Heterogeneity in effective size across the genome: effects on the Inverse Instantaneous Coalescence Rate (IICR) and implications for demographic inference under linked selection

Simon Boitard, Armando Arredondo, Lounès Chikhi, Olivier Mazet

Heterogeneity in effective size across the genome: effects on the Inverse Instantaneous Coalescence Rate (IICR) and implications for demographic inference under linked selection

Simon Boitard[1], Armando Arredondo[2],
Lounès Chikhi[3,4] & Olivier Mazet[2]

1: INRAE, Centre de Biologie et de Gestion des Populations (CBGP), Montpellier
2: INSA, Institut de Mathématiques de Toulouse (IMT), Toulouse
3: CNRS, Evolution et Diversité Biologique (EDB), Toulouse
4: Instituto Gulbenkian de Ciência (IGC), Oeiras, Portugal

Rencontres Alphy / AIEM, March 14-16, 2022

Context

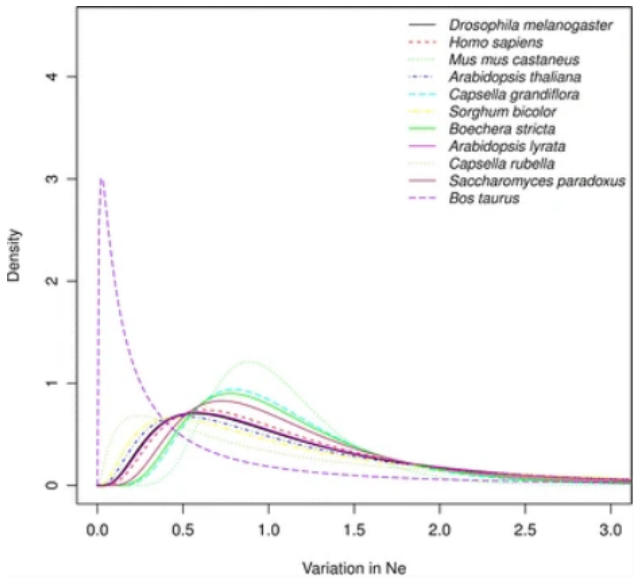
- **Genetic diversity** shaped by
  - Genetic Drift (Kimura, 1968; 1983).
  - Population demography and structure (Nei *et al*, 1974; Wakeley, 1999; Wall *et al*, 2002).
  - **Linked selection** (Hill and Robertson, 1966), **including ...**
  - Selective sweeps (**SW**; Maynard Smith and Haigh, 1974; Gillespie, 1991).
  - Background selection (**BGS**; Charlesworth *et al*, 1993).
  - Balancing selection (**BalS**; Ford 1975).
- **Respective role of these forces**? (Kern and Hahn 2018; Jensen *et al* 2019).

Linked selection and demographic inference

- Linked selection **pervasive** (Elyashiv *et al*, 2016; Pouyet *et al*, 2018).
- Linked selection **biases demographic inference** (Ewing and Jensen, 2016; Schrider *et al*, 2016; Pouyet *et al*, 2018; Johri *et al* 2021)

