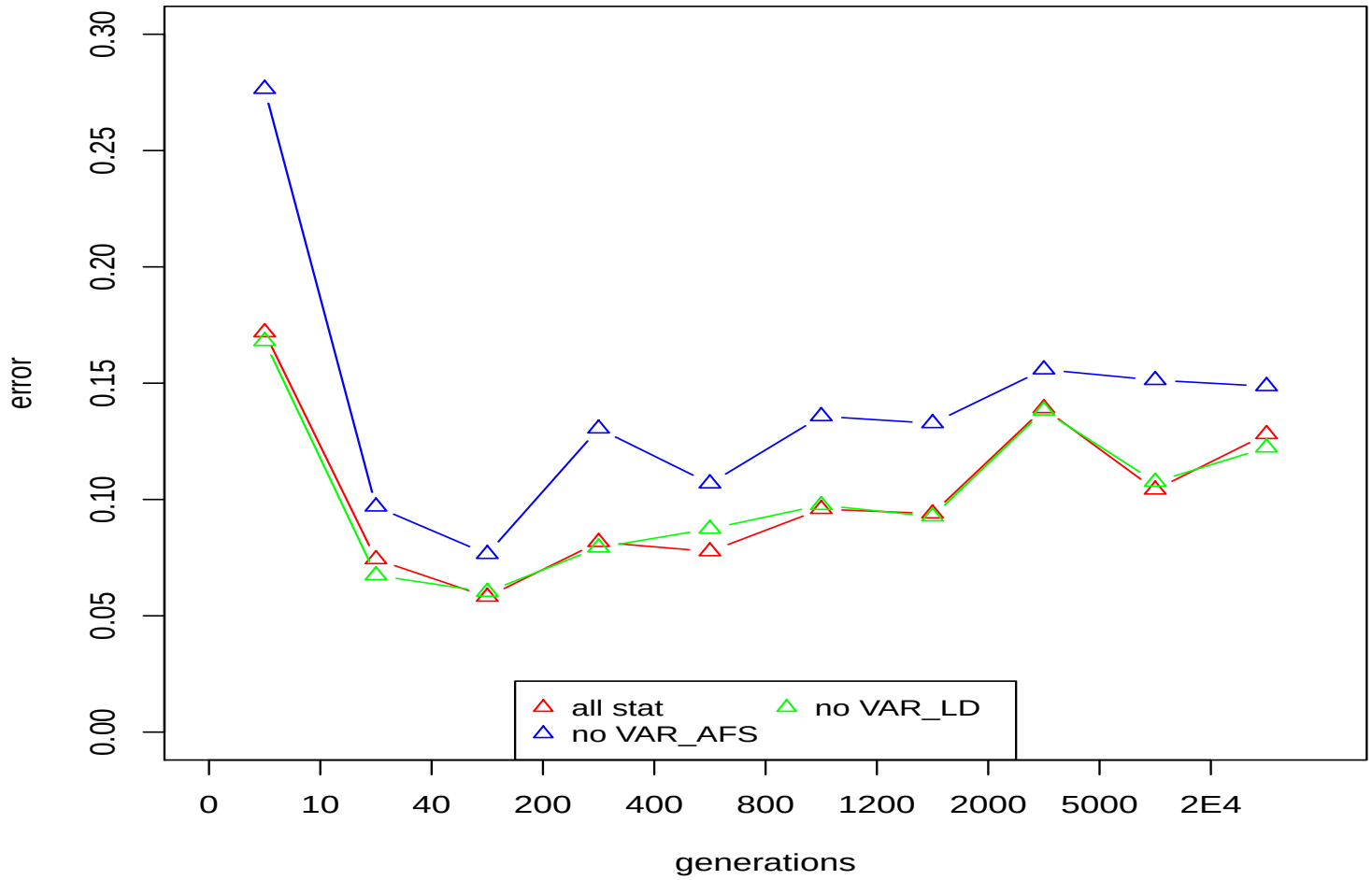


Estimation error $\frac{\sum_i (\theta_i - \hat{\theta}_i)^2}{I * Var(\theta_i)}$ based on 100 CV replicates.

