Inferring demography and selection from genomic time series data

Simon Boitard

INRAE, CBGP, Montpellier, France

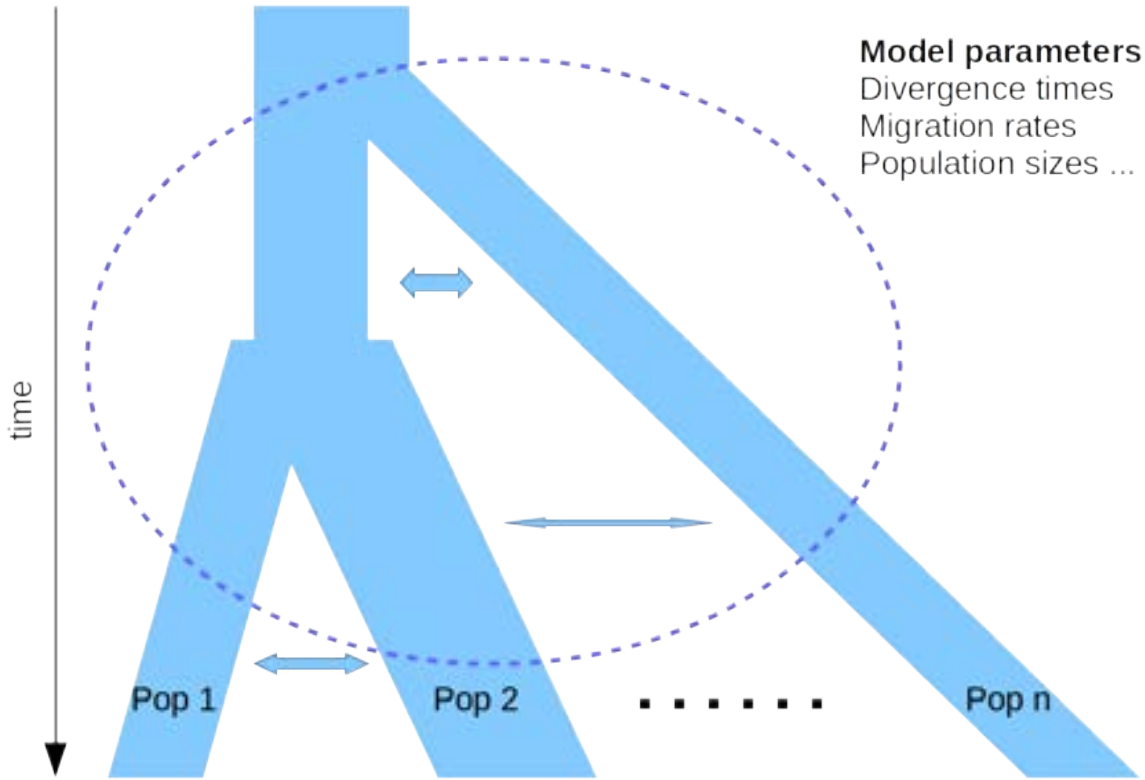
Séminaire CBGP
14th January 2025

Outline

Plan

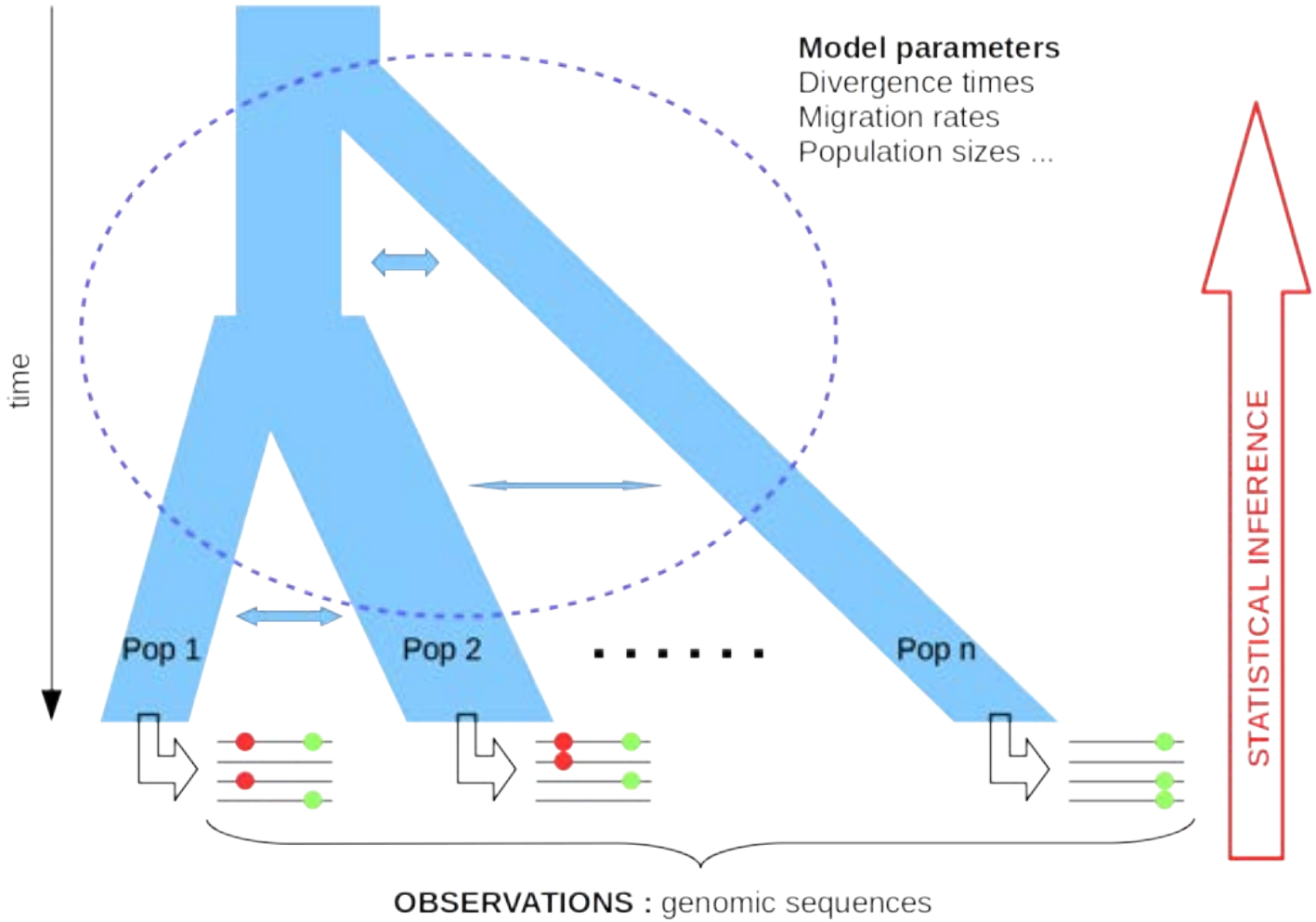
Standard Population Genetics Inference

from **molecular data sampled at a single time**.