The variable N_e hypothesis

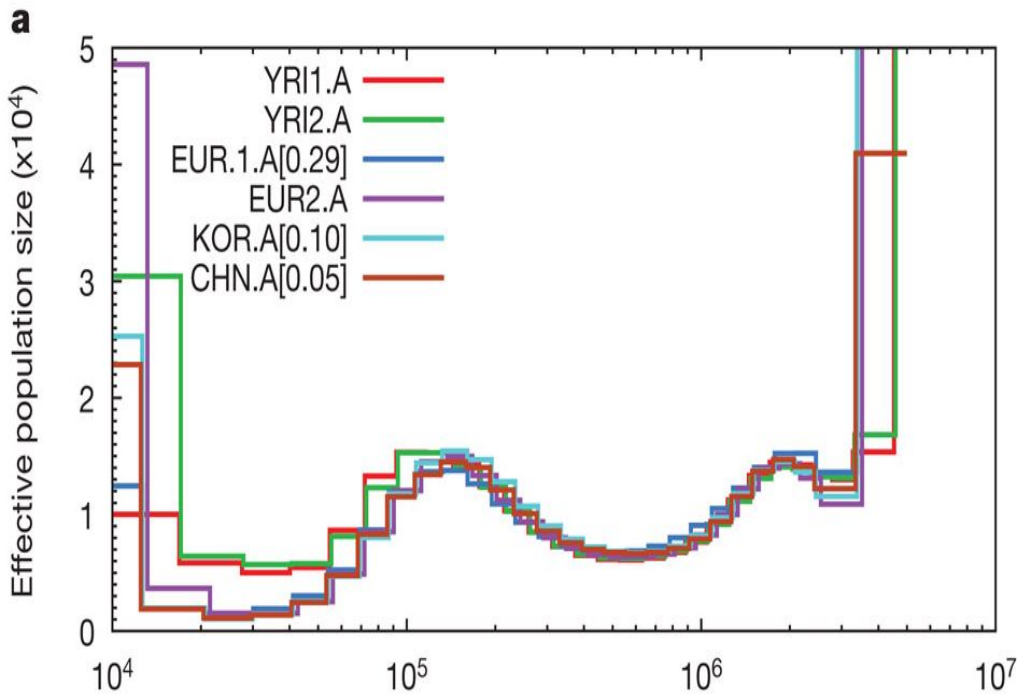
- Linked selection modelled by a change of N_e (Hill and Robertson, 1966; Charlesworth, 2009):
  - decreasing for BGS and SW.
  - increasing for BalS.
- **Variable levels of N_e genome-wide** (Gossmann *et al*, 2011; Jimenez-Mena *et al*, 2016) reflecting variations of selection constraints.



Objectives

Characterize the effects of linked selection on genome-wide patterns of genetic diversity.

- Model linked selection through several classes of N_e.
- Study the genome-wide distribution of coalescence times (T_2).
- Consequences for demographic inference with PSMC (Li and Durbin, 2011) or MSMC (Schiffels and Durbin, 2014).



Inverse Instantaneous Coalescence Rate (IICR)