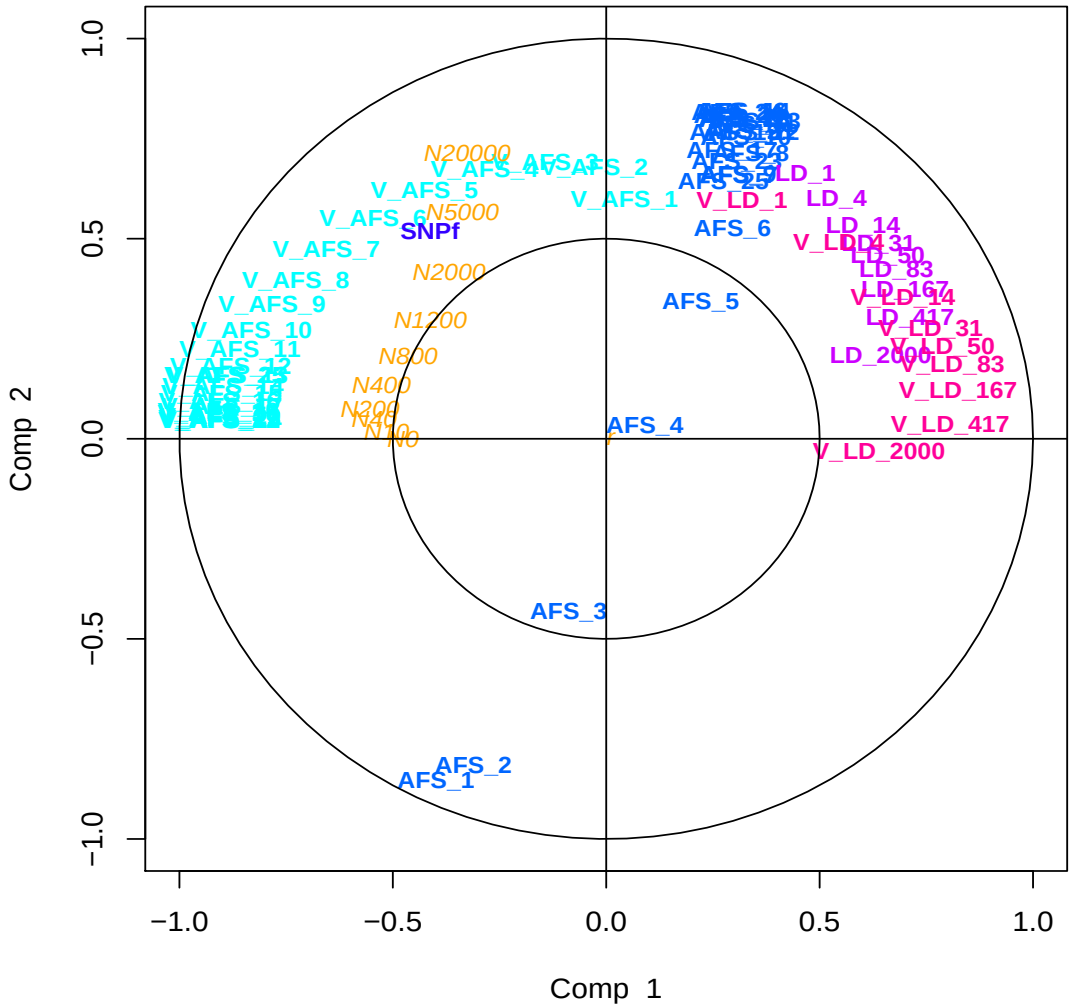
Influence of AFS and LD statistics - Cross Validation