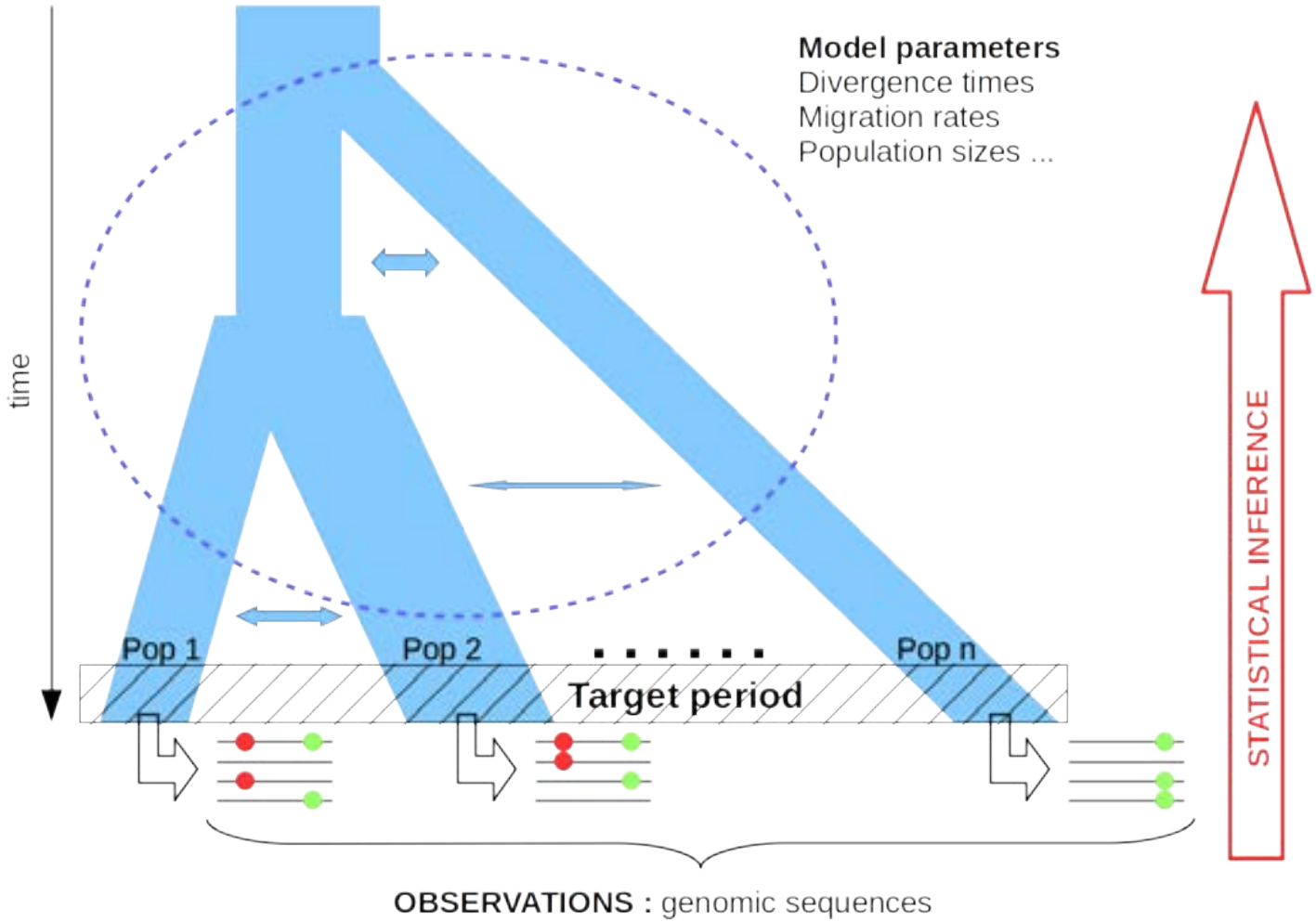
Standard Population Genetics Inference

from **molecular data sampled at a single time**.



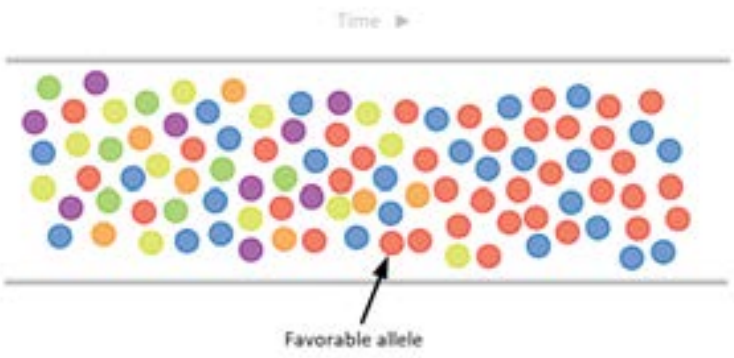
Standard Population Genetics Inference

from **molecular data sampled at a single time**.

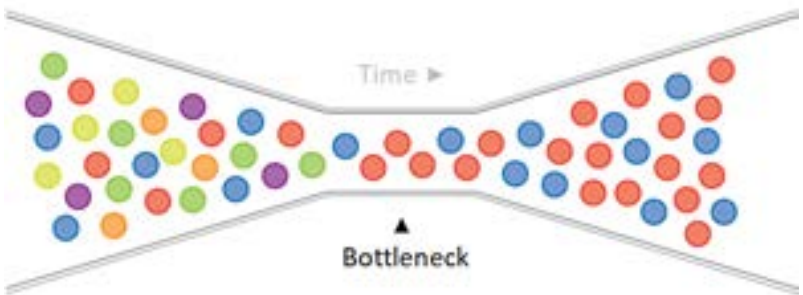


Confounding effects of demography and selection

Selective sweep

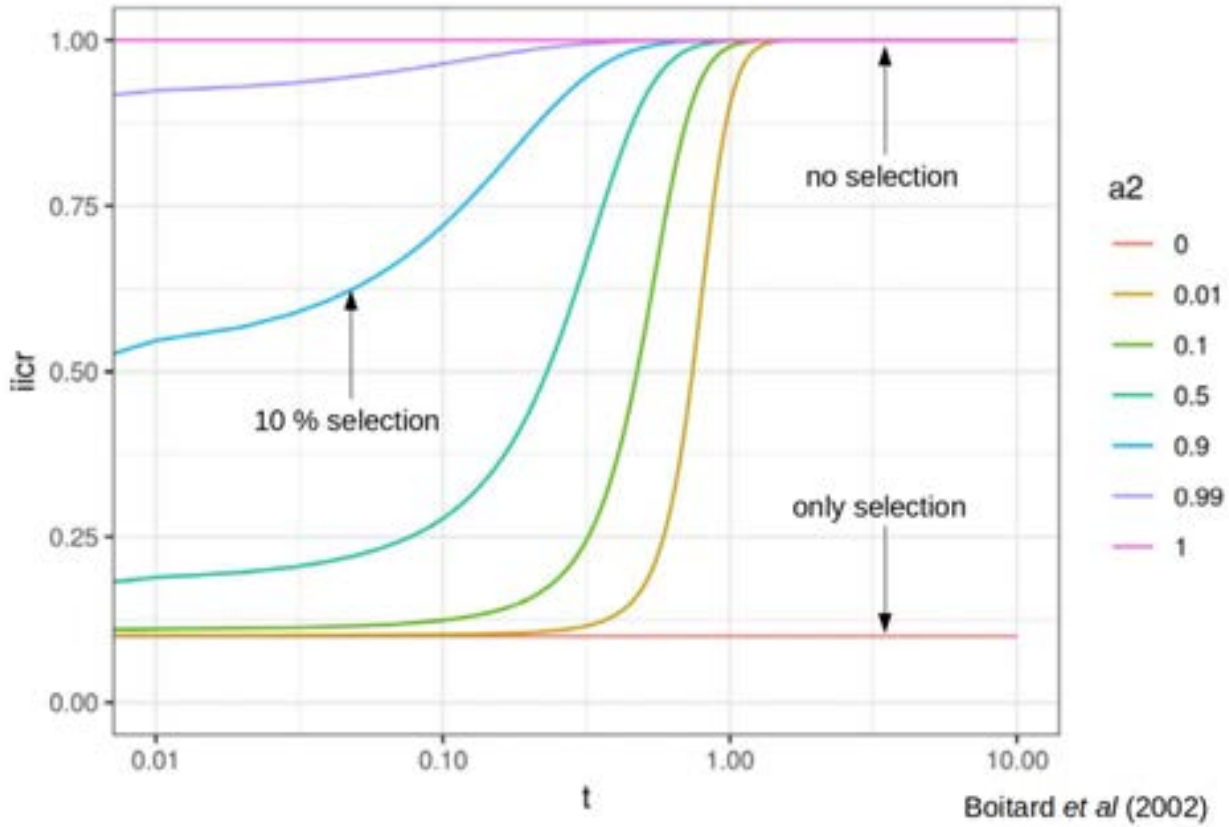


Bottleneck



Ignoring the true demography can lead to wrongly detect selection

Confounding effects of demography and selection



Ignoring selection can bias population size inference

Genomic time series

Various contexts and temporal scales:

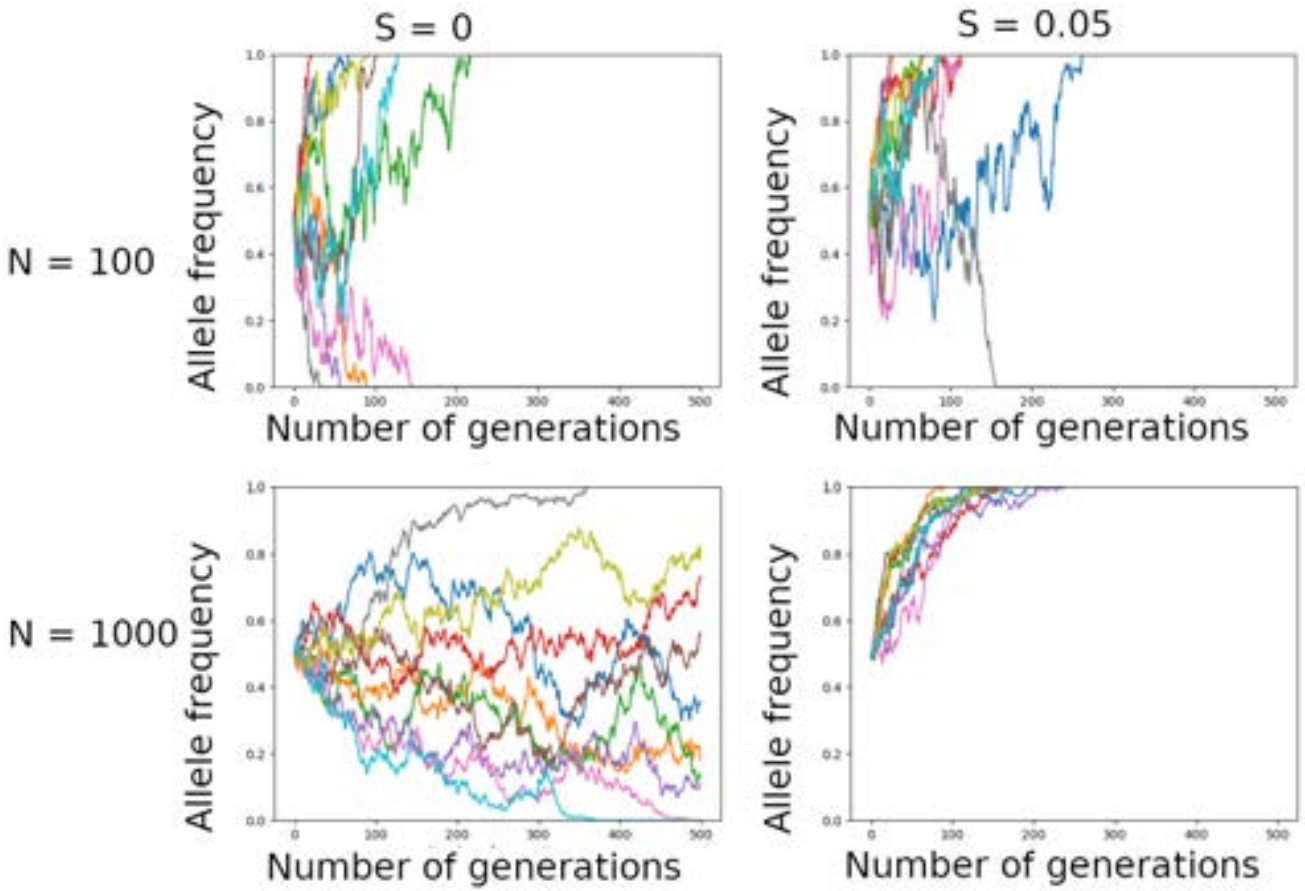
Experimental evolution

Monitoring of wild populations

Ancient DNA

Genomic time series

Temporal trajectories of allele frequencies informative about both demography and selection.



Summary

Genomic time series

- Arise in various contexts and temporal scales.
- Focus on a specific period of the evolutionary history.
- Allow (in principle!) disentangling demographic and selective effects within this period.

Plan

Acknowledgements

Time series methodology:

- Cyriel Paris & Bertrand Servin, INRAE, GenPhySE, Toulouse, France
- Miguel de Navasués & Mathieu Uhl, Paul Bunel, CBGP

Fly experiment:

- Lily Cesari, Candice Deschamps, Arnaud Estoup, Julien Foucaud, Mathieu Gautier, Emilie Mendes, Laure Olazcuagua & Nicolas Rode . . . CBGP

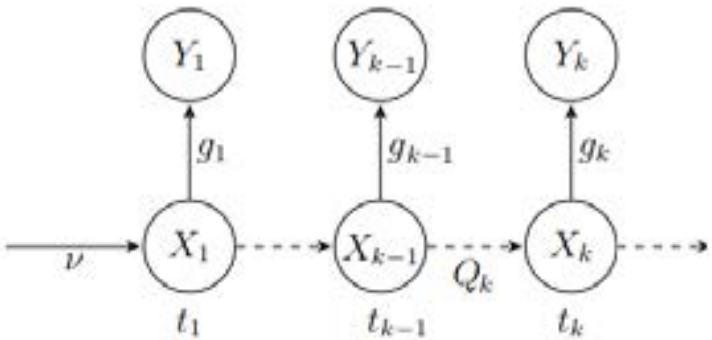
Molecular biology:

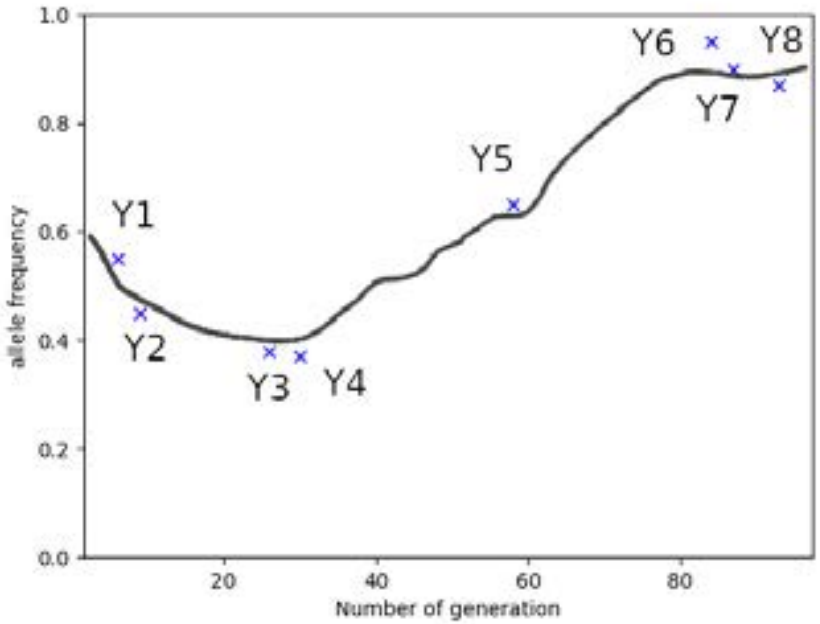
- Anne Loiseau, CBGP

Read mapping and variant calling:

- Mathieu Gautier

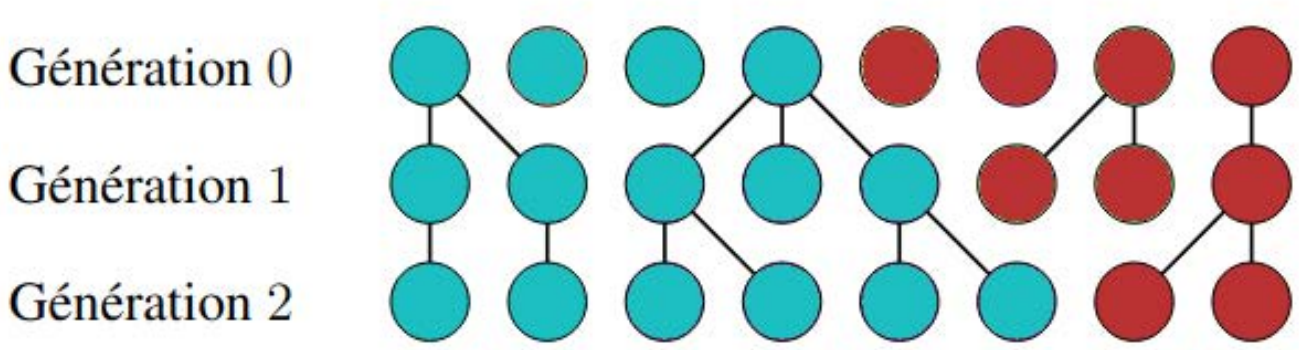
Hidden Markov Model (HMM) (Bollback 2008)





- X_k population allele frequency at time t_k (hidden)
- Y_k sampled allele frequency at time t_k (observed)
- Q_k transition matrix from time t_{k-1} to time t_k

Wright-Fisher model



- Panmictic population, constant size N, non overlapping generations
- Neutral evolution : all alleles sampled with the same probability

$$E[X_{t+1}] = X_t$$

- Selection : one allele more likely sampled due to higher fitness

$$E[X_{t+1}] = f_s(X_t)$$

HMM Transition matrix

- Depends on N, s and $t_{k-1} - t_k$.
- Example for $N = 4$:

$$Q = \begin{array}{c} \\ 0/4 \\ 1/4 \\ 2/4 \\ 3/4 \\ 4/4 \end{array} \begin{array}{c} \begin{array}{ccccc} 0/4 & 1/4 & 2/4 & 3/4 & 4/4 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0.32 & 0.42 & 0.21 & 0.05 & 0.04 \\ 0.06 & 0.25 & 0.38 & 0.25 & 0.06 \\ 0.004 & 0.05 & 0.21 & 0.42 & 0.32 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