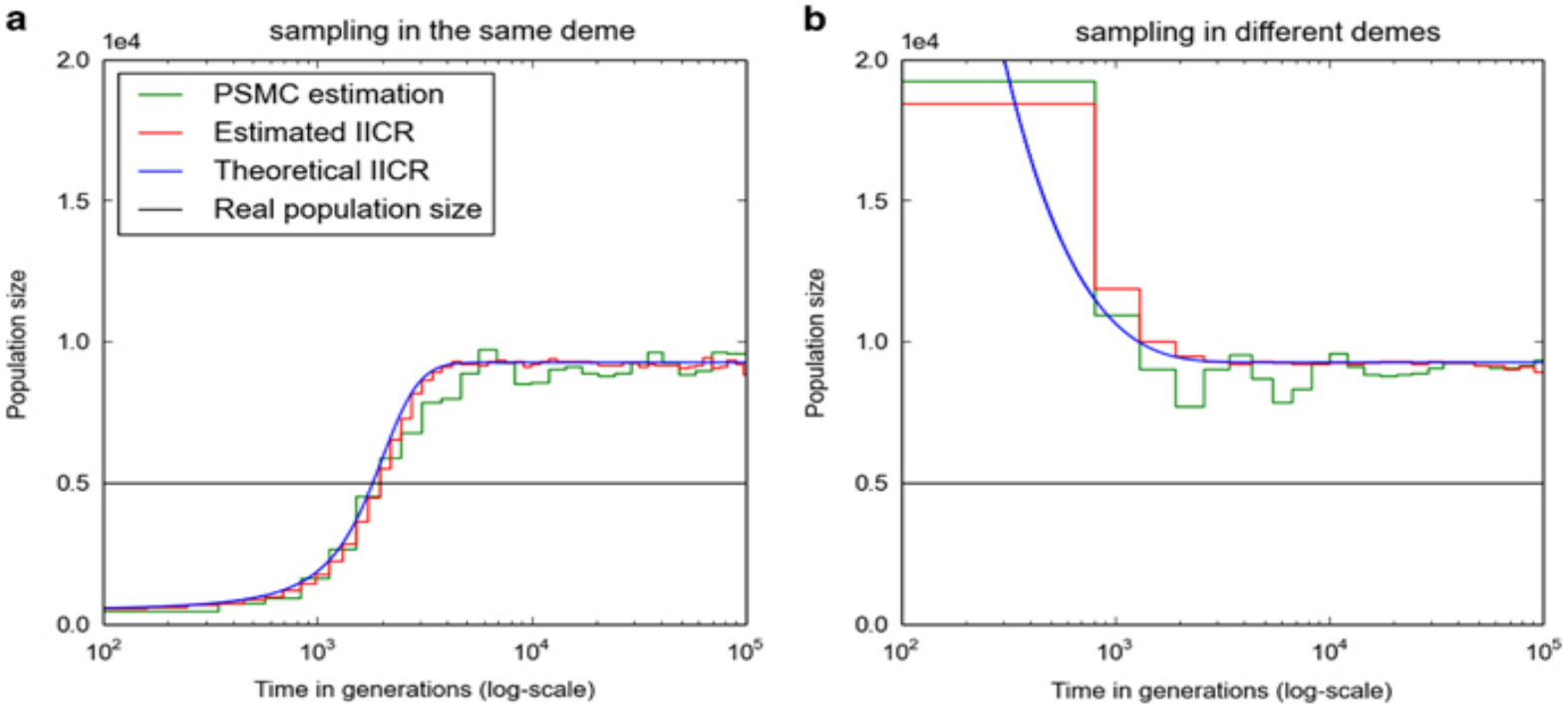
- Arbitrary evolution model.
- T_2 coalescence time between two haploid sequences, in units of $2N$ generations.
- Let $F(t) = \mathbb{P}(T_2 \leq t)$, and $F'(t) = f(t)$.
- $$IICR(t) = \frac{1 - F(t)}{f(t)} \quad \text{(Mazet } \textit{et al}\text{, 2016)}$$

IICR properties

- IICR $\Leftrightarrow$ T_2 distribution.
- Can be obtained numerically (Rodriguez *et al*, 2018) for a given evolution model.
- Convenient because :
  - inferred from genomic data by PSMC or MSMC.
  - direct interpretation **under panmixia and neutrality** : population size history.

Population structure

- PSMC does estimate the IICR but ...
- ... the IICR does not reflect population size history.



Mazet *et al* (2016), Chikhi *et al* (2018).

Panmictic model with "selection"

- K genomic classes with relative proportion a_i.
- Class i evolves under the WF model with λ_iN diploids.