Influence of AFS and LD statistics - Cross Validation

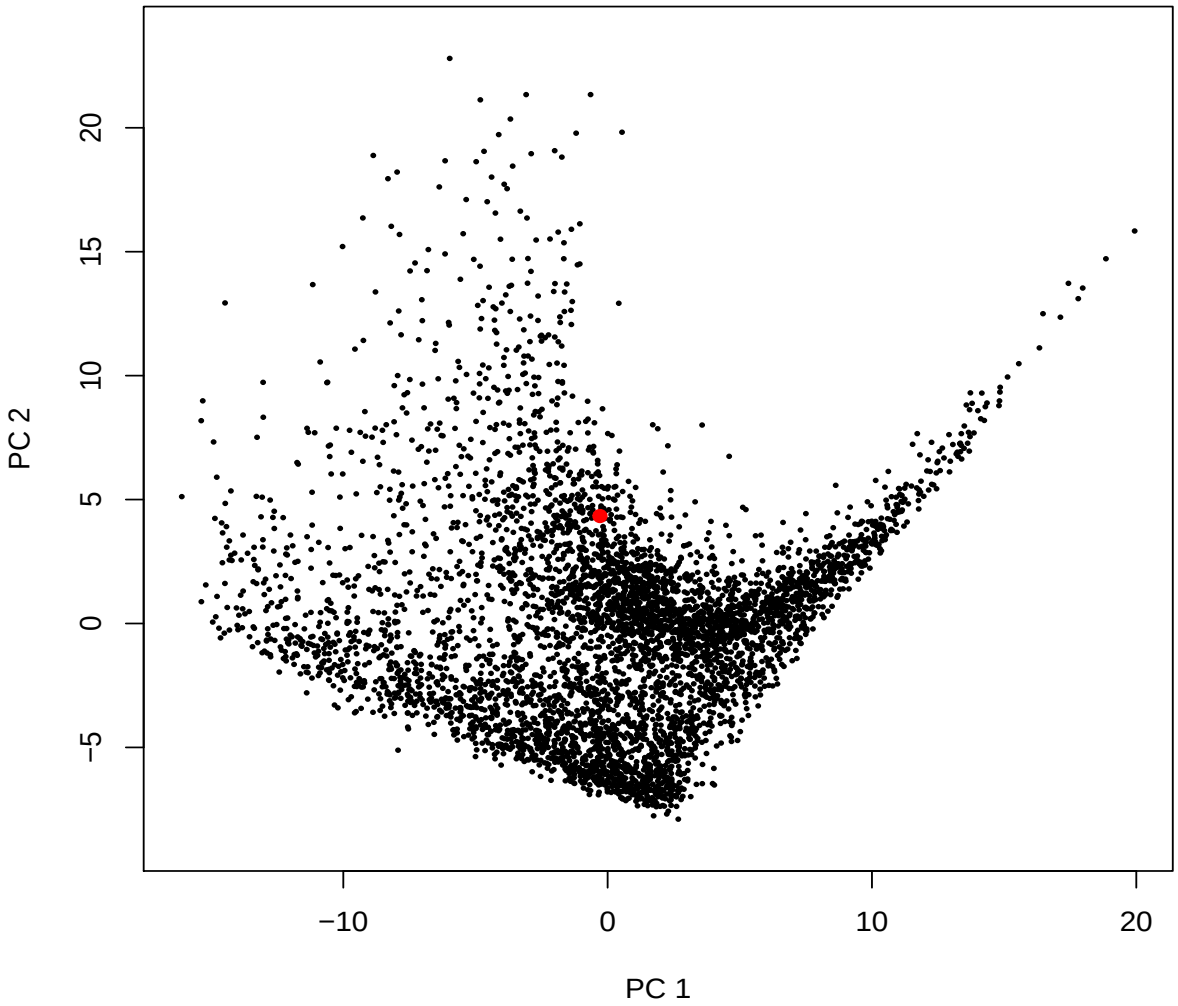


Influence of AFS and LD statistics - PLS regression

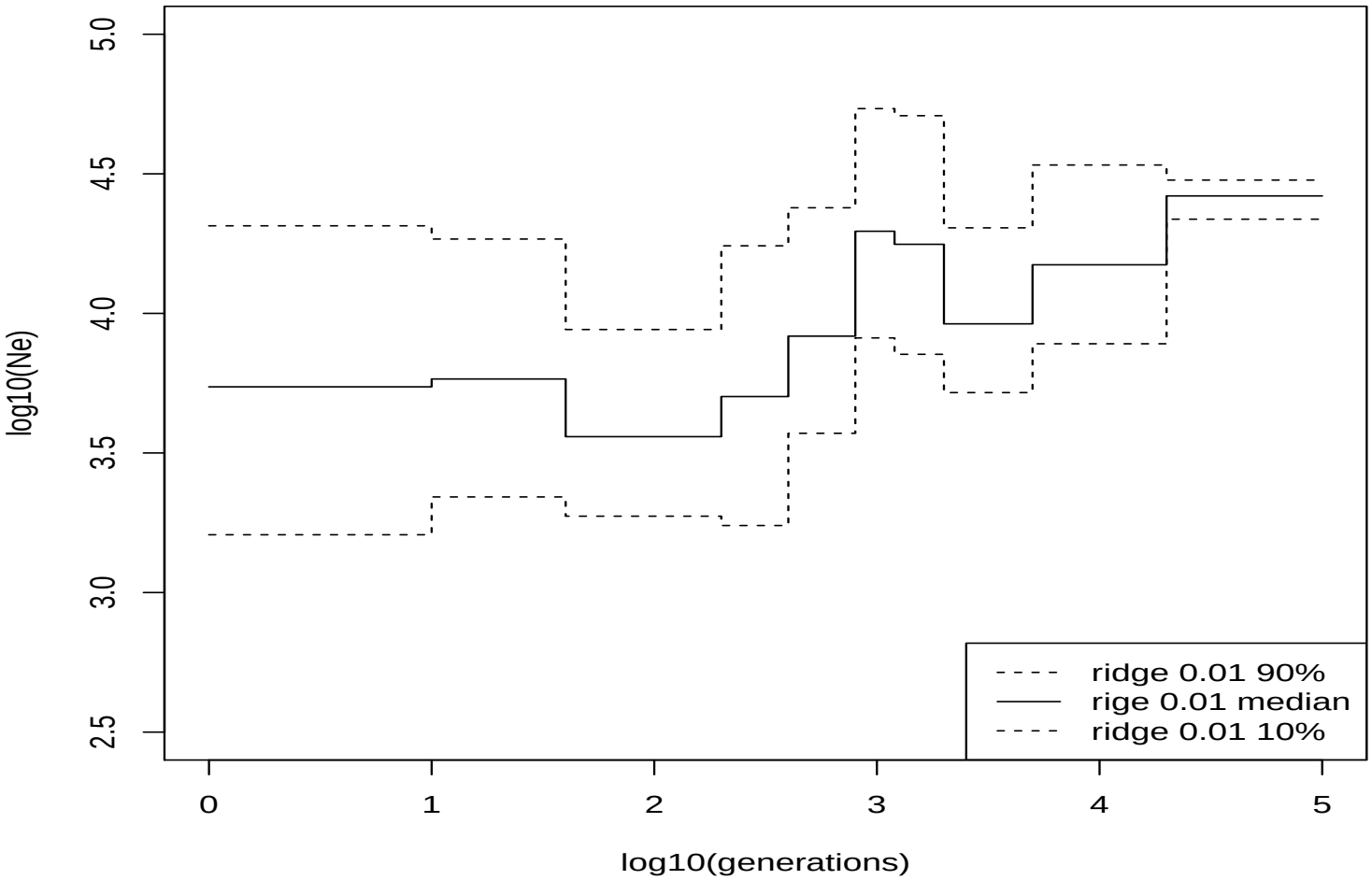


Outline

