Inference of selection from genetic time series using various parametric approximations to the Wright-Fisher model

Cyriel Paris, Bertrand Servin, Simon Boitard

Inference of selection from genetic time series using various parametric approximations to the Wright-Fisher model

Cyriel Paris, Bertrand Servin, Simon Boitard

INRA, Génétique Physiologie et Systèmes d'Elevage (GenPhySE), Toulouse

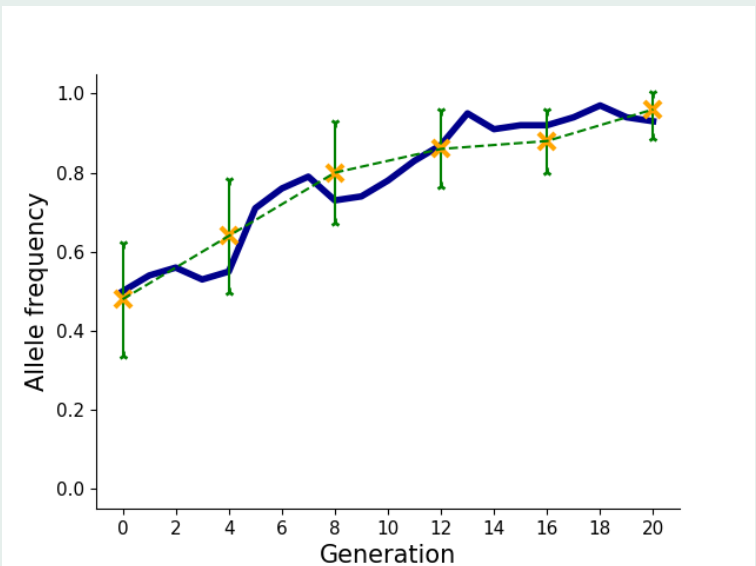
Workshop 'Evolutionary genomics and life history traits'
labex CeMeb
Montpellier, 16 octobre 2019

Introduction

Context

- Increasing availibility of genomic time series data: experimental evolution, gene banks, ancient DNA ...
- **Objective:** detect selection and estimate its intensity

Selection and time series



Data

- One time series per SNP
- Sample size
- Reference allele count

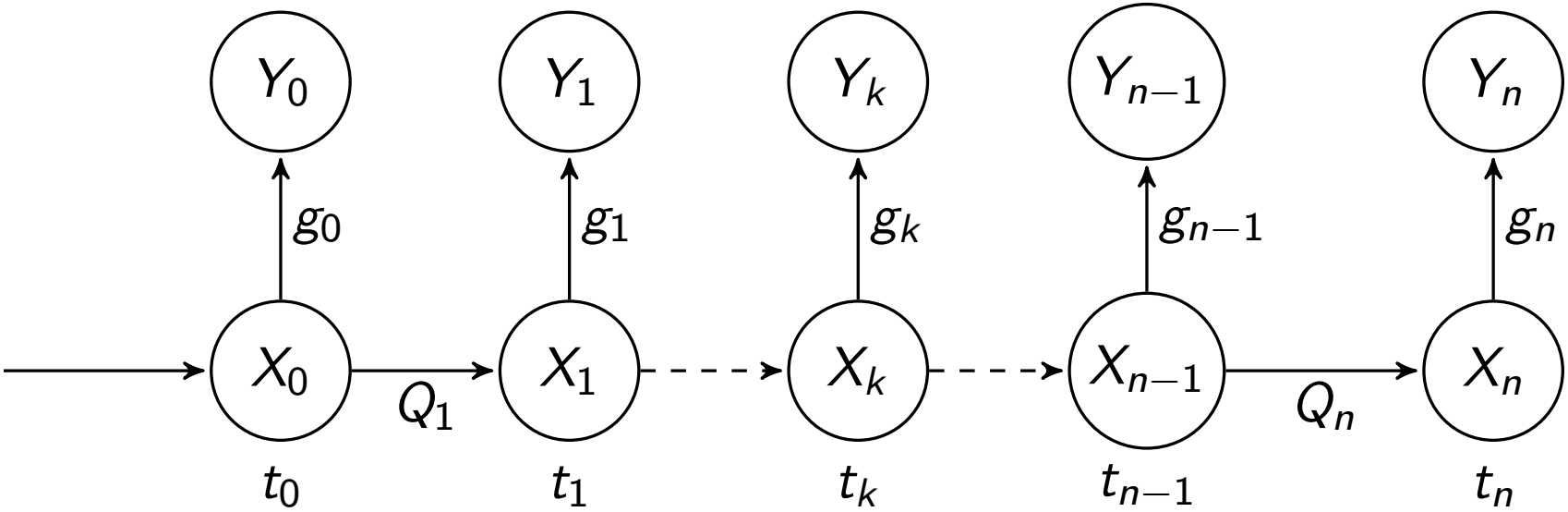
Need to model

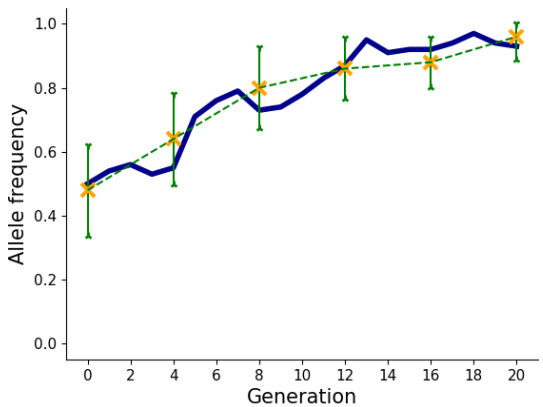
- Allele frequency evolution
- Sampling variability

Outline

Outline

Hidden Markov Model (Bollback *et al*, 2008)

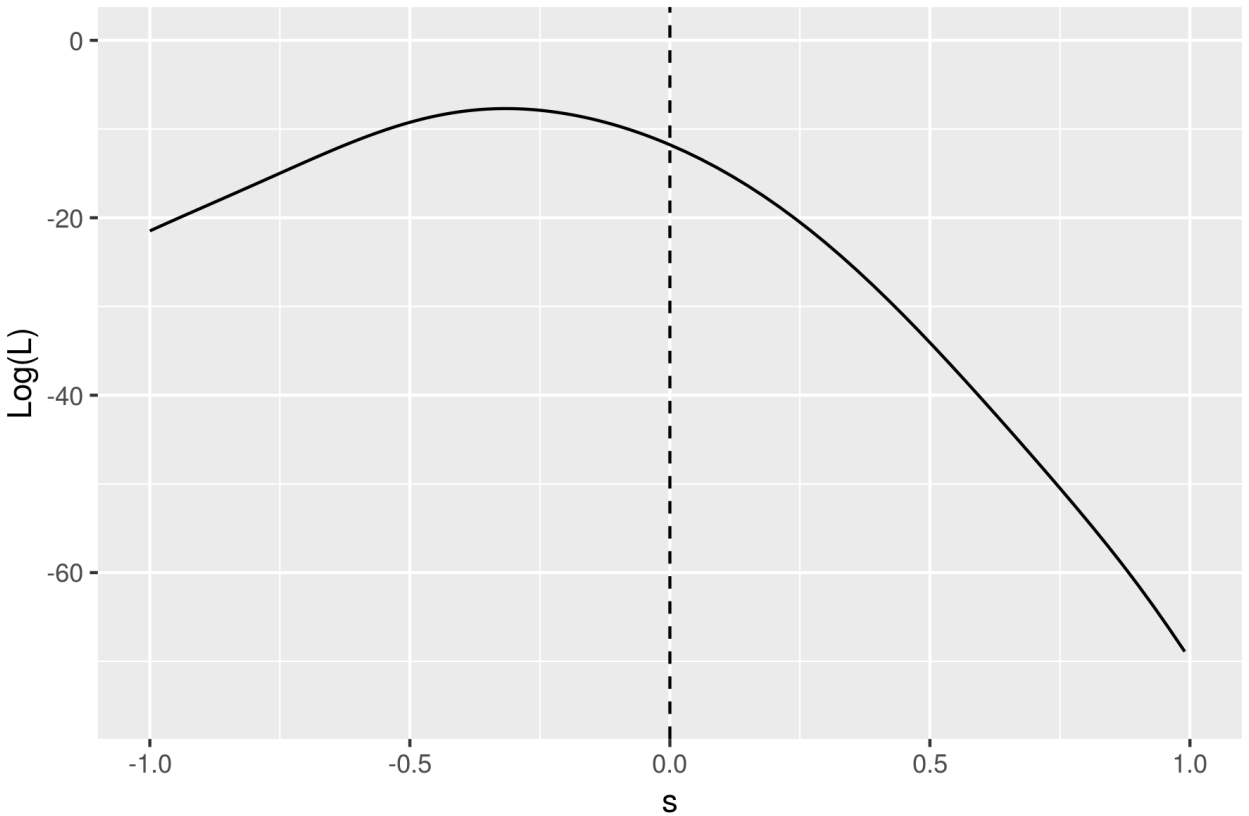




- Hidden states $\{X_k\}_{k\geq 0}$: Population allele frequencies, Markov chain, **transition** $Q = Q^{s,h,N}$
- Observations $\{Y_k\}_{k\geq 0}$: Sample allele counts, independent given $\{X_k\}_{k\geq 0}$, **emission g**

Selection Inference

- Likelihood $L(y_0, \ldots, y_n; Q, g)$ efficiently computed using the Forward algorithm.
- $L(y_0, \ldots, y_n; Q, g) = L(y_0, \ldots, y_n; s, N, h)$
  $\rightarrow$ evaluate on a grid for different values of s, N, h.