$$T_2^i \sim \mathcal{E}(\mu_i), \qquad \mu_i = 1/\lambda_i$$

- Genome-wide distribution of T_2

$$f(t) = \sum_{i=1}^{K} a_i \mu_i e^{-\mu_i t}$$

- IICR

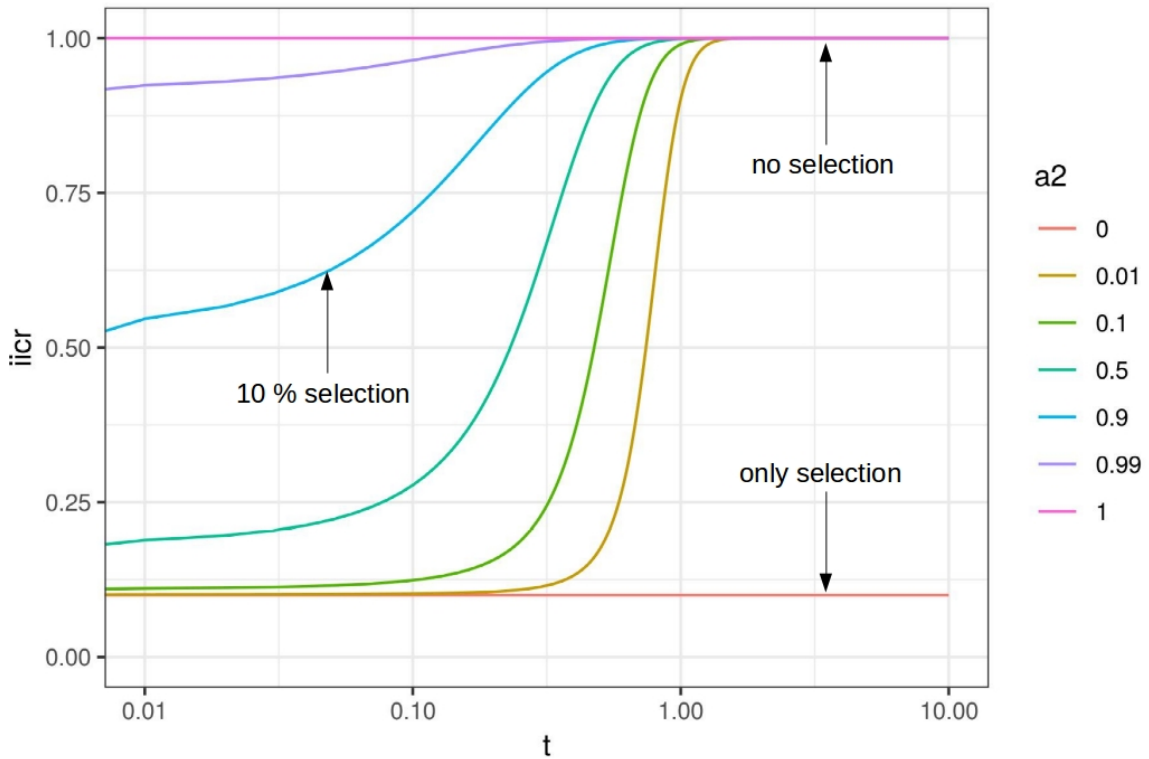
$$IICR(t) = -\frac{1 - F(t)}{f(t)} = \frac{\sum_{i=1}^{K} a_i e^{-\mu_i t}}{\sum_{i=1}^{K} a_i \mu_i e^{-\mu_i t}}.$$

Mathematical properties

- $IICR(t)$ increasing (backward in time)
  $\rightarrow$ spurious signal of **population size decline**.
- Starting value:
$$IICR(0) = \frac{1}{\sum_{i=1}^{K} \frac{a_i}{\lambda_i}}$$
- $IICR(t) \rightarrow \lambda_{max}$ as $t \rightarrow +\infty$