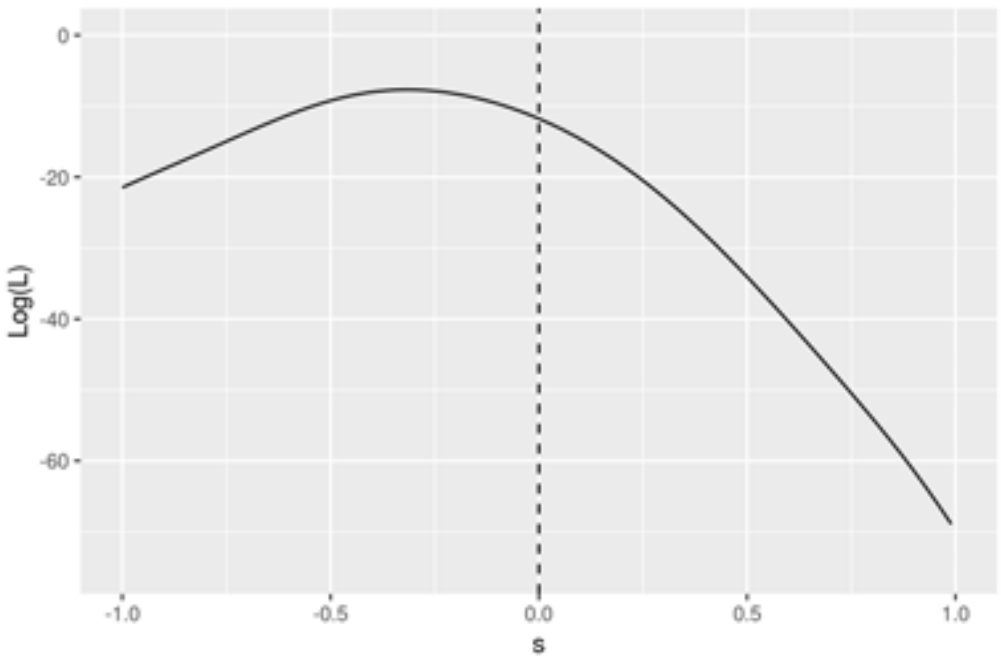
Demography and Selection Inference

Exact (and fast) computation of the likelihood

$$P(Y_1, ..., Y_n|N, s) = P(\overline{Y}|N, s) = P(\overline{Y}|Q_1(N, s), ..., Q_n(N, s))$$

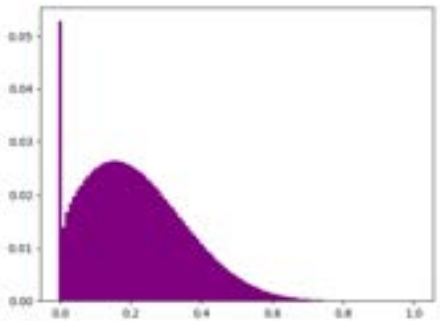
for any values of N and s

1. **Inference of N** : connsider p independent loci and optimize $P(\overline{Y}_1|N, s_1 = 0)P(\overline{Y}_2|N, s_2 = 0) \dots P(\overline{Y}_p|N, s_p = 0)$ over N.
2. **Inference of s** : for each locus i, optimize $P(\overline{Y}_i|\hat{N}, s_i)$ over s_i.

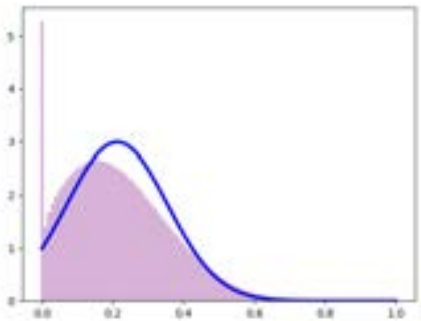


Wright-Fisher approximations

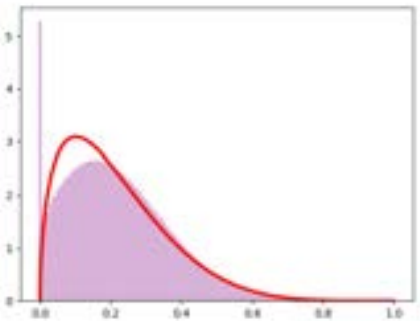
- Wright-Fisher model limited to $N \approx 500$ for numerical reasons (Q of size $N \times N$).
- Continuous approximations



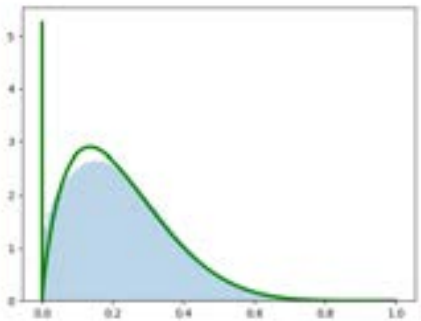
Wright-Fisher



Gaussian



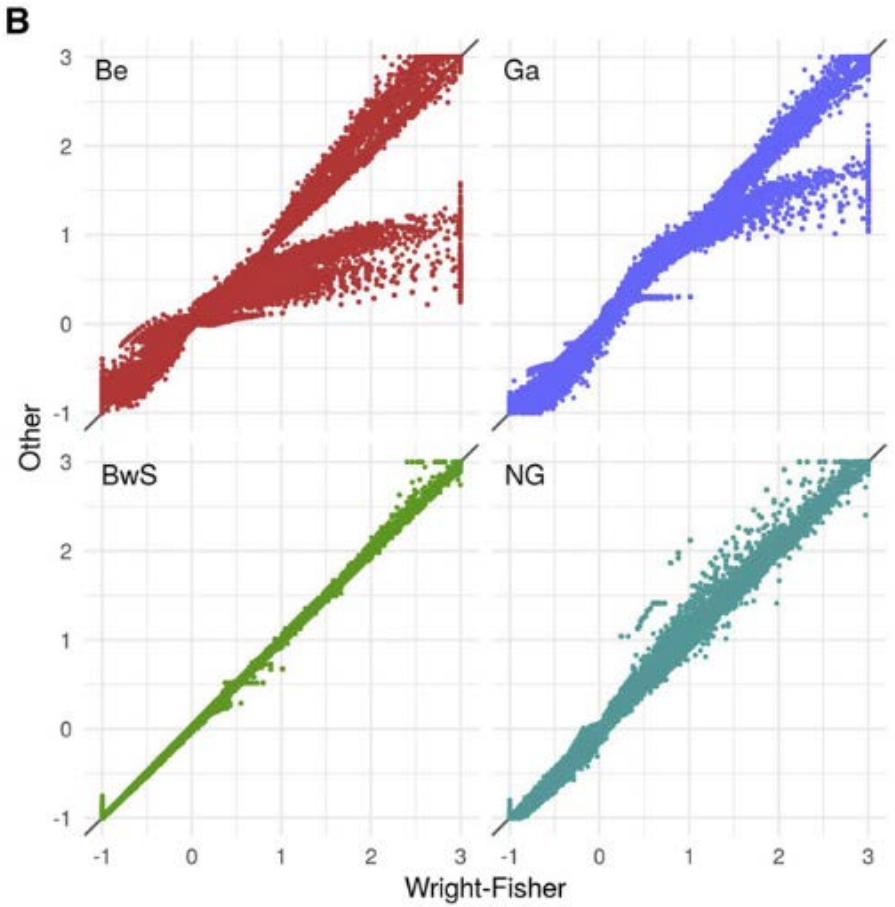
Beta



Beta with Spikes

Wright-Fisher approximations

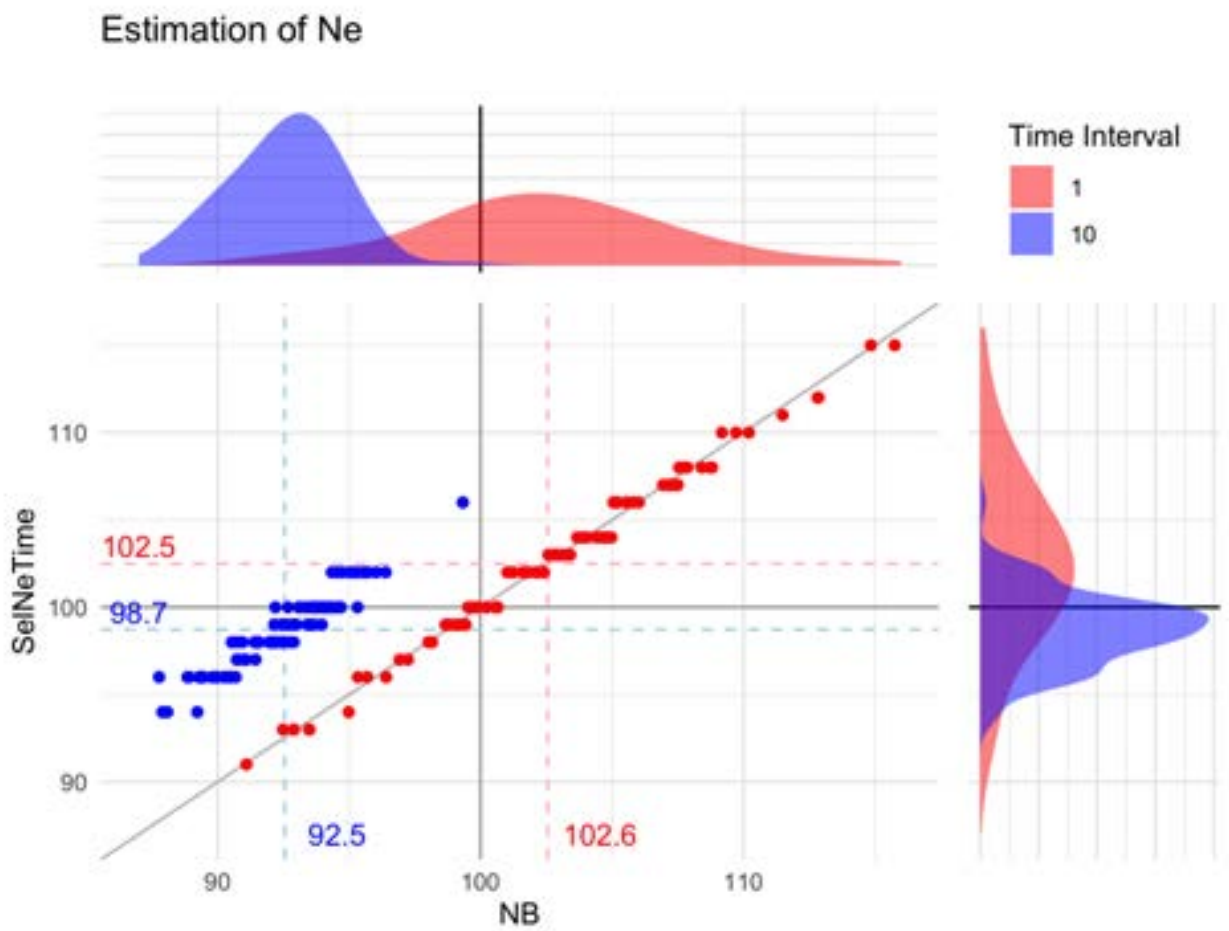
The Beta with Spikes distribution (Tataru *et al* 2019) is a very good approximation (Paris *et al*, 2019).