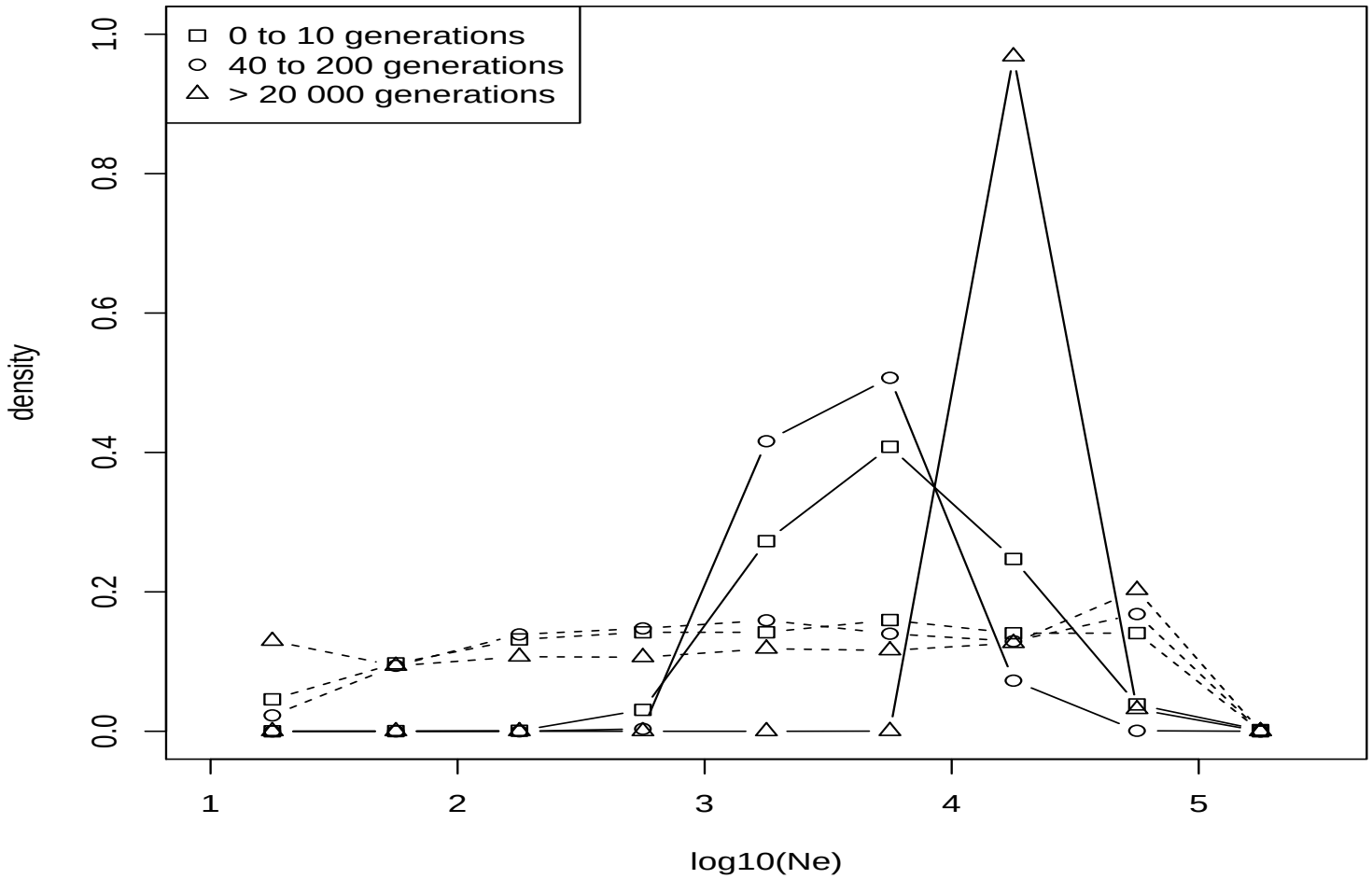
Prior check



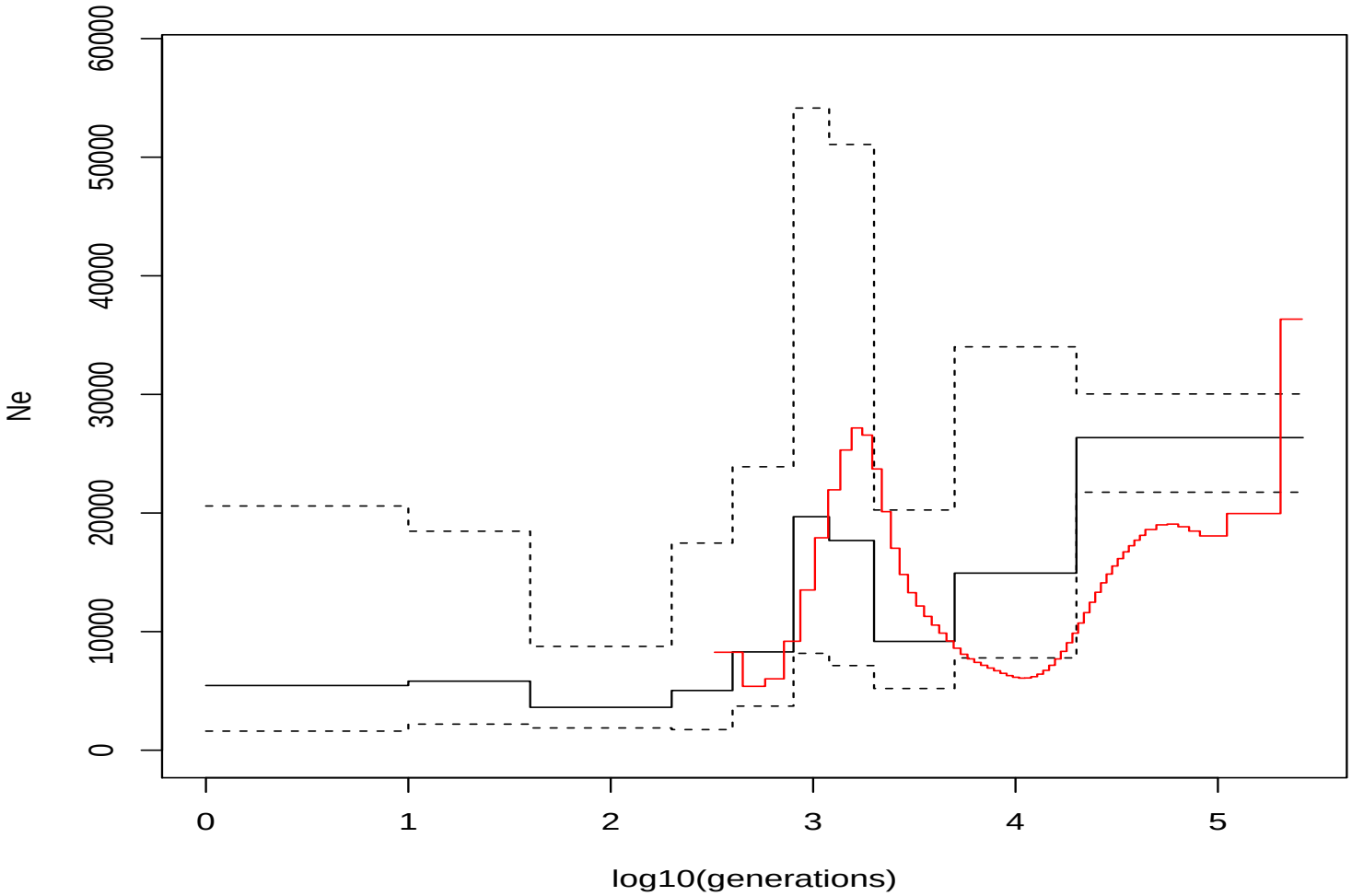
Estimated dynamics



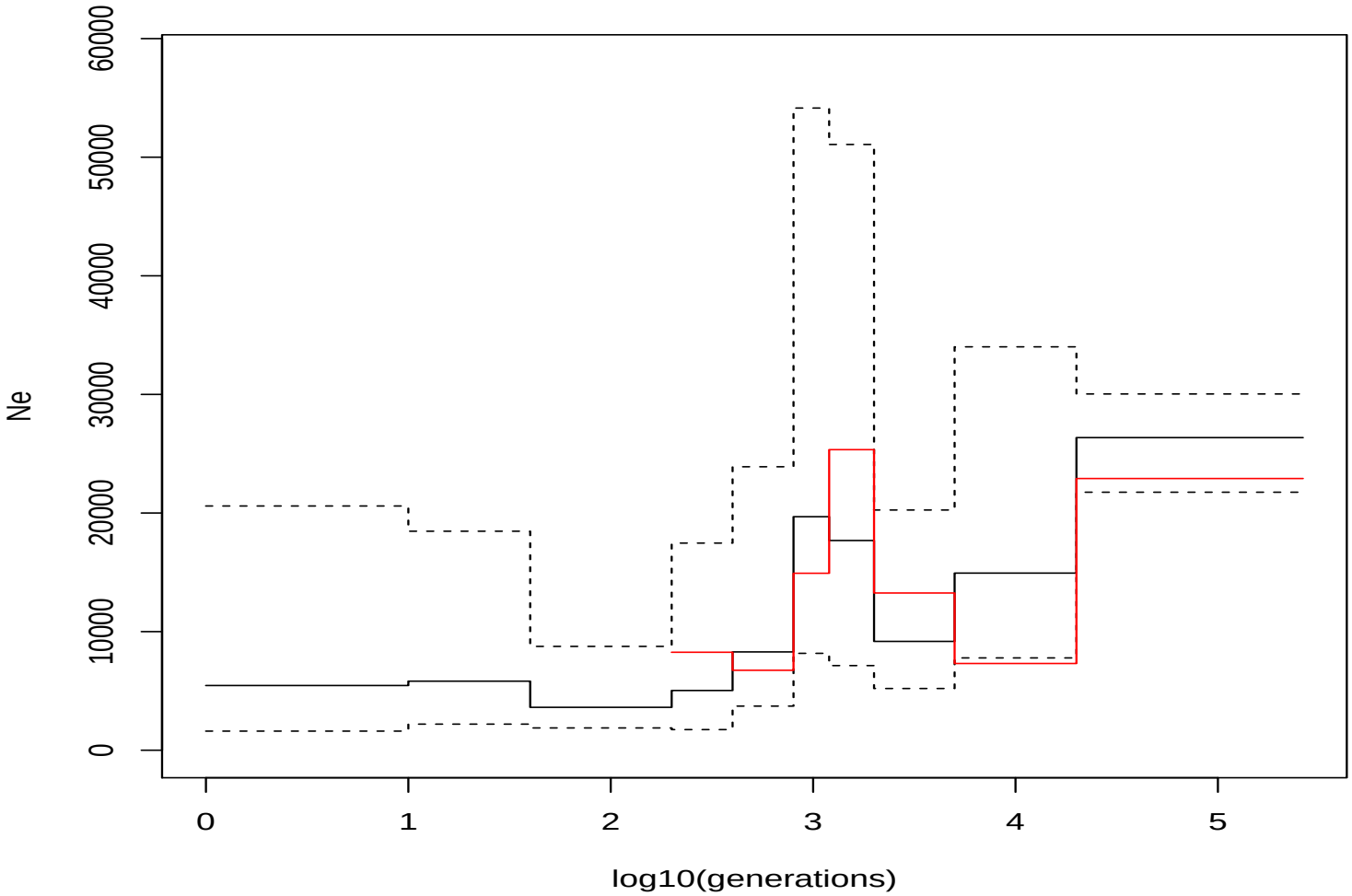
Data is informative



Comparison with PSMC



Comparison with PSMC



Outline

Conclusions

- The approach seems to work (low cross validation errors, sensible credible intervals).
- Combining AFS and LD is useful.
- Variance of AFS is useful, but variance of LD is not.
- Estimated demography is quite consistent with PSMC, but credible intervals are rather large.
- Estimation of recent population size seems too large (> 1000). Influence of sequencing errors (MacLeod *et al*, 2013)?