Two classes of N_e

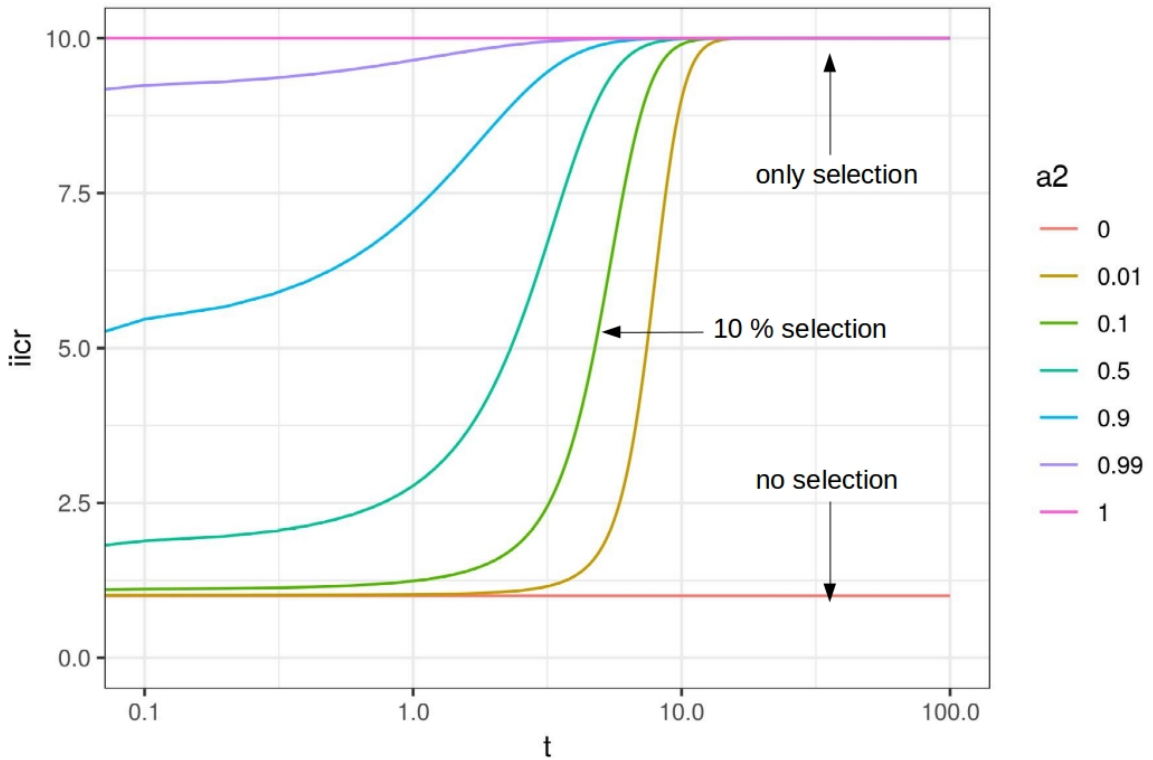
- $\lambda_1 = 0.1$ (BGS/SW), $\lambda_2 = 1$ (neutral).



$\rightarrow$ **BGS/SW** impact **recent past** IICR.

Two classes of N_e

- $\lambda_1 = 1$ (neutral), $\lambda_2 = 10$ (BalS).



$\rightarrow$ **BalS** impacts intermediate to **ancient past** IICR, even if in **small proportion**.