python package selnetime (Mathieu Uhl *et al*)

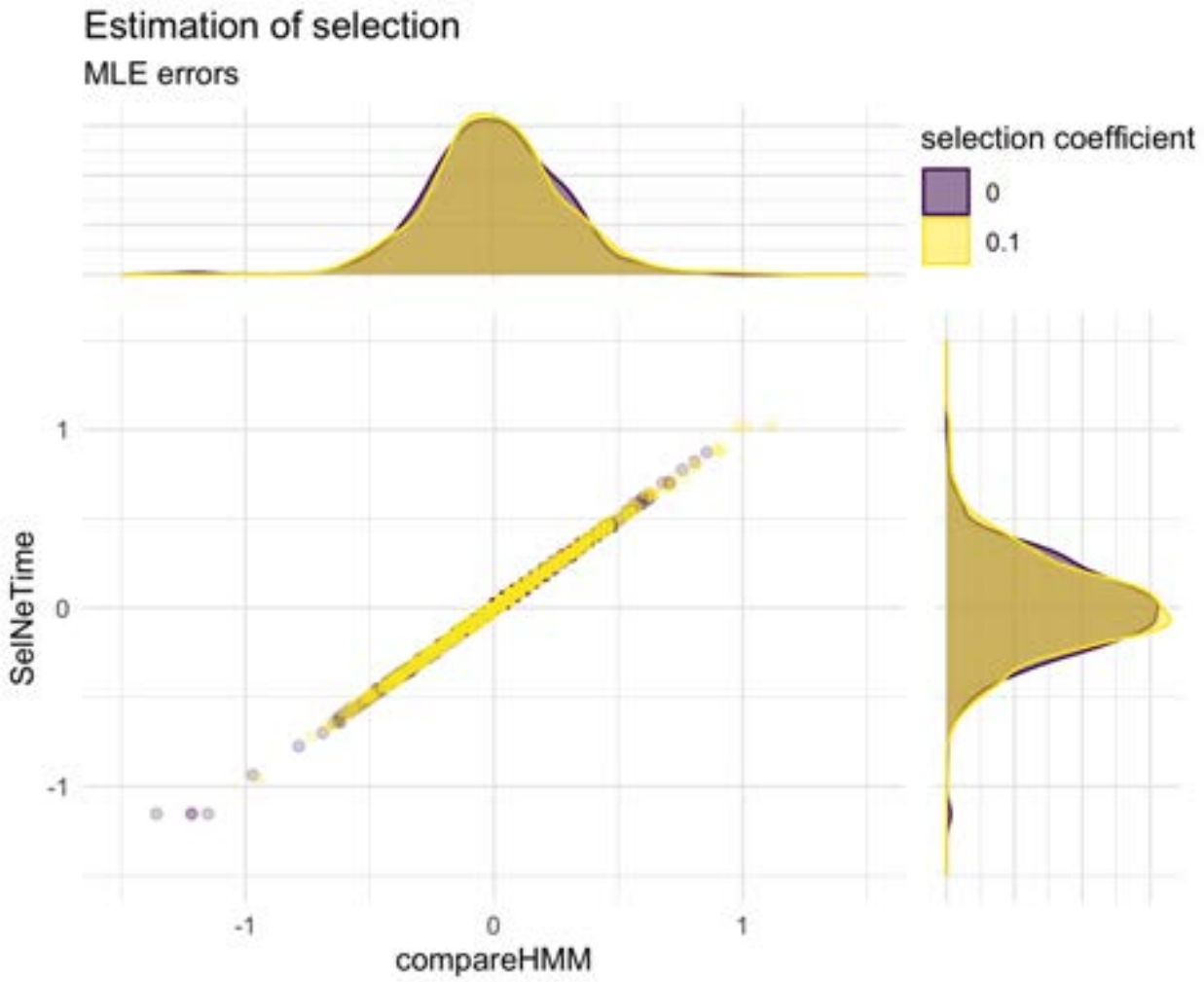
- **Models** Beta-with-Spikes and Wright-Fisher transitions.
- **Infers** N assuming $s = 0$ and / or s given N.
- **Simulate** genomic time series.
- **Install** https://pypi.org/project/selnetime/
- **Source code** https://forgemia.inra.fr/simon.boitard/snt
- **Software note** on BioRxiv, under review in PCI Math Comp Biol.

Estimation of N



- 10 sampling times, $s = 0$, $N = 100$, 1000 loci.
- Better estimation with the BwS than with the Beta model (Hui and Burt 2015) for large δ_t (blue).

Estimation of s



- $t = 1 \ldots 10$, $N = 100$, BwS model.
- Unbiased estimation of s, as in Paris *et al* (2019).

Computing time

Nb. loci	compareHMM estimation of s	SelNeTime estimation of s	SelNeTime estimation of N
100	39.27s	28.7s	6.12s
1000	360.01s	36.21s	14.23s
10000	3530.47s	96.95s	87.83s

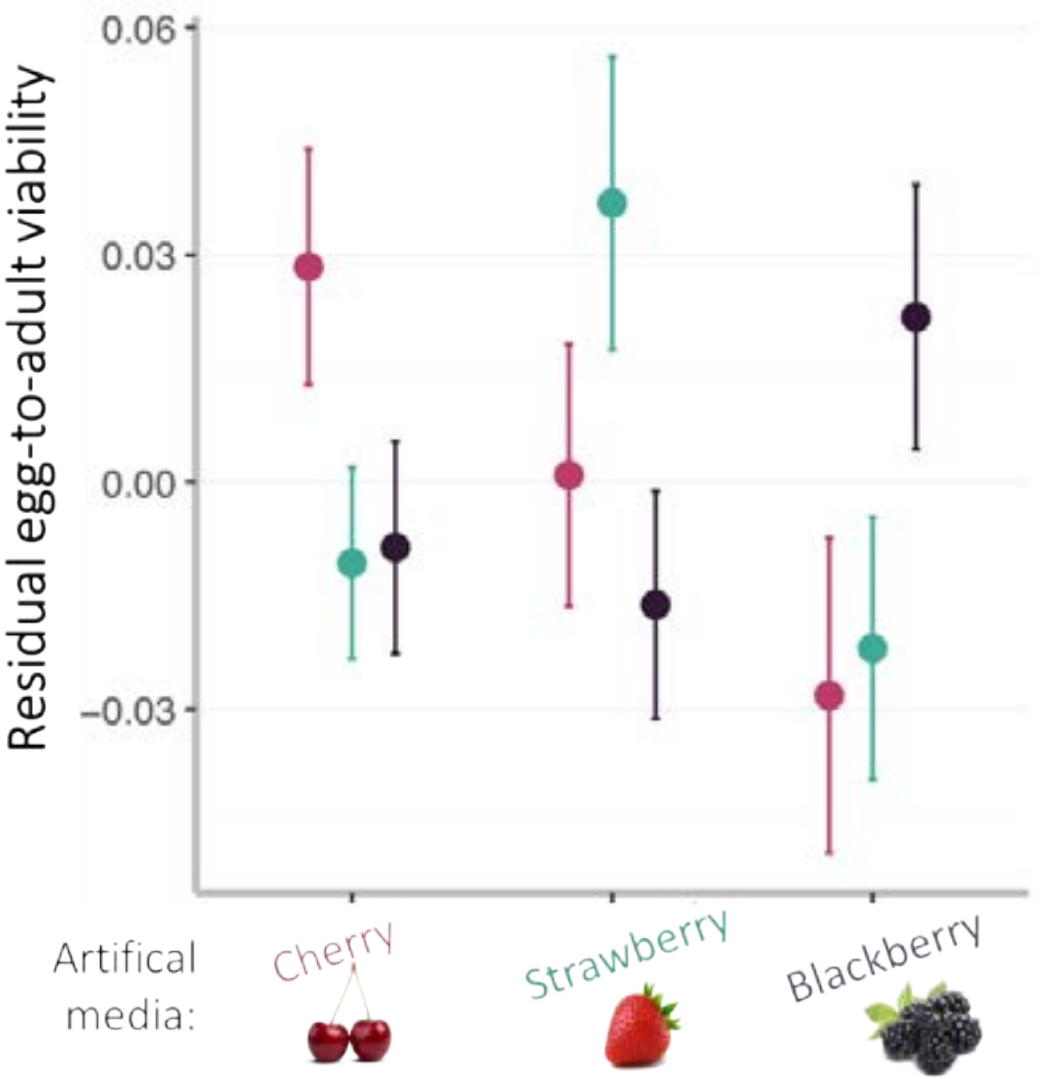
- 10 sampling times, $dt = 10$, one core.
- Fixed time to compute all transitions (28s) + 0.007s per locus.