Selection inference

- Estimate s by maximum likelihood
- Test selection using the Likelihood Ratio Statistic

$$2(\log L(y_0, \ldots, y_n; \hat{s}, N, h) - \log L(y_0, \ldots, y_n; 0, N, h))$$

  chi-quare distributed under $s = 0$ (neutrality)
- **Computation of Q?**

Reference model: Wright-Fisher

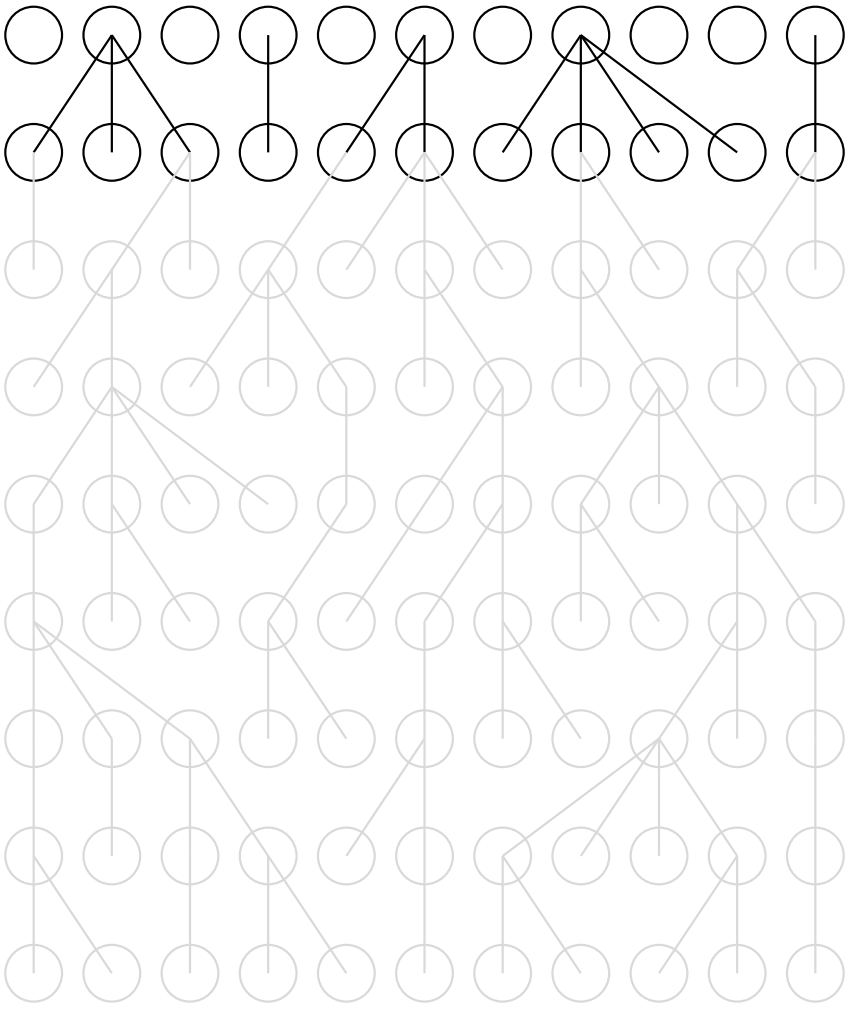
Hypotheses

- Panmixia, non overlapping generations
- Constant population size (N)
- X_t: allele frequency at generation t
- $X_{t+1}|X_t \sim \frac{1}{N}\mathcal{B}(N, X_t)$

Adding Selection

A_1A_1	A_1A_0	A_0A_0
$1+s$	$1+sh$	1
- f_s: fitness function
- $X_{t+1}|X_t \sim \frac{1}{N}\mathcal{B}(N, f_s(X_t))$, P
- $X_{t+u}|X_t$, $Q = P^u$

Reference model: Wright-Fisher



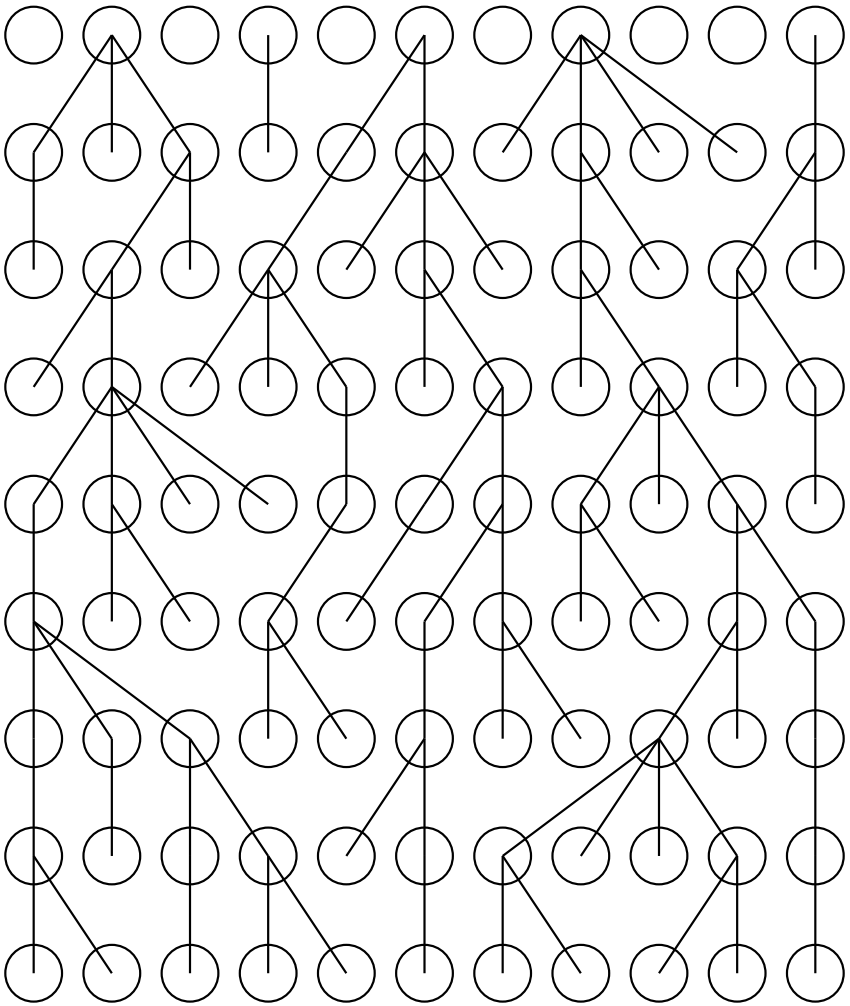
Hypotheses

- Panmixia, non overlapping generations
- Constant population size (N)
- X_t: allele frequency at generation t
- $X_{t+1}|X_t \sim \frac{1}{N}\mathcal{B}(N, X_t)$

Adding Selection

A_1A_1	A_1A_0	A_0A_0
$1+s$	$1+sh$	1
- f_s: fitness function
- $X_{t+1}|X_t \sim \frac{1}{N}\mathcal{B}(N, f_s(X_t))$, P
- $X_{t+u}|X_t$, $Q = P^u$

Reference model: Wright-Fisher



Hypotheses

- Panmixia, non overlapping generations
- Constant population size (N)
- X_t: allele frequency at generation t
- $X_{t+1}|X_t \sim \frac{1}{N}\mathcal{B}(N, X_t)$

Adding Selection

-

A_1A_1	A_1A_0	A_0A_0
$1+s$	$1+sh$	1

- f_s: fitness function
- $X_{t+1}|X_t \sim \frac{1}{N}\mathcal{B}(N, f_s(X_t))$, P
- $X_{t+u}|X_t$, $Q = P^u$