More complex models

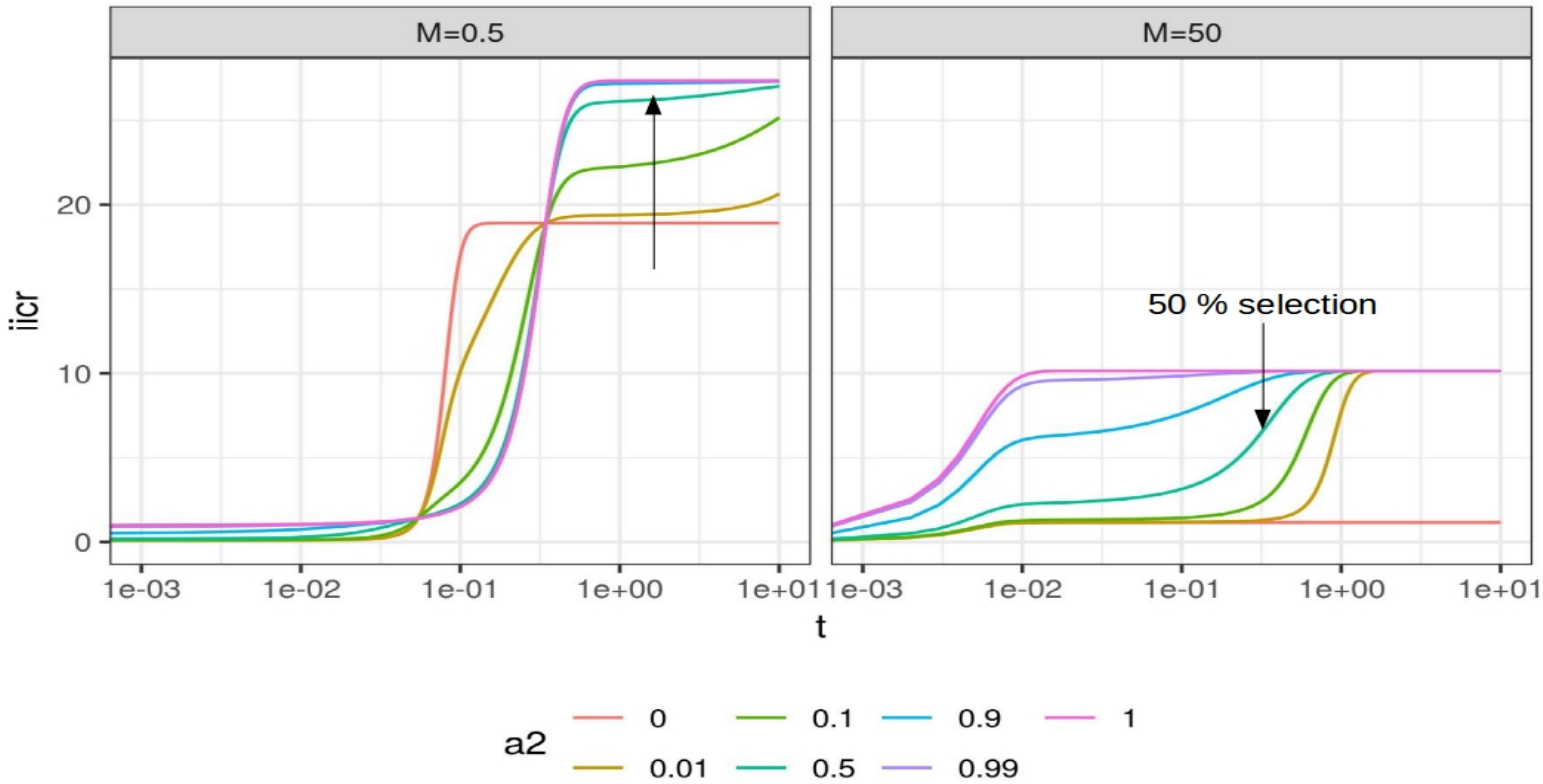
$$IICR(t) = \frac{\sum_{i=1}^{K} a_i(1 - F_i(t))}{\sum_{i=1}^{K} a_i f_i(t)}$$

$f_i()$ pdf of T_2.

- Genuine population size changes.
- Population structure (stationary or non stationary).
- Transient (rather than reccurent) selection.