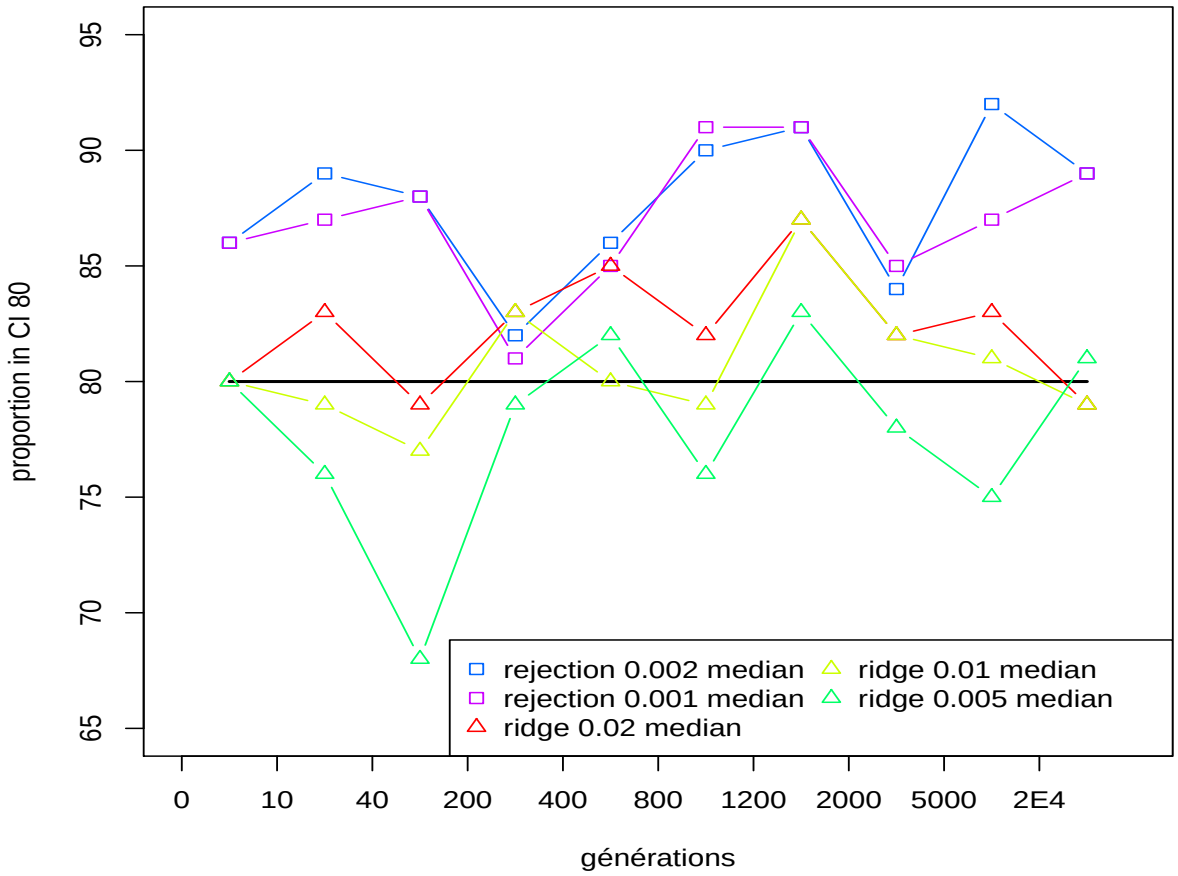
Perspectives

- Objective definition of time intervals.
- ABC with more segments ($L = 100$)?
- ABC based on more replicates? Second more local step?
- Estimation with PLS?

Acknowledgements

- Stanislas Sochacki (Ecole Polytechnique).
- Lounes Chikhi (University Toulouse III), Willy Rodriguez, Olivier Mazet, Simona Grusea (INSA Toulouse).
- Bertrand Servin (INRA, Toulouse).
- 1000 bull genomes project.

Credible Intervals



Proportion in CI 80 $\frac{1}{I}\sum_i 1(\hat{q}_{10}(\theta_i) <= \theta_i <= \hat{q}_{90}(\theta_i))$.

Summary statistics

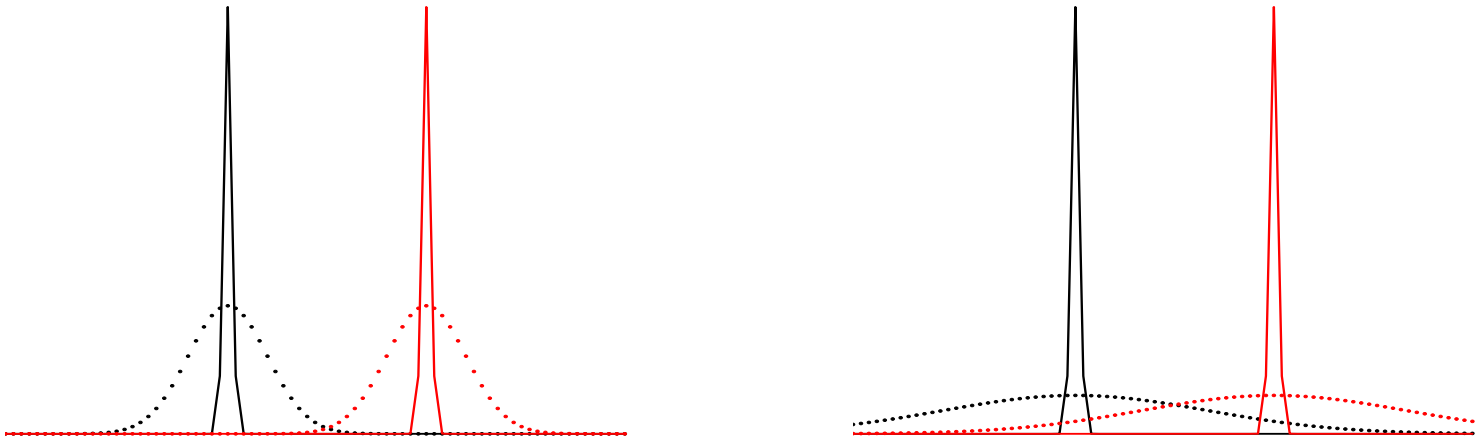
- Proportion of SNPs : $f = \mathbb{P}(x > 0)$, x number of copies of the minor allele.
- Allele frequency spectrum (AFS) : $\mathbb{P}(x = i | x > 0)$ for i from 1 to 25.
- Variance of AFS : $std(d_i) * f$ for i from 1 to 25, d_i distance between two consecutive sites with i copies of the minor allele.
- Linkage disequilibrium (LD) : $\mathbb{E}[r^2(d)]$ and $std[r^2(d)]$, $r^2(d)$ LD between SNPs at distance d.
- d=1kb, 4kb, ...2Mb, corresponding to time intervals in the model. Ex : d=1kb $\rightarrow c = 10^{-5}$M $\rightarrow t = \frac{1}{2c} = 50000$.

Number of segments

- For each position i, S_i i.i.d with $\mathbb{E}[S_i \mid \theta]$, $Var(S_i \mid \theta)$.
- Our statistics are averages, i.e. $S_L = \frac{1}{L}\sum_{i=1}^{L} S_i$
  $\rightarrow \mathbb{E}[S_L \mid \theta] = \mathbb{E}[S_i \mid \theta]$, $Var(S_L \mid \theta) = \frac{1}{L} Var(S_i \mid \theta)$
- $Var(S_{genome} \mid \theta) = \frac{1}{3*10^9} Var(S_i \mid \theta)$
- $Var(S_{50*2Mb} \mid \theta) = \frac{1}{5} Var(S_{10*2Mb} \mid \theta)$
- When does this variance become too large?