Wright-Fisher's Limitations

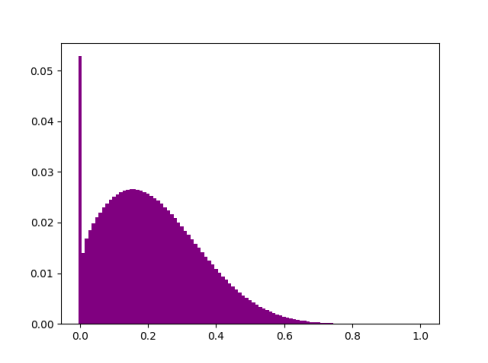
- Wright-Fisher model:
  - Matrix Q has dimension $N \times N$
  - Computing time: $Q = P^u$ + forward algorithm
  - Memory load
- Alternative: find a good approximation
  - Control Q dimension
  - Fits Wright-Fisher transitions
  - Mimics the statistical properties obtained with Wright-Fisher transitions (parameter estimation and detection power)

The method of moments (Lacerda and Seoighe, 2014)

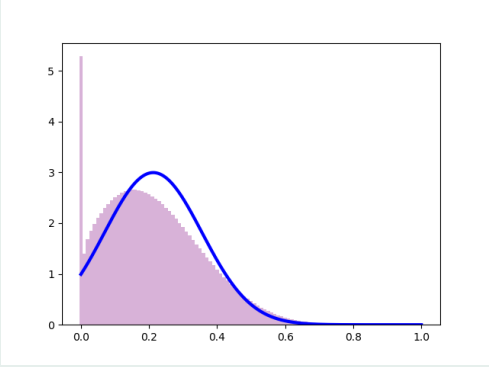
- Choose a parametric distribution
- Set parameters such that its moments fit those of the Wright-Fisher (WF).
- **Gaussian model (Ga):** Popular approximation for short time and intermediate allele frequency (Cavalli-Sforza *et al*, 1964; Lacerda and Seoighe, 2014; Terhorst *et al.*, 2015).
- **Nicholson Gaussian model (NG):** Allocate the mass outside $(0, 1)$ to point masses in 0 and 1 (Nicholson *et al*, 2002).
- **Beta model (Be):** Wide variety of shapes, distributed in $[0, 1]$ (Balding and Nichols, 1995; Hui and Burt, 2015).
- **Beta with spikes model (BwS):** Adds fixation probabilities to Beta model (Tataru *et al*, 2016).

Moment fitting examples

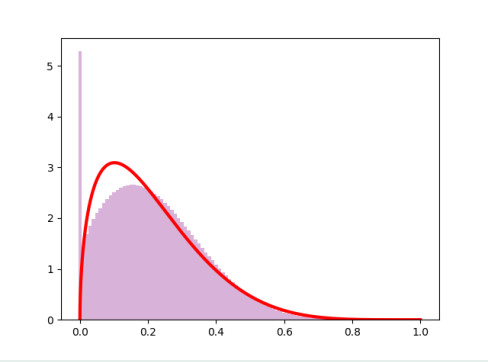
Wright-Fisher distribution



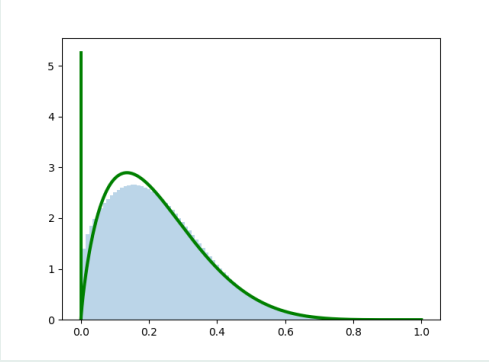
Gaussian distribution



Beta distribution



Beta with spikes distribution

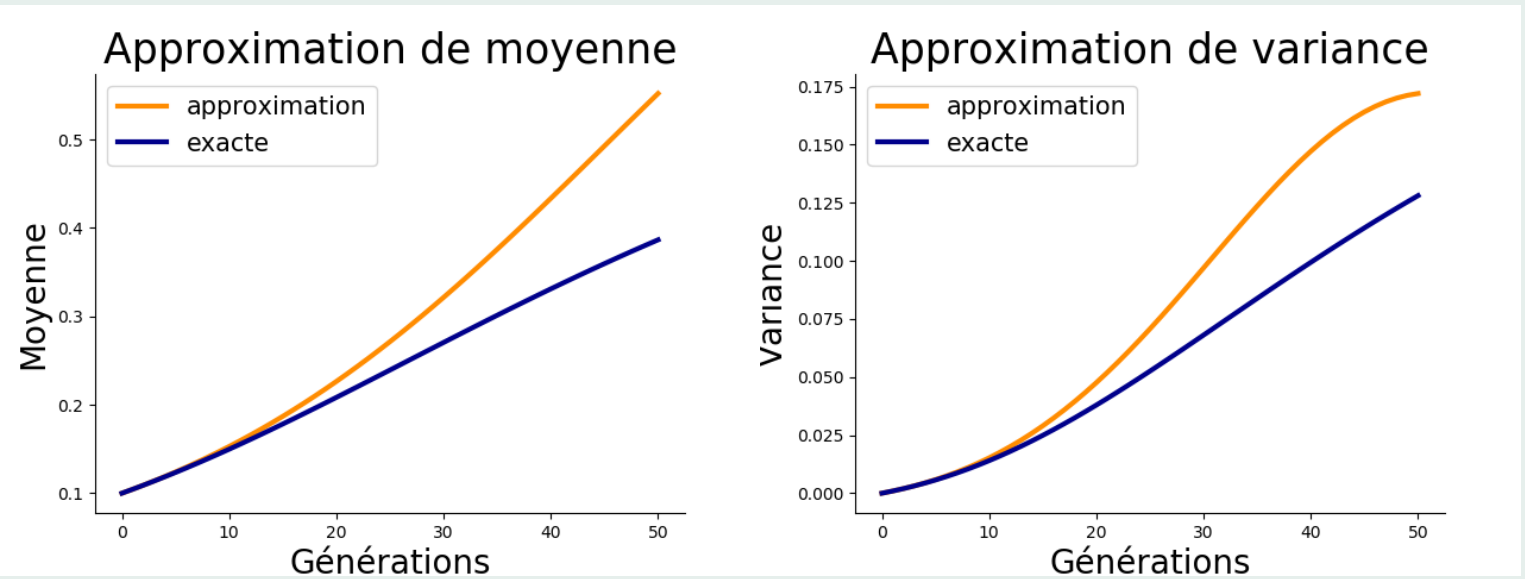


Moment approximation: Delta-method

Taylor approximation (Lacerda and Seoighe, 2014)

$$\begin{array}{rcl@{\qquad}rcl} \mu_{n+1} & \approx & f(\mu_n) & \mu_n & \approx & E(X_n) \\ \sigma^2_{n+1} & \approx & f'(\mu_n)^2\sigma^2_n & \sigma^2_n & \approx & Var(X_n) \end{array}$$

Accuracy ($x_0 = 0.1, N = 100, s = 0.1$)



- BwS: fixation proba. also approximated (Tataru *et al*, 2016).