Example 1: stationary n island model

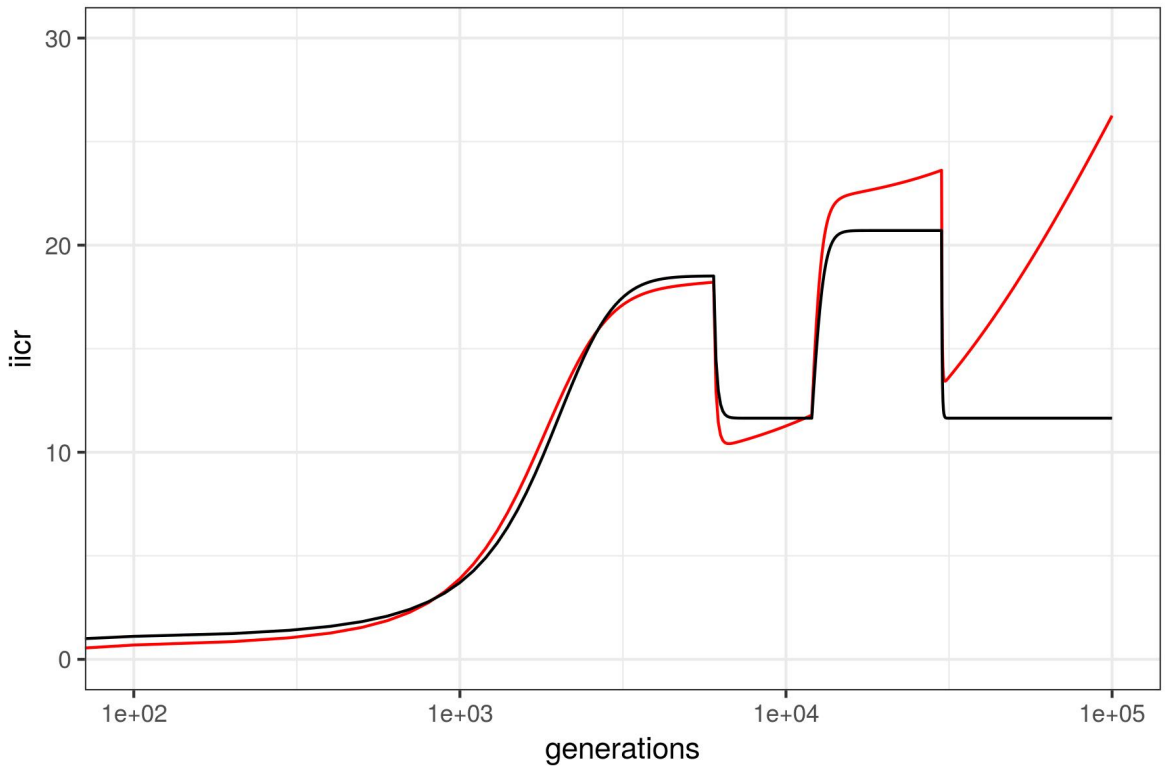
- $n = 10$ islands, two classes: $\lambda_1 = 0.1$, $\lambda_2 = 1$.



→ **Selection effect weaker than under panmixia** for $M \leq 1$.

Example 2: human evolution model (Mazet *et al*, 2016)

- $n = 10$, **non stationary** M (4 epochs).
- $+ \lambda$ distribution estimated by Gossmann *et al* (2011) (red).



$\rightarrow$ **Main effect** of selection in the **ancient past**.

Main conclusions

- Under panmixia, **linked selection** leaves a **signature of population decline** from the largest genomic N_e (with PSMC or MSMC).
- **BGS/SW impact recent past IICR, BalS ancient past**.
- **Larger effect of BalS**, even if in much lower proportions.
- Linked selection partly **masked by population strucure**.
- Ref: Genetics, 220(3), iyac008.

Perspectives

- **Joint inference of demography and linked selection** parameters?
- Approximate model of linked selection, but :
  - **fast** IICR evaluation.
  - **flexible** modelling: several forms of selection, population structure, demographic changes, transient or fluctuating selection ...
- Model genomic variations of mutation rate, gene flow ...

Advertizing

PhD positions in population genomics available in september 2022.