Perspectives

- **Joint estimation** of demography and selection to avoid biases.
- Variable population size or selection intensity.
- PhD Paul Bunel (2024 - 2027, CBGP / GenPhySE).

Plan

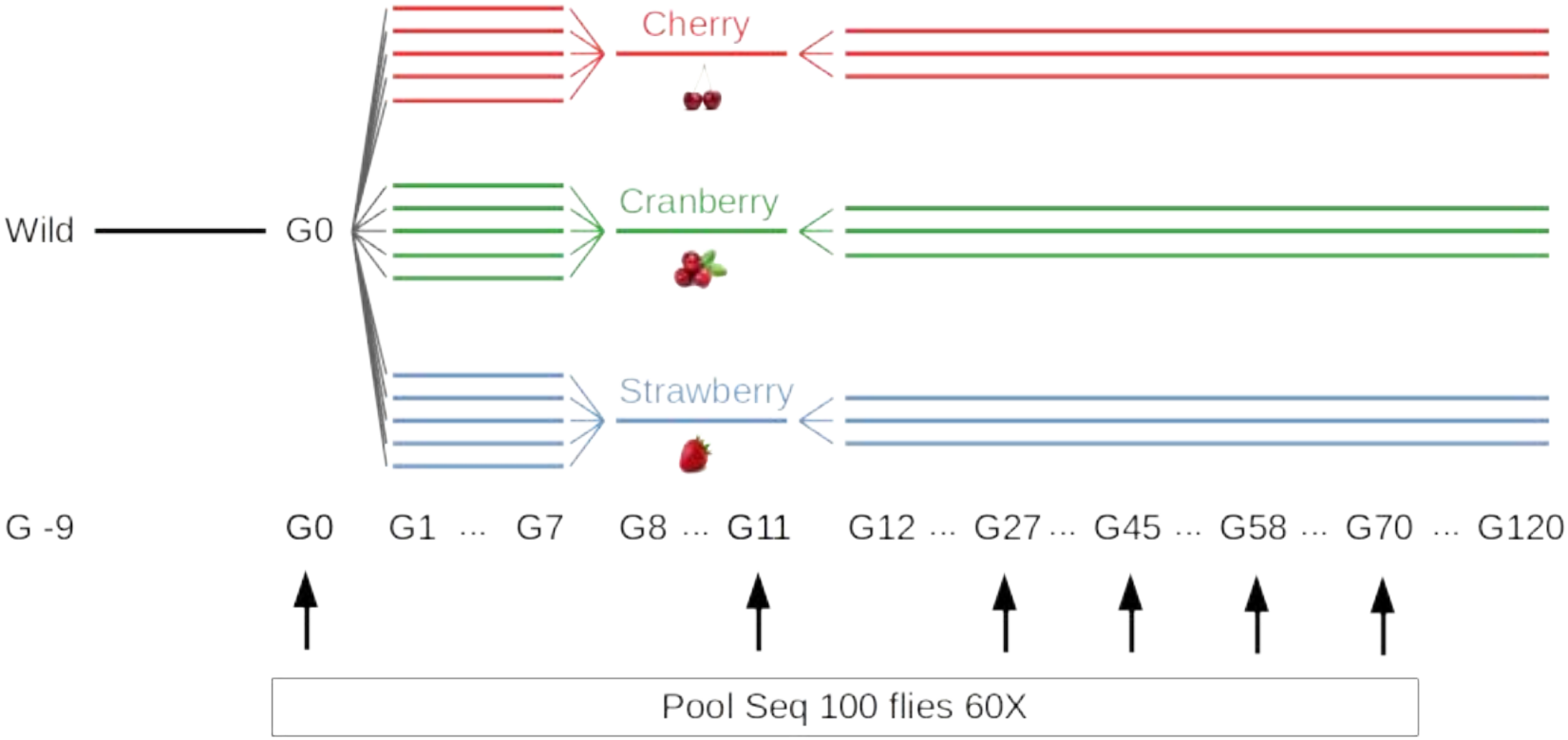
Local adaptation to host plant (Olazcuagua *et al*, 2022)



Local adaptation test:
$P = 0.005$

Fly population from:
- Cherry
- Strawberry
- Blackberry

Evolve & Resequence experiment



Acknowledgements

Time series methodology:

- Cyriel Paris & Bertrand Servin, INRAE, GenPhySE, Toulouse, France
- Miguel de Navasués & Mathieu Uhl, Paul Bunel, CBGP

Fly experiment:

- Lily Cesari, Candice Deschamps, Arnaud Estoup, Julien Foucaud, Mathieu Gautier, Emilie Mendes, Laure Olazcuagua & Nicolas Rode . . . CBGP

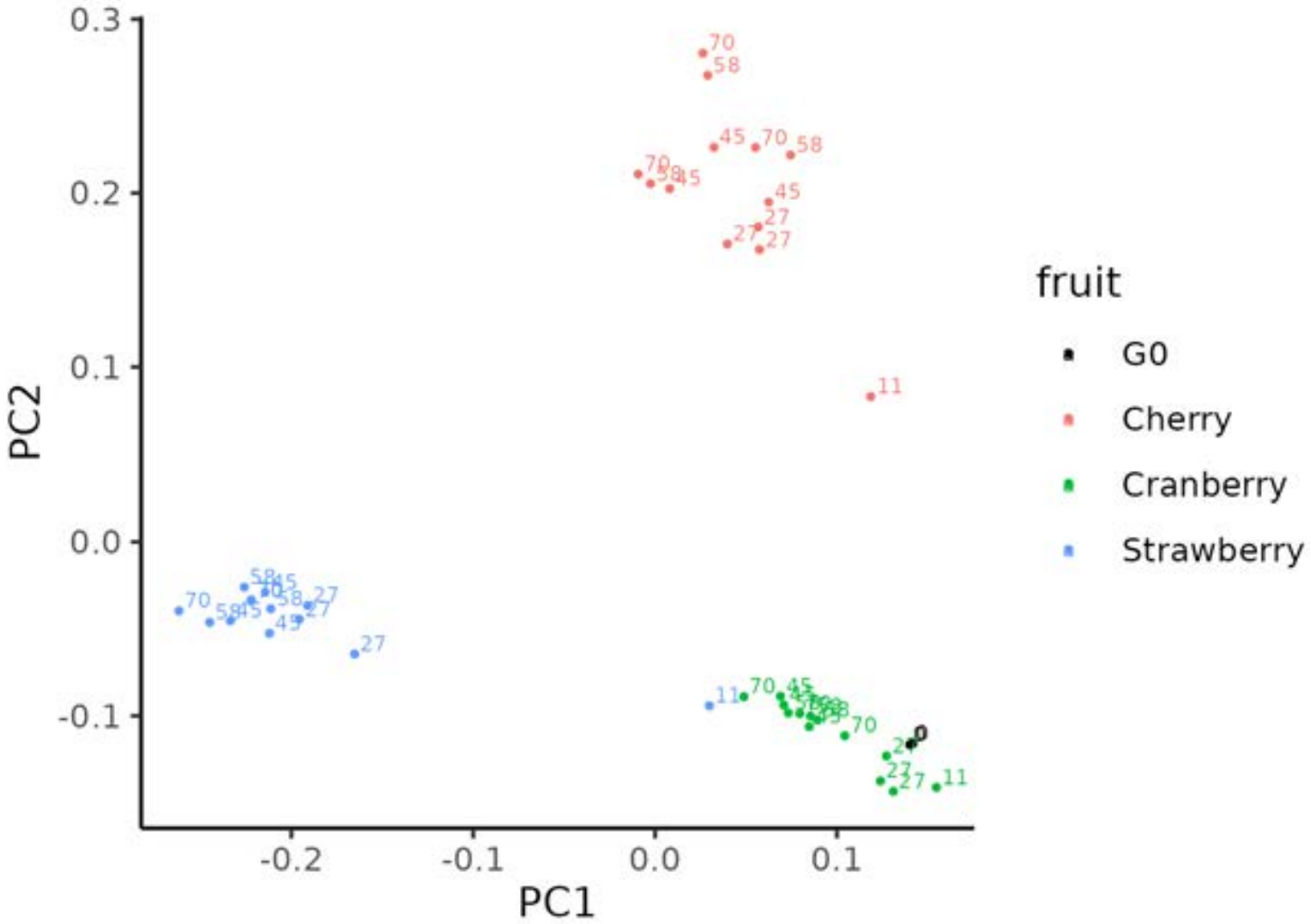
Molecular biology:

- Anne Loiseau, CBGP

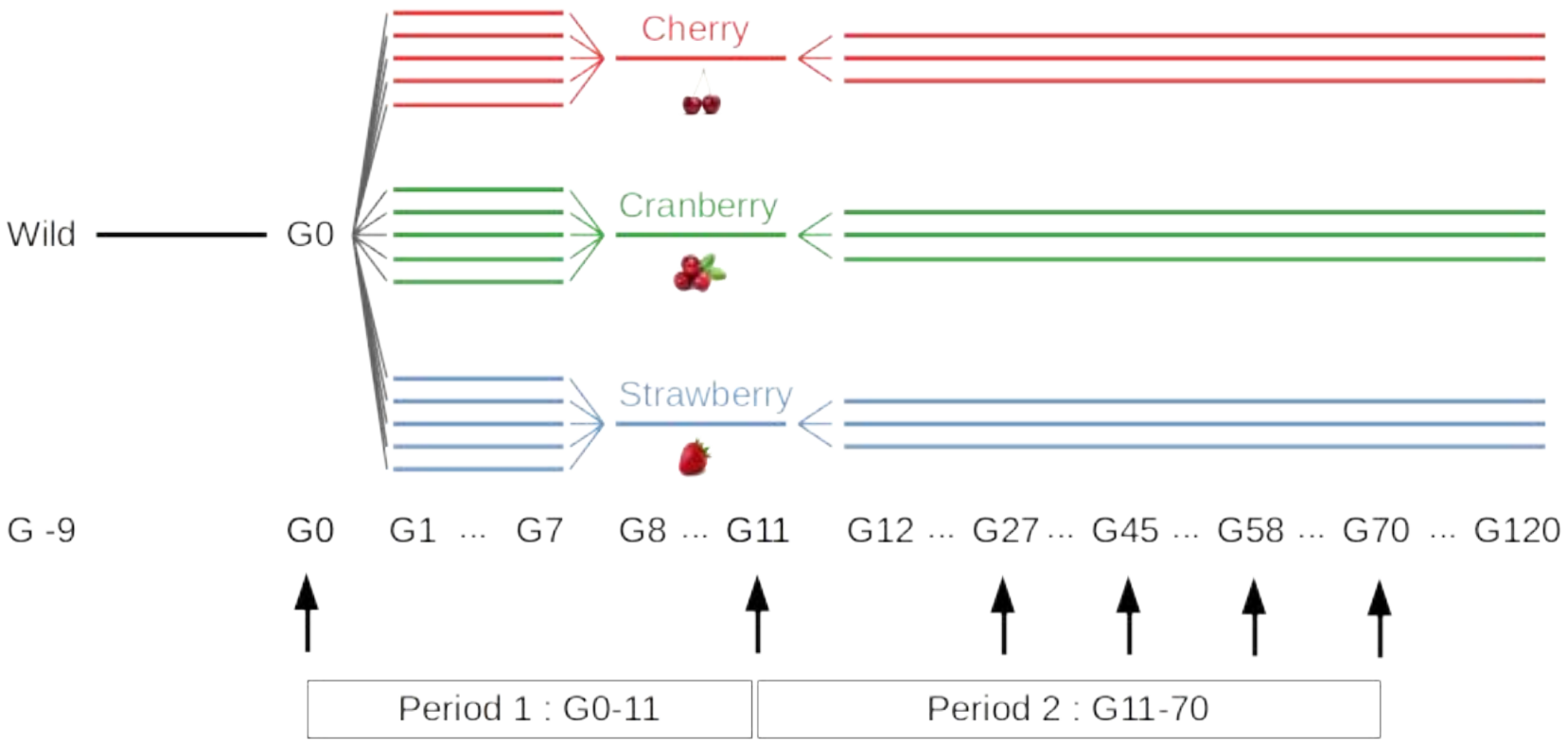
Read mapping and variant calling:

- Mathieu Gautier

Genetic diversity structuring

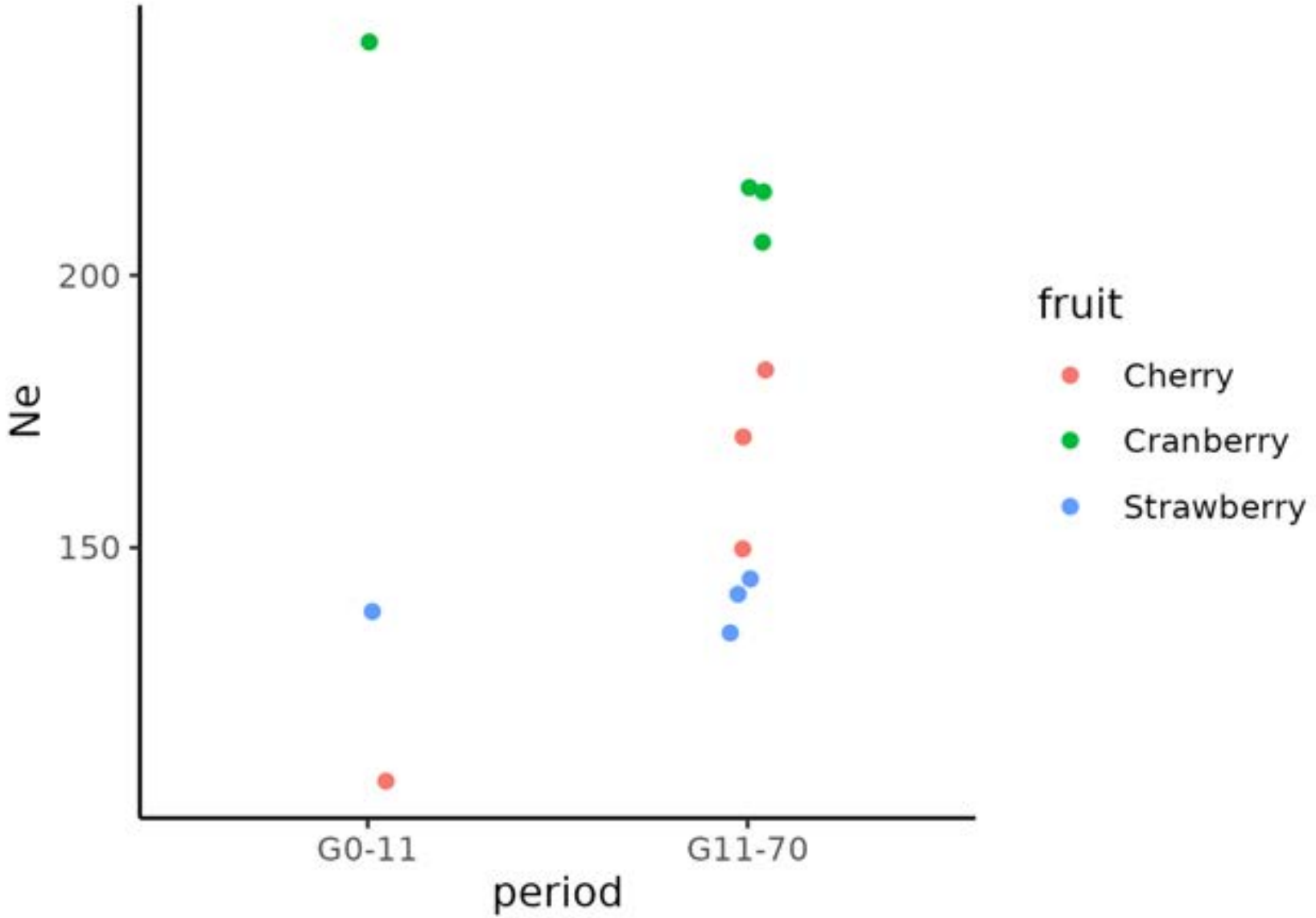


Analysis of Evolve & Resequence data

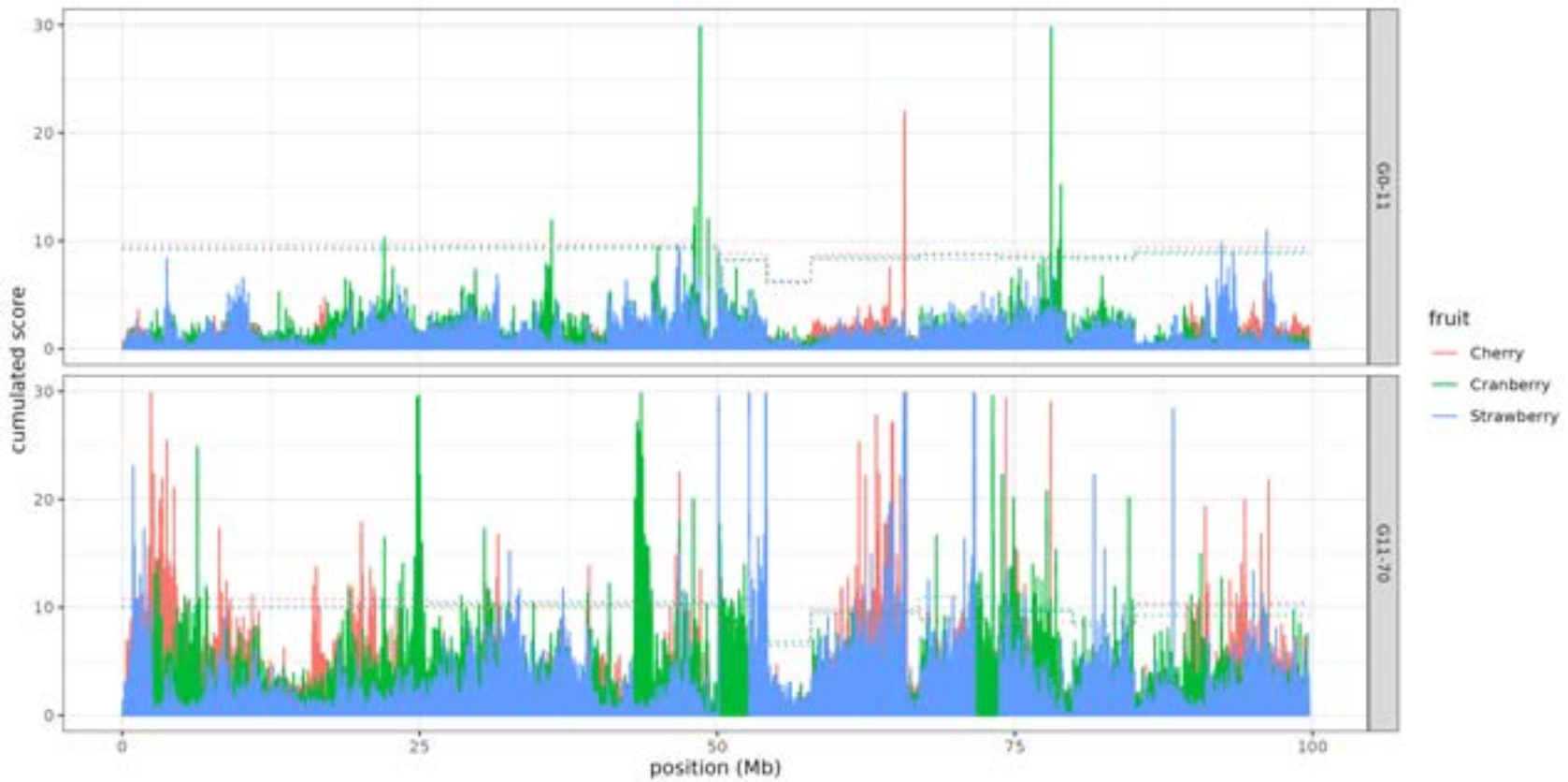


3 lines in period 1, 9 lines (3 per fruit) in period 2

Inferred N

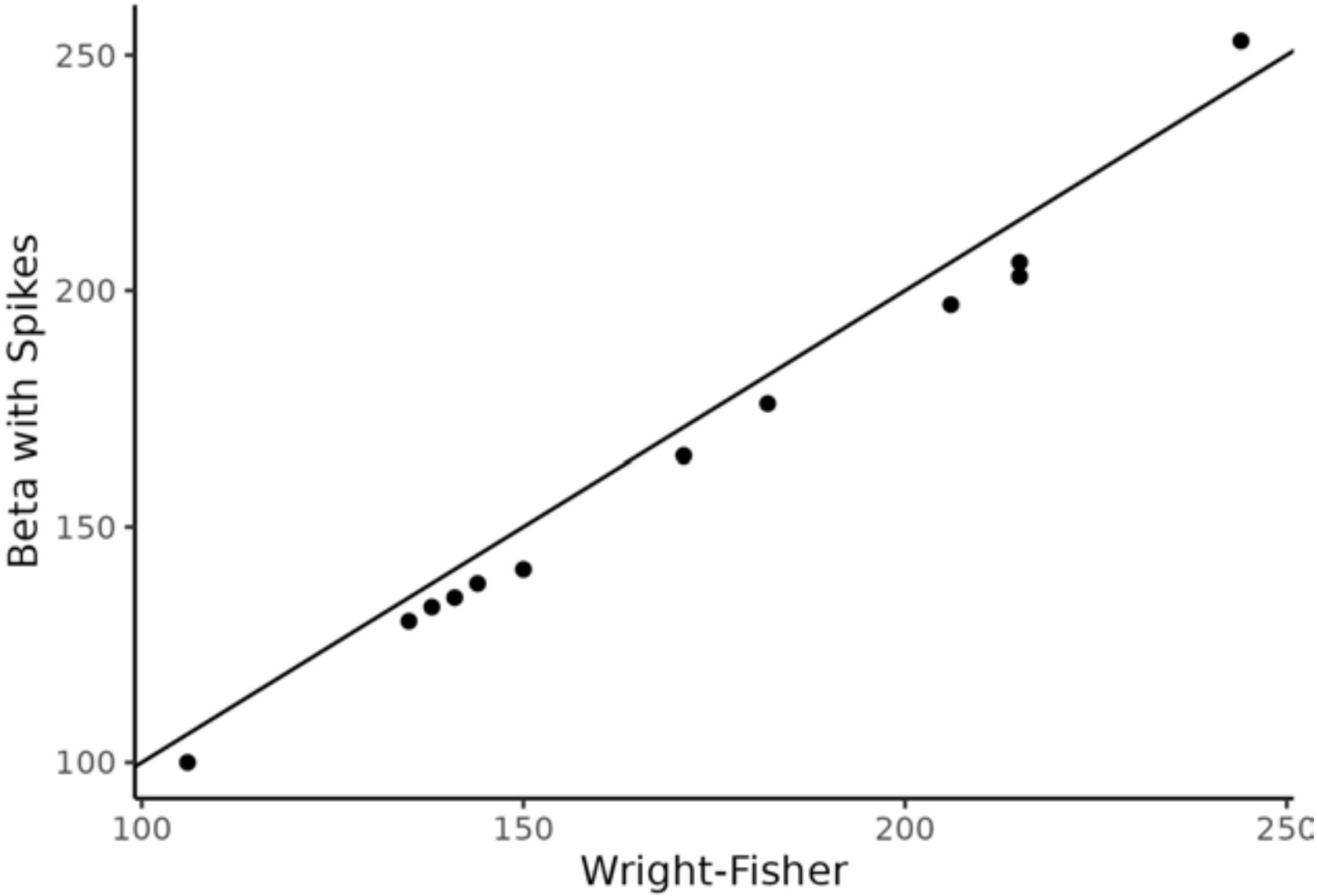

200
150
Ne
G0-11
G11-70
period
fruit
Cherry
Cranberry
Strawberry

Candidate regions under selection ($\hat{s} \neq 0$)



p-values obtained from the HMM and 'cumulated' using a local score approach (Fariello *et al*, 2017).

Inferred N


Beta with Spikes
250
200
150
100
100
150
200
250
Wright-Fisher

Perspectives

- Explore candidate regions, especially those that are specific to one single fruit.
- Compare with candidate regions detected on wild populations PoolSeq data from different fruits.

Acknowledgements

Time series methodology:

- Cyriel Paris & Bertrand Servin, INRAE, GenPhySE, Toulouse, France
- Miguel de Navasués & Mathieu Uhl, Paul Bunel, CBGP

Fly experiment:

- Lily Cesari, Candice Deschamps, Arnaud Estoup, Julien Foucaud, Mathieu Gautier, Emilie Mendes, Laure Olazcuagua & Nicolas Rode . . . CBGP

Molecular biology:

- Anne Loiseau, CBGP

Read mapping and variant calling:

- Mathieu Gautier