Outline

Simulations 1: fit to the WF model

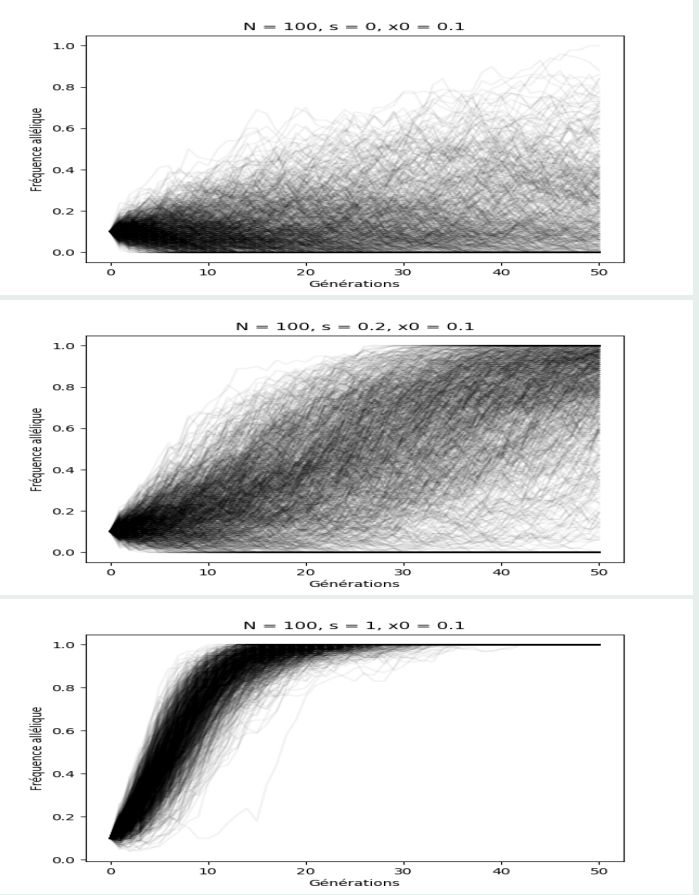
Simulated process

- Wright-Fisher with selection

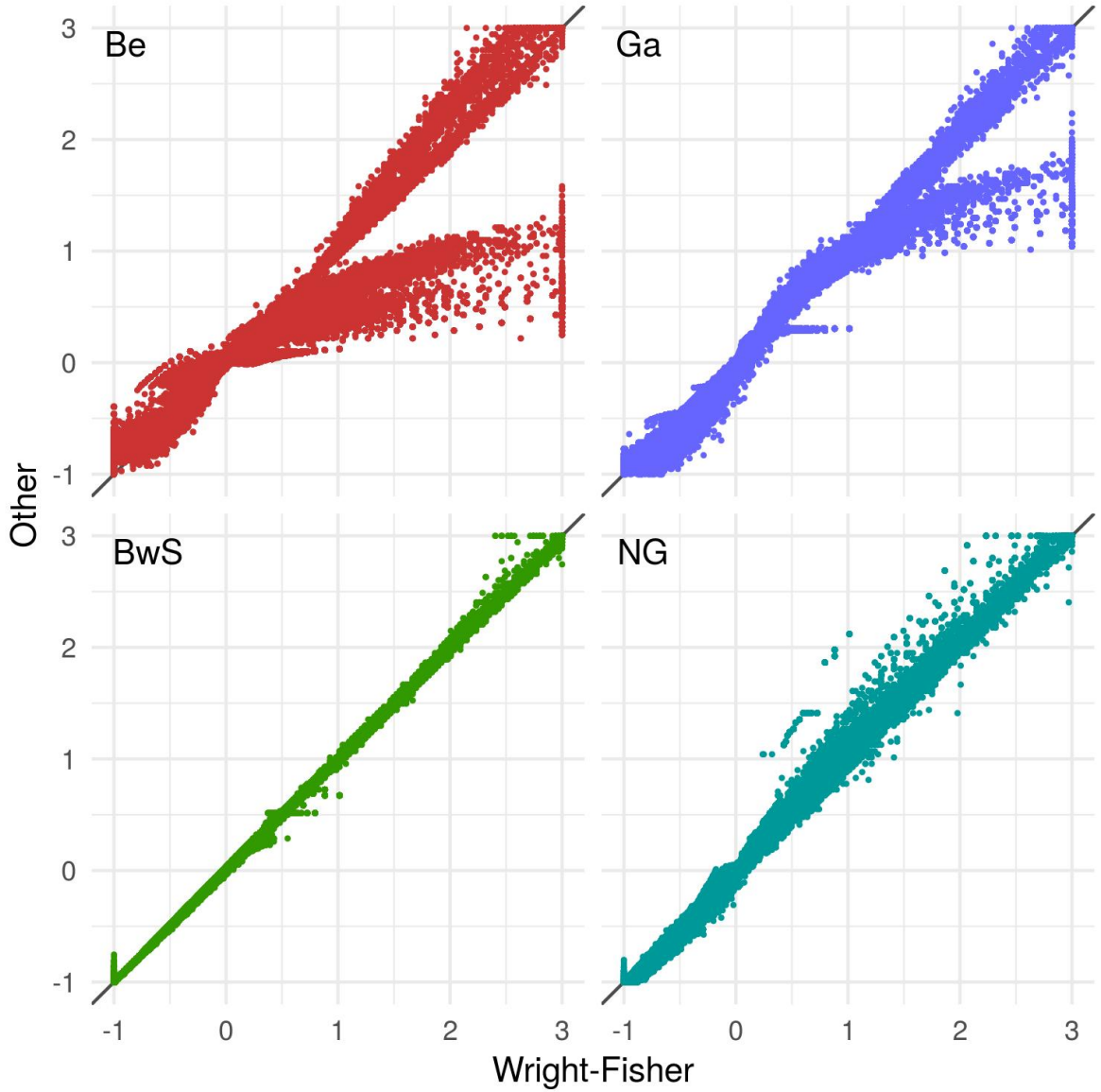
Parameters

- $N = 100$
- $x_0 = 0.1$ or 0.5
- $s = 0$ to 1
- 10 sampling dates
- 30 alleles per sample
- Total evolution time: $T = 9$ to 180 generations