Number of segments

$Var(S_L \mid \theta)$ must remain small compared to $Var_\theta(\mathbb{E}[S_L \mid \theta])$



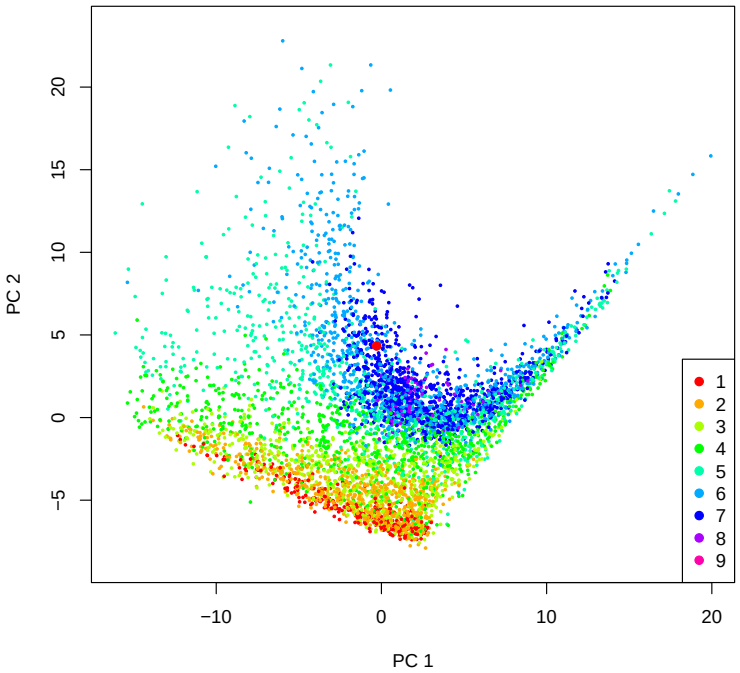
- Computation of $Var(S_L \mid \theta)/Var_\theta(\mathbb{E}[S_L \mid \theta])$ for 1000 θ_i values sampled from the prior distribution.
- For each θ_i, 50 replicates of S_L.

Number of segments

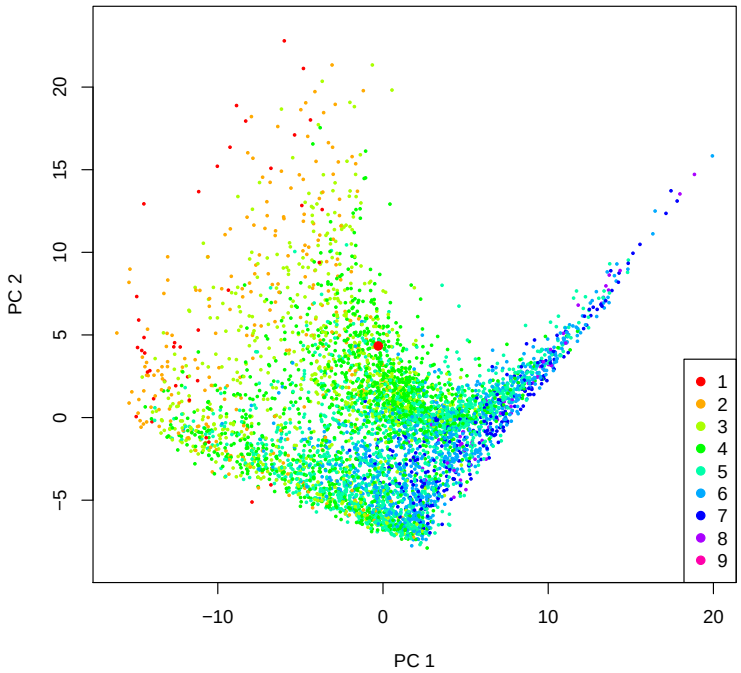
Distribution of $Var(S_L \mid \theta)/Var_\theta(\mathbb{E}[S_L \mid \theta])$

Statistic	$L = 10 * 2Mb$		$L = 50 * 2Mb$	
	q_{90}	prop < 0.1	q_{90}	prop < 0.1
AFS_1	0.14	0.87	0.03	0.97
AFS_2	0.79	0.53	0.16	0.82
AFS_{25}	3.68	0.47	0.73	0.67
VAR_AFS_1	< 0.01	0.97	< 0.01	0.98
VAR_AFS_{25}	0.19	0.83	0.04	0.97
LD_{2Mb}	1.2	0.54	0.24	0.82
LD_{1kb}	0.64	0.78	0.13	0.89
VAR_LD_{2Mb}	4.71	0.04	0.94	0.22
VAR_LD_{1kb}	1.94	0.63	0.39	0.75
x	< 0.01	1	< 0.01	1

Prior check



$\overline{N}$



$N_0 - N_\infty$