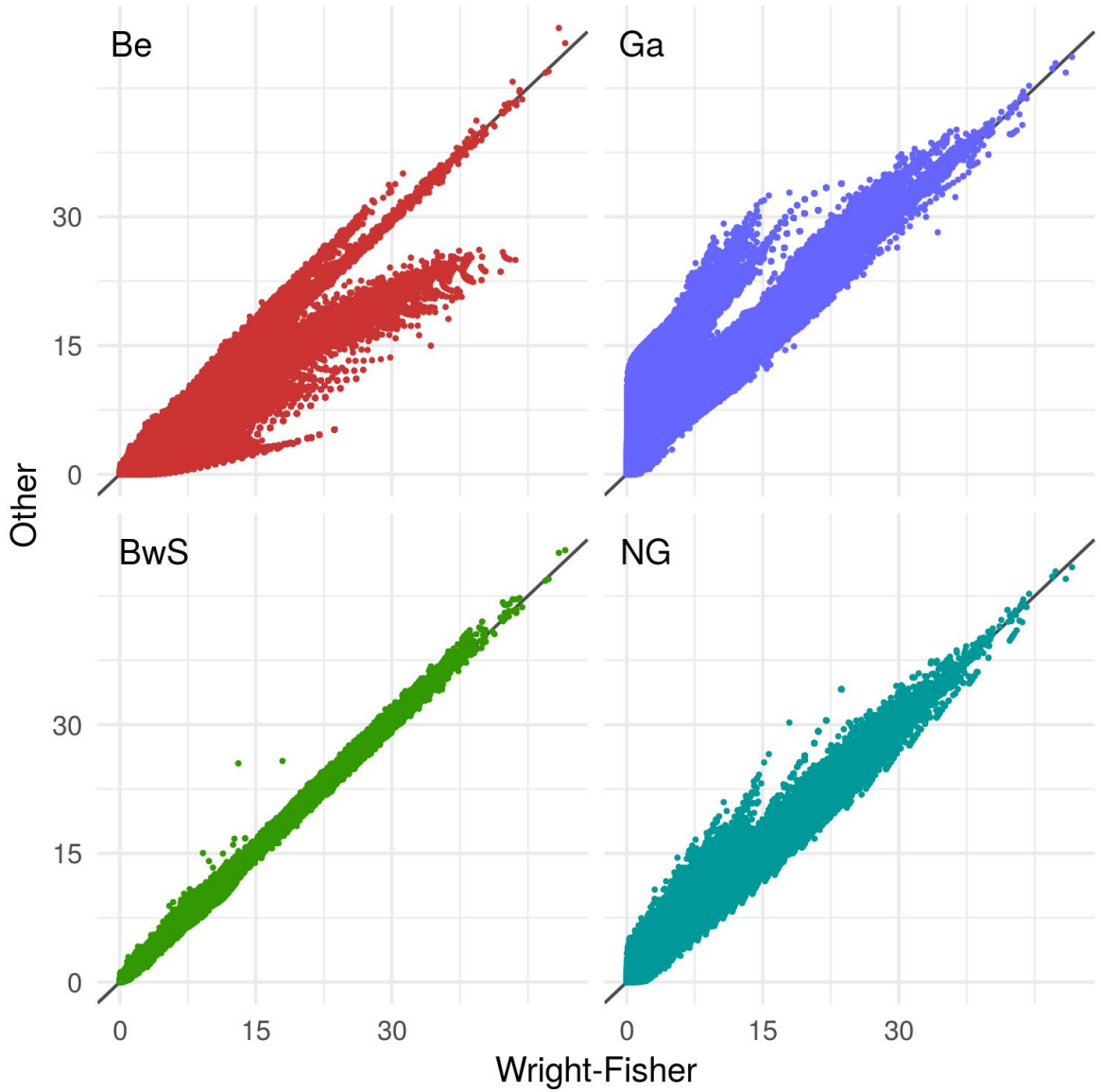
Simulation examples



Maximum Likelihood Estimation



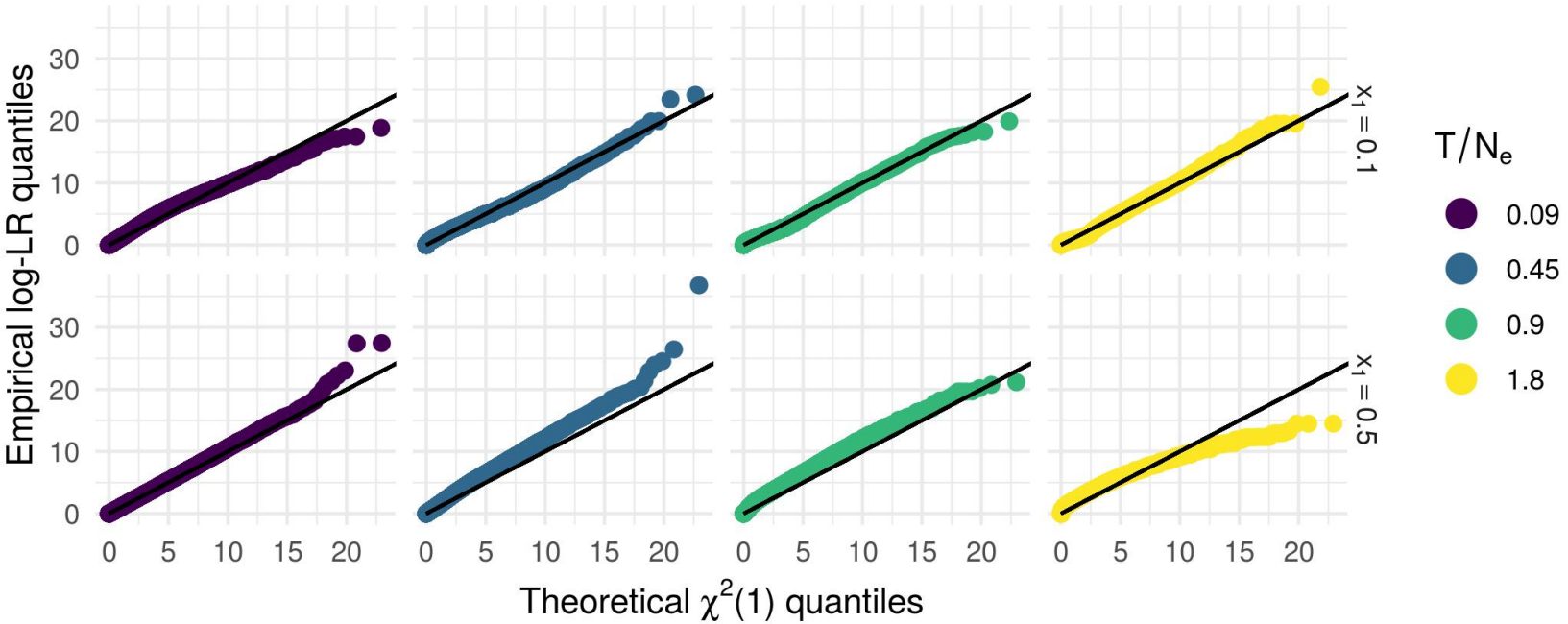
Likelihood Ratio Statistic



Simulations 2: statistical properties

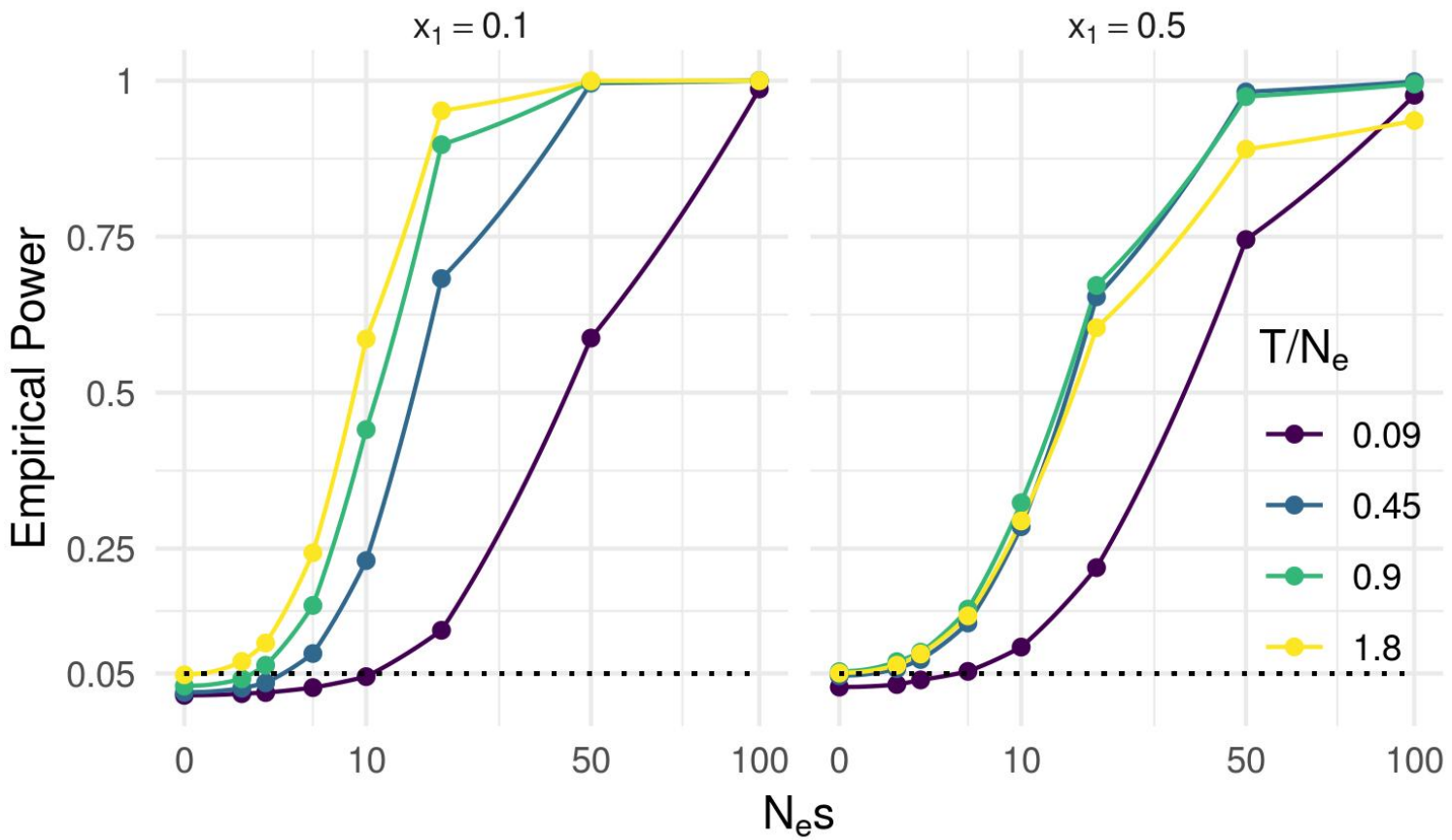
- Data analysis only with BwS transitions.
- $N = 100$, 1000 or 10000.
- x_0 and sampling design as before.
- T/N as before: from 0.09 to 1.8.
- Ns as before: from 0 to 100.

Likelihood Ratio Test calibration



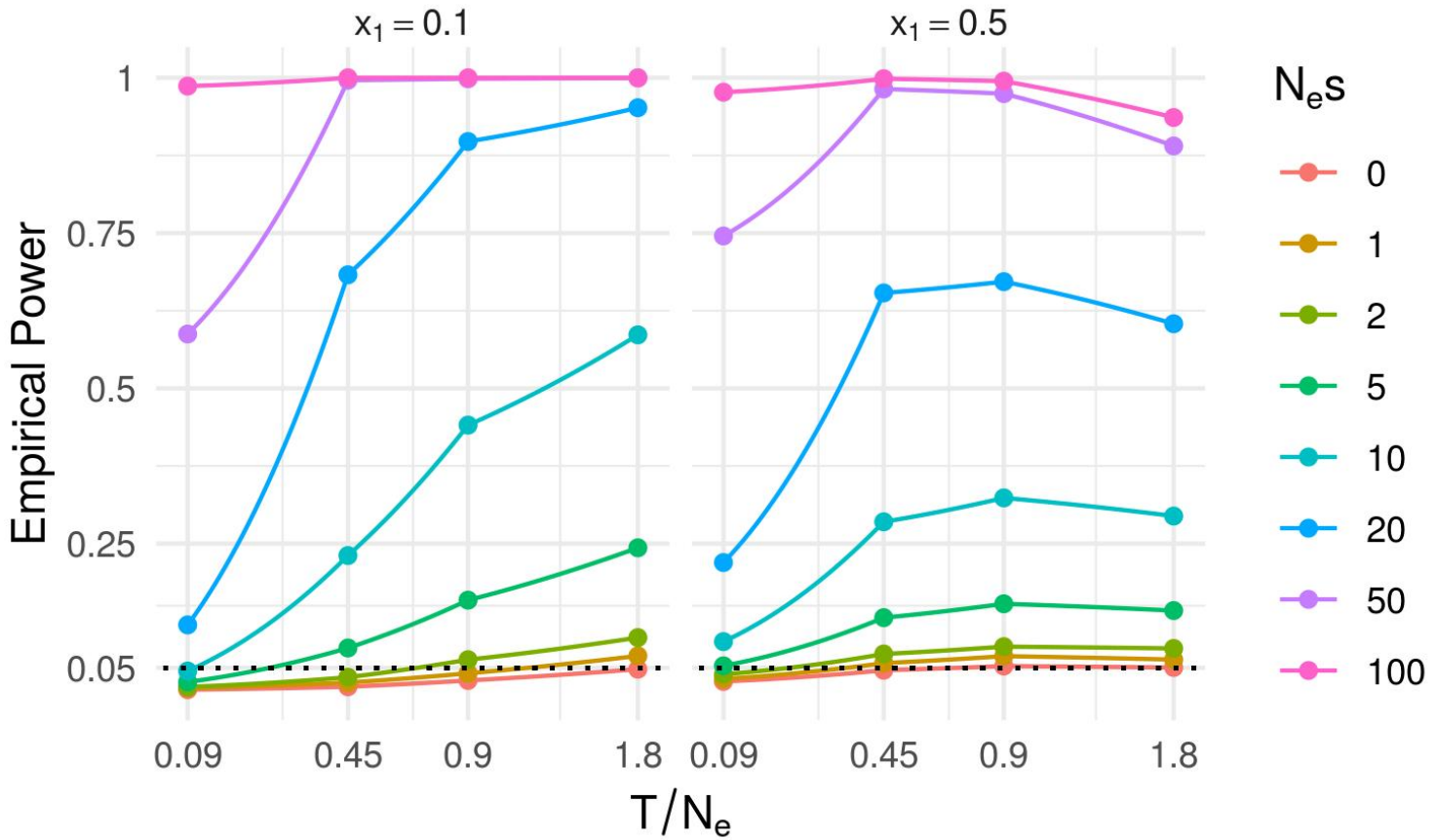
- Neutral simulations
- Good fit to the chi-quare distribution, except for $(T = 1.8, x_0 = 0.5)$.

Detection Power

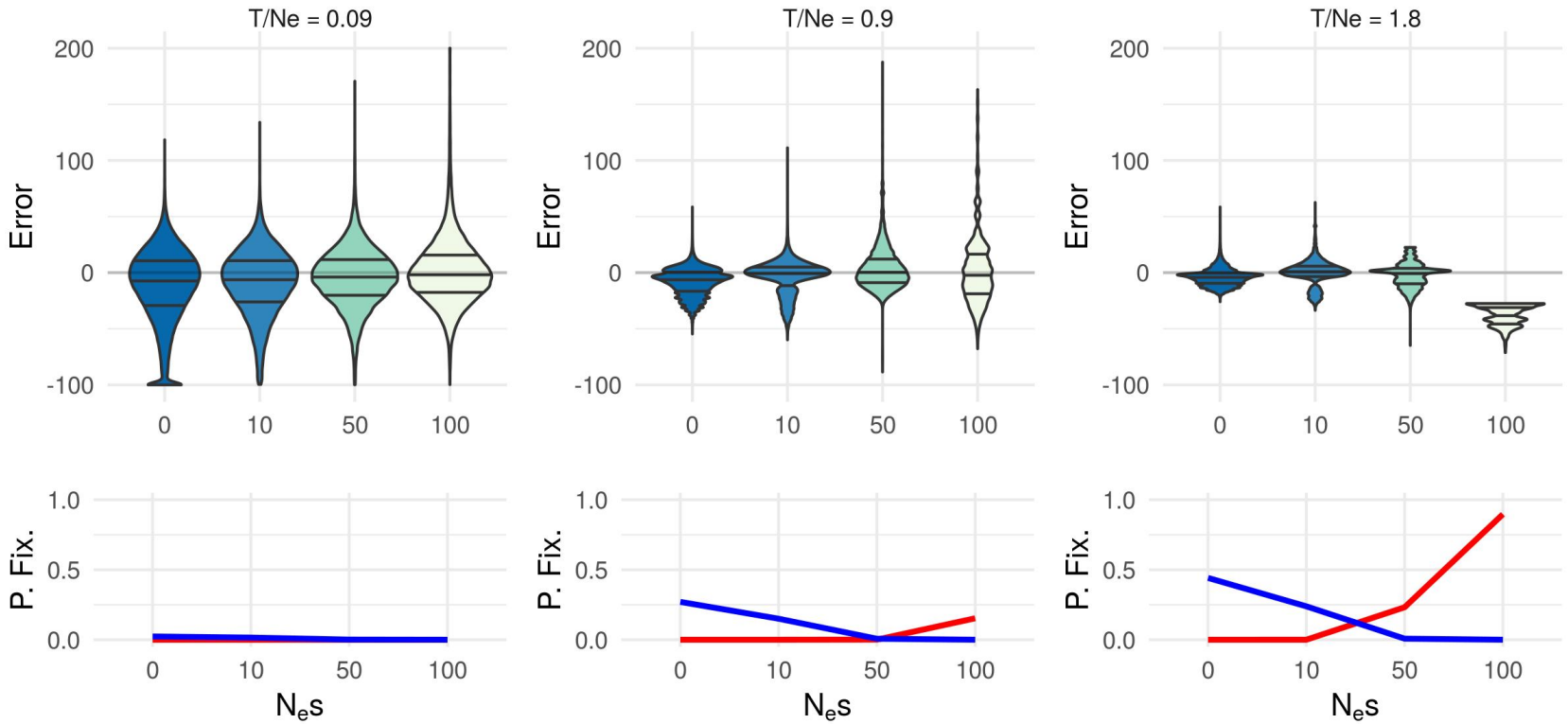


- Increases with Ns, T/N (up to a certain point).

Detection Power



Estimation of s



- Unbiased except for $(Ns, T/N)$ large (fixation probabilities).
- Lower variance for large T/N.

Outline

Motivation

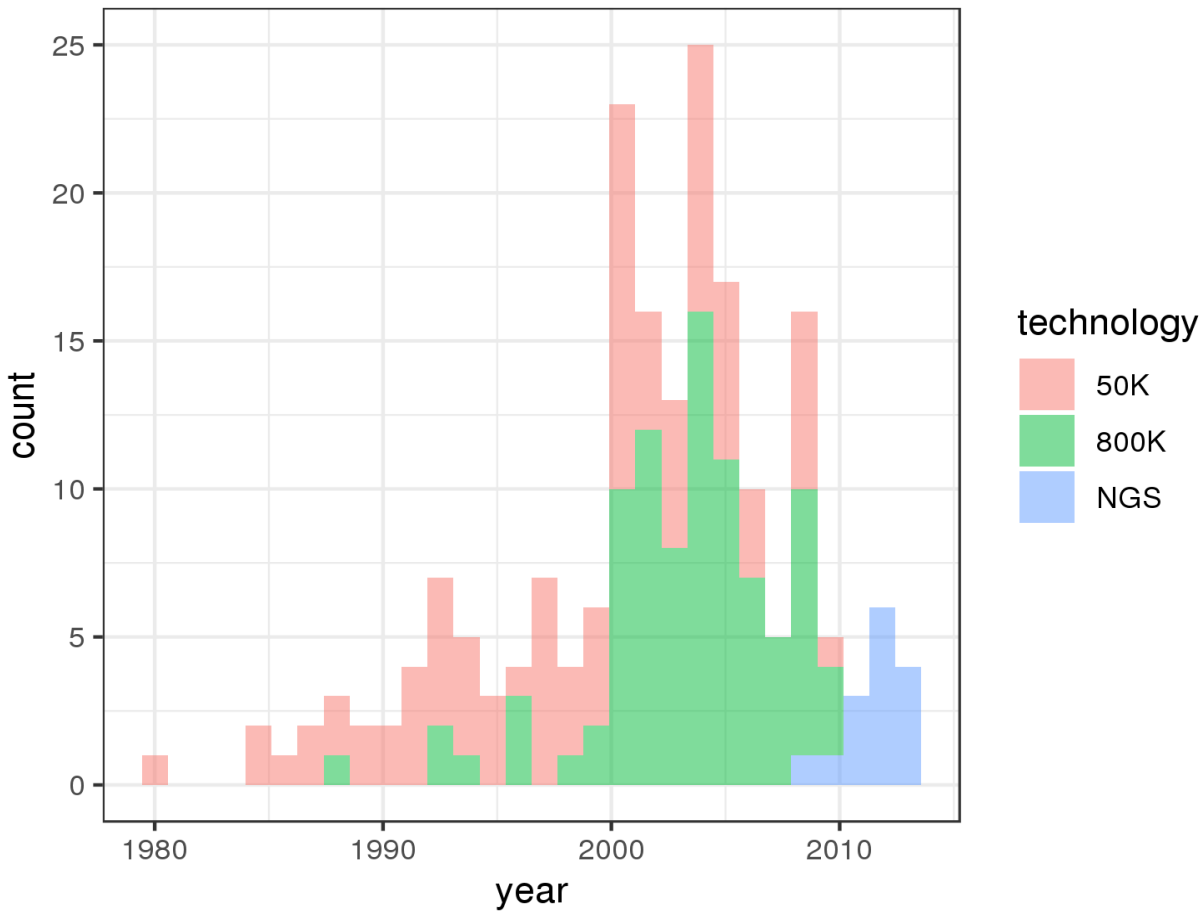
- Past positive selection can be detected from present time molecular data.
- Annotation of selection events more difficult: onset and intensity of selection? adaptive trait?
- Usefulness of time series (biobanks) to (i) detect recent selection or (ii) annotate signatures of historical selection?

Case study: Asturiana de los Valles

- Spanish autochtonous beef cattle breed.
- North of Spain (Asturias).
- Semi-extensive breeding conditions.
- Evolution of genetic diversity along 35 years (1980 – 2015).

Data

- Genotype data (50K or 800K) for 137 animals.
- 15 animals sequenced at $\approx$ 8X coverage.
- After quality filters, 14,328,987 SNPs detected from NGS data, 35,656 in common with the chips.



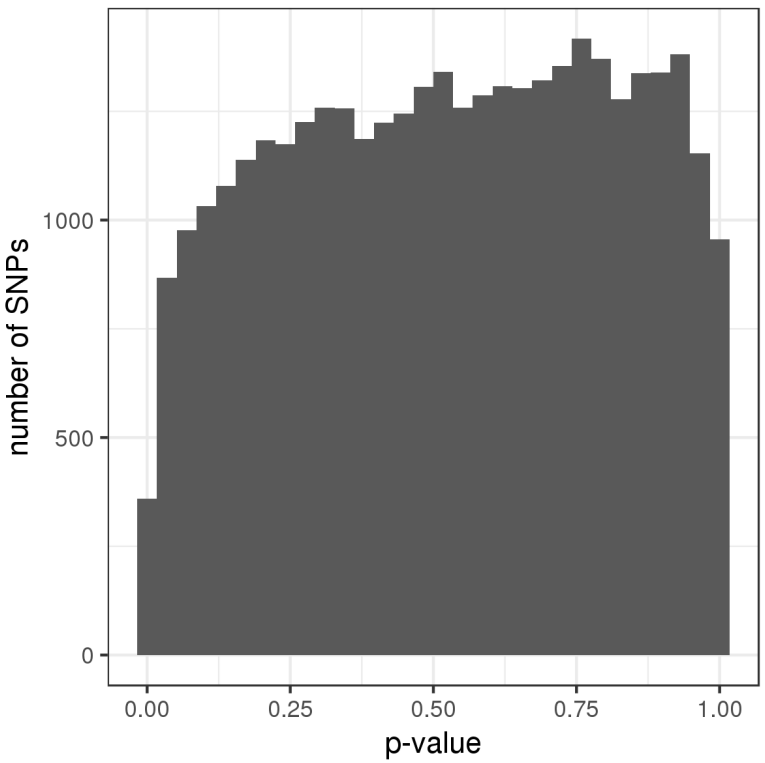
Final time series dataset

- Generation time of 4 years.
- Inbred or related animals removed.

Generation	1	2	3	4	5	6	7	8	9
First year	1980	1984	1988	1992	1996	2000	2004	2008	2012
Initial sample size	1	5	11	15	21	40	38	12	10
Final sample size	0	4	8	13	17	28	29	9	9

Selection signatures: time series approach

- Population size estimated with R package NB (Hui and Burt, 2015): $N = 400$ diploids.
- No evidence of selection at the single SNP level.
- 5 regions enriched in low p-values using a local score approach (Fariello *et al*, 2017).



Candidate regions

- Candidate genes related to carcass and meat traits (RBPMS2, OAZ2) or milk traits (ARFIP1).

Chr	Start (bp)	End (bp)	Length (bp)	Nb SNP	Genes
10	45387461	45564676	177215	7	RBPMS2, OAZ2 ZNF609, RF00413
13	41414256	41529941	115685	3	-
17	4675045	4750693	75648	4	FHDC1, ARFIP1
17	31268164	31632465	364301	8	-
22	39414833	39491373	76540	3	PTPRG

Selection signatures: present time data approach

- Signatures of historical selection detected from NGS data only (15 animals from generations 8 and 9).
- nSL statistic (Ferrer-Admetlla *et al*, 2014): looks for long haplotypes at high frequency.
- No proper p-value associated to a given nSL score: outlier approach.
- Focus on SNPs in common with the chip data: 5 regions with 'outlier' p-values.

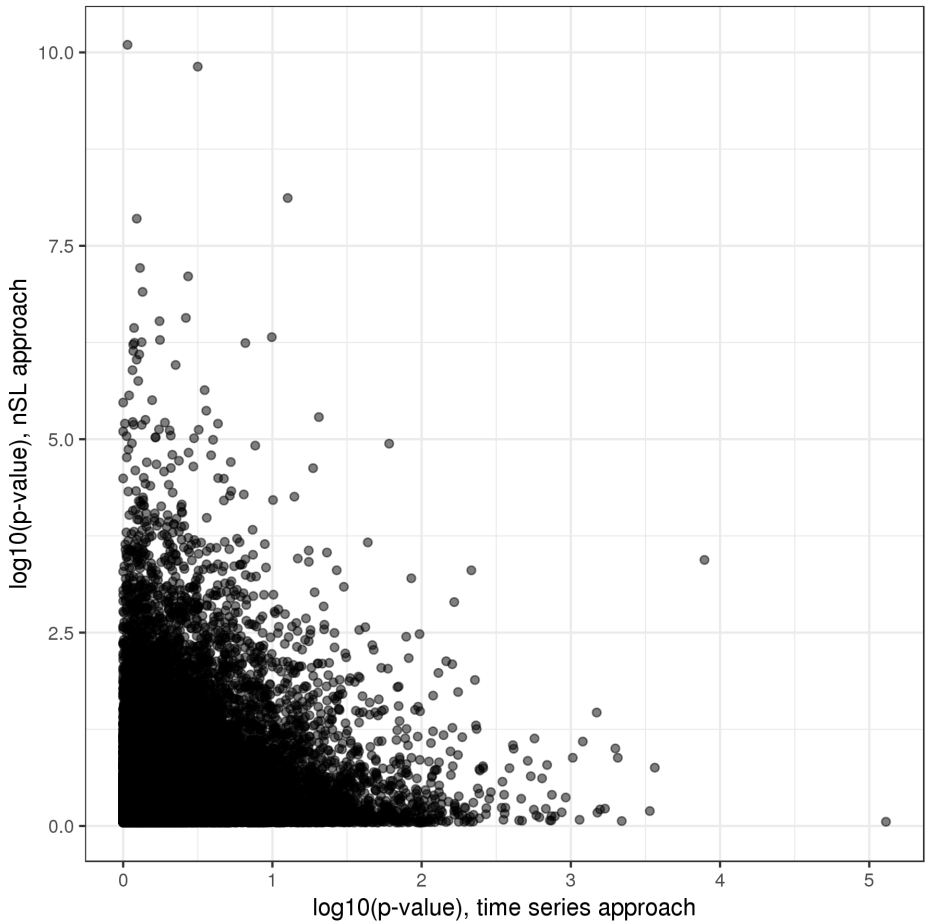
Selection signatures: present time data approach

Chr	Start (bp)	End (bp)	Nb SNP	log10(pval)	Genes
2	7169804	7270116	2	9.82	COL5A2, COL3A1
2	8476975	8476975	1	8.12	-
6	55360713	55360713	1	7.85	-
10	98290813	98290813	1	10.10	FLRT2
25	13647777	13647777	1	7.10	PARN, BFAR, PLA2G10

- One region close to the MSTN gene (Chr2, $\approx$ 6,2Mb), whose allele nt821(del11) associated to double muscling has been selected in Asturiana (Dunner *et al*, 2003).
- FLRT2: embryonic development, associated to calf birth weight in Holstein (Cole *et al*, 2014).

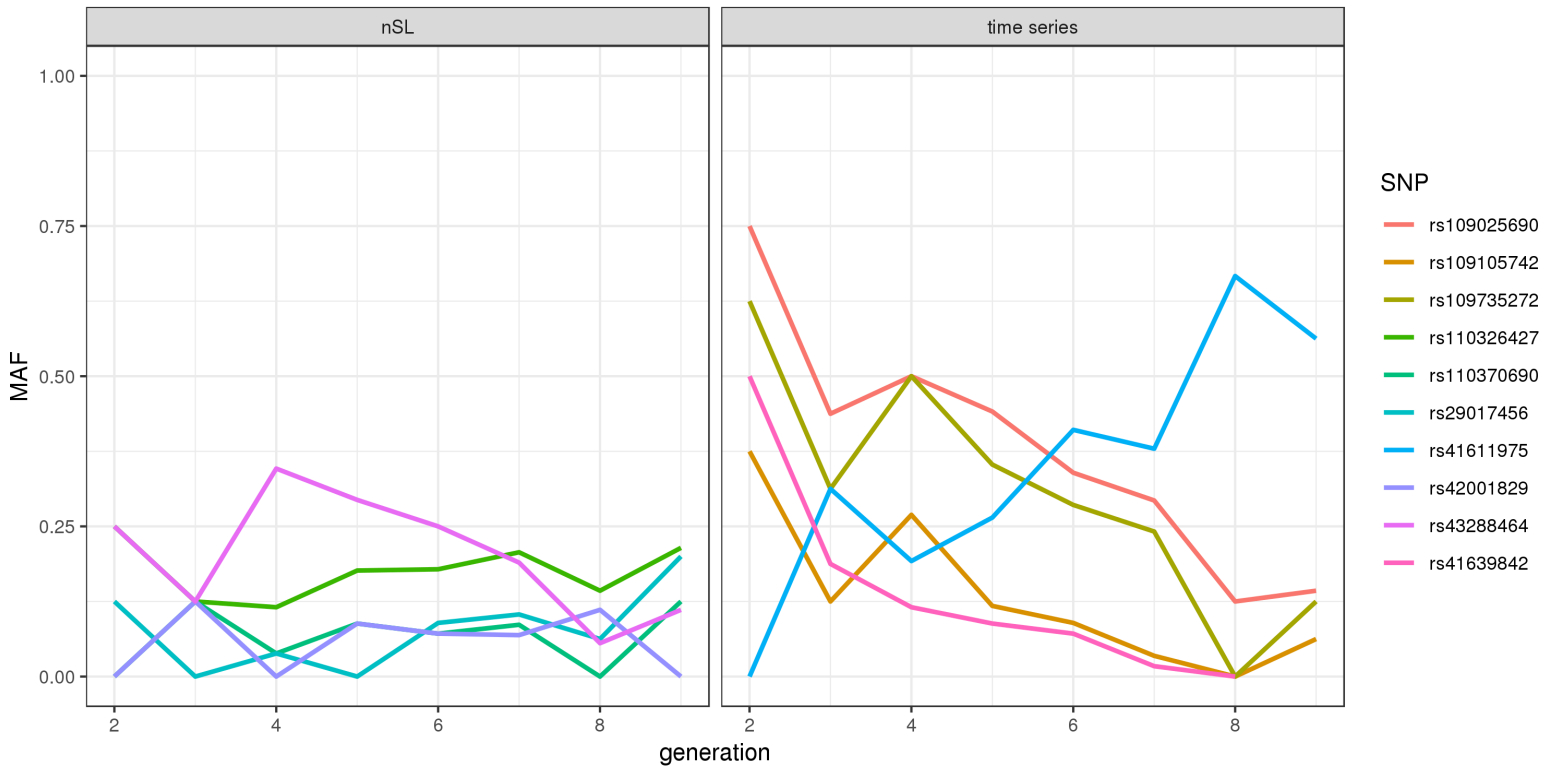
Comparison between approaches

- Almost no correlation between the p-values obtained from the time series and the nSL approach.



Comparison between approaches

- The top 5 SNPs of the time series approach show a clear shift in allele frequency since 1980.
- The top 5 nSL SNPs show a stable allele frequency since 1980, suggesting selection at these loci is older.



Outline

Data (Le Bihan-Duval *et al*, 2018)

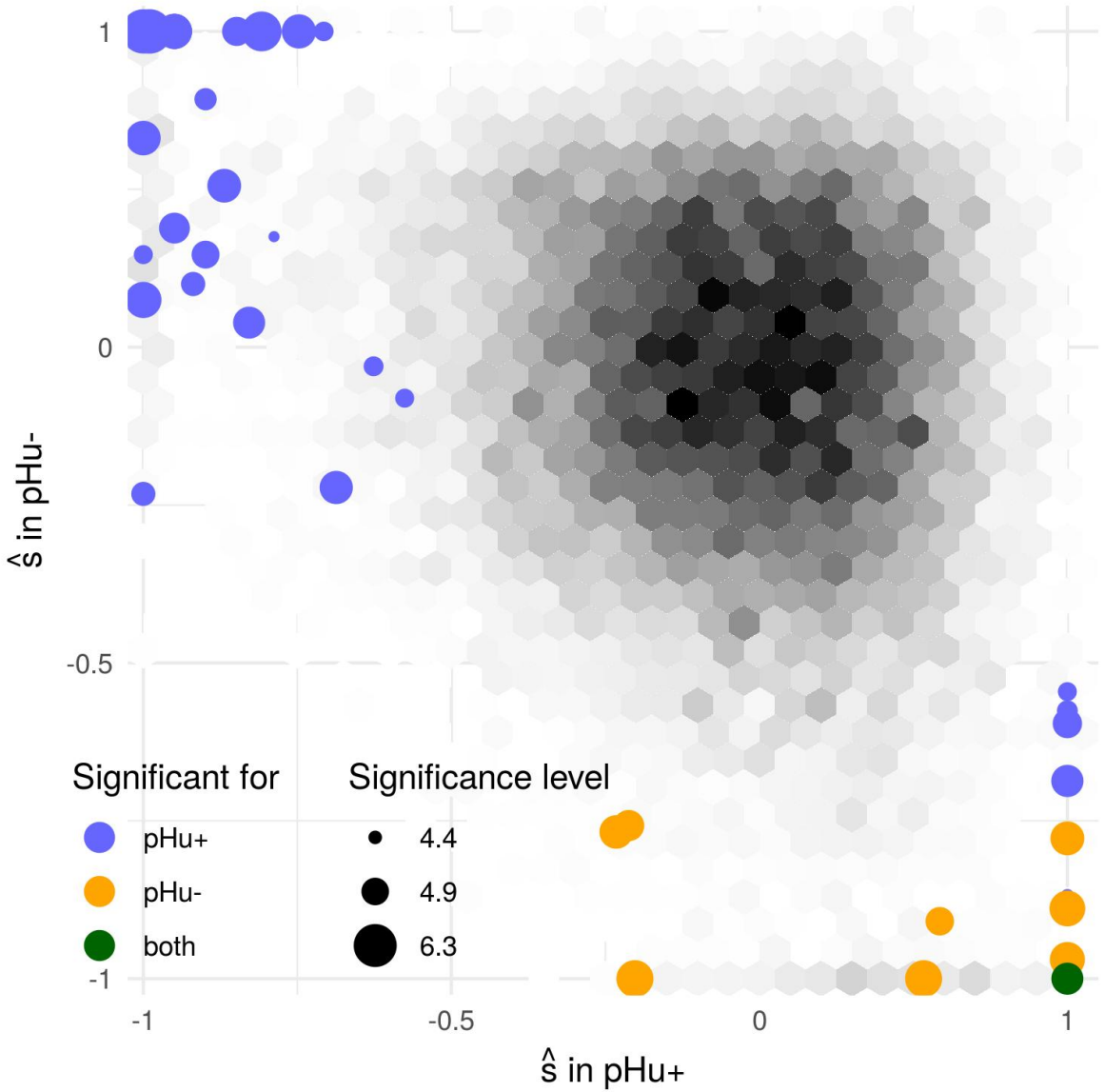
- Two chicken lines, 5 generations of divergent selection for low or high muscular ultimate pH (pHu).
- Genotypes at 40,199 SNPs.
- Haploid sample sizes:

Dates (in generations)	0	1	2	3	4	5
pHu+	102	28	42	56	54	60
pHu-	102	34	46	44	54	56

Results

- Estimated population sizes: $N = 157$ haploids in pHu+ and $N = 123$ with pHu-.
- 39 significant SNPs detected, at a FDR of 5%.
- 12 regions: 8 in pHu+, 1 in pHu-, 3 in both.
- Comparison with tests based on the differentiation between lines : FLK (Bonhomme *et al*, 2009) and hapFLK (Fariello *et al*, 2013):
  - 6 regions in common, 6 new.
  - New regions: allele frequency change only in one line.

Results


1
0
-0.5
-1
-1
-0.5
0
1
ŝ in pHu-
ŝ in pHu+
Significant for
pHu+
pHu-
both
Significance level
4.4
4.9
6.3

Outline

Conclusions

- Beta with Spikes: perfect fit to the WF (for the problem considered here), very good alternative for large N (computation time).
- Inclusion of fixation probabilities.
- $Ns = 10$ generally sufficient to detect selection from time series, but depends on T/N, x_0 and sampling design.
- Estimation of s difficult for $(T/N, Ns) \geq (1, 100)$.
- Time series complimentary to present time data and allow a better annotation of selection events.
- Publication in G3 (early online), code in github.

Perspectives

- Joint estimation of N and s (genome-wide data).
- Optimize the code.
- Account for variations of N and s.
- Extend to halotype data.

Acknowledgements

- IMAGE project, European Union's Horizon 2020 Research and Innovation Programme, grant agreement no. 677353.
- Haptitude project, INRA métagrogramme SelGen.
- Thomas Bataillon & Paula Tataru (BIRC, Aarhus, Danemark).
- Susana Dunner & Natalia Sevane (UCM, Madrid, Spain).
- Kenza Bazi-Kabbaj & Maria Bernard (INRA GABI, Jouy-en-Josas).
- Genotoul bioinformatics platform Toulouse Midi-Pyrénées.