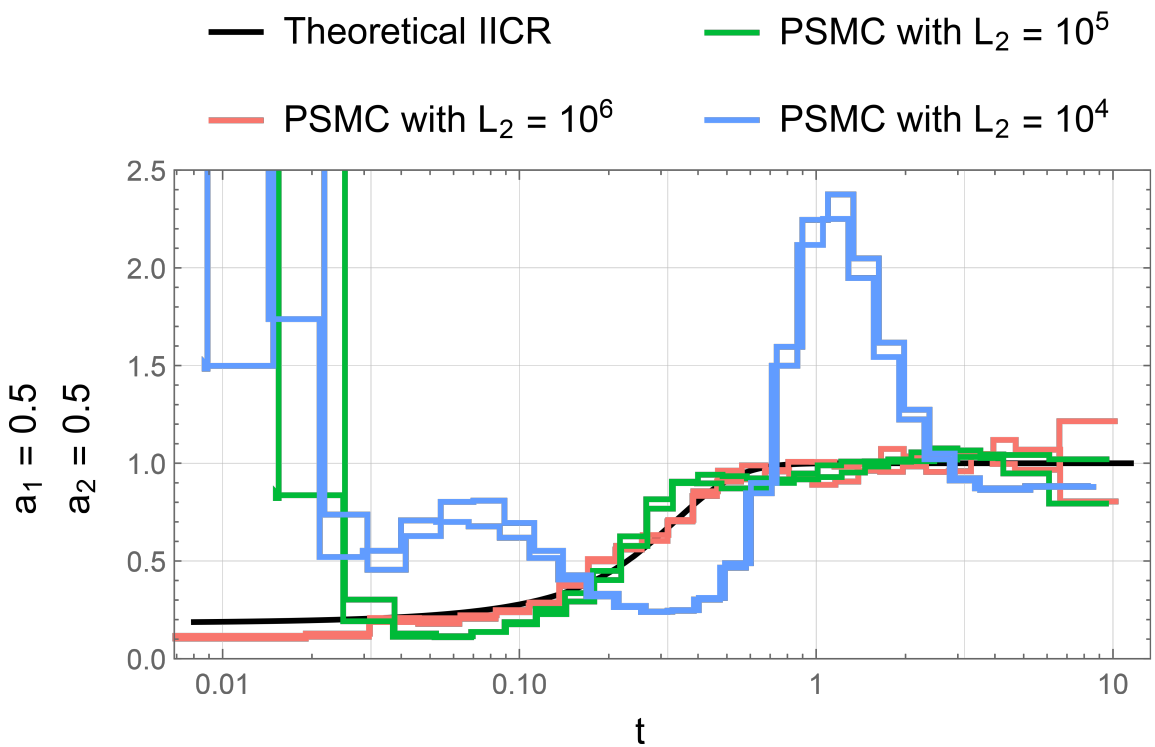
- Inference in spatial models.
- Inference from genomic time series.
- Genomic offset.

Why BGS does not leave a signal of expansion?

- The **decline signature** comes for the **variability of N_e along the genome**, not considered by
  - single locus models (Zeng and Charlesworth, 2011; Walczak *et al*, 2012).
  - models with **regular and short scale** alternance of neutral and selected loci (Johri *et al*, 2021).
- **Variable N_e hypothesis** less appropriate for samples properties like the **Site Frequency Spectrum** (Ewing and Jensen, 2016).
  - excess of singletons in BGS vs neutral models (Charlesworth *et al*, 1993).

IICR prediction and PSMC estimation

- Panmictic model with $\lambda_1 = 0.1$, $\lambda_2 = 1$.
- PSMC applied to genomes, simulated with variable N_e (class 1, class 2, class 1 . . .).

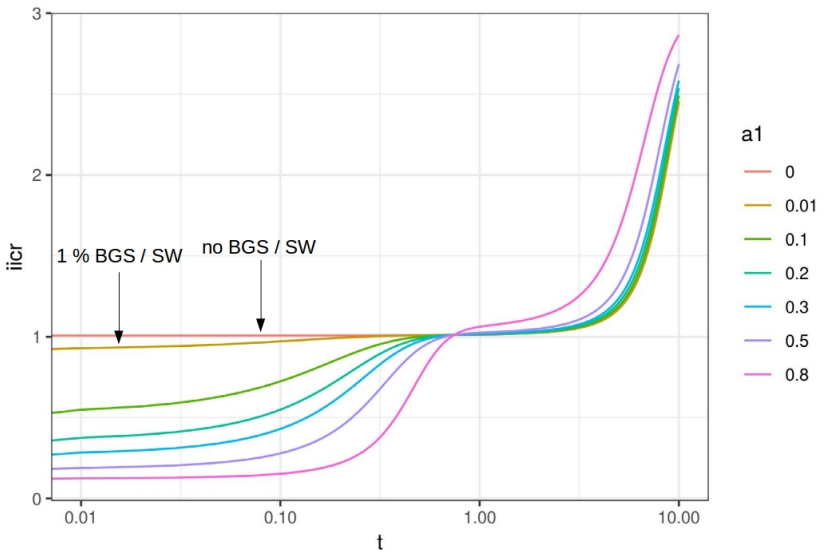


$\rightarrow$ consistent for **large scale N_e variartions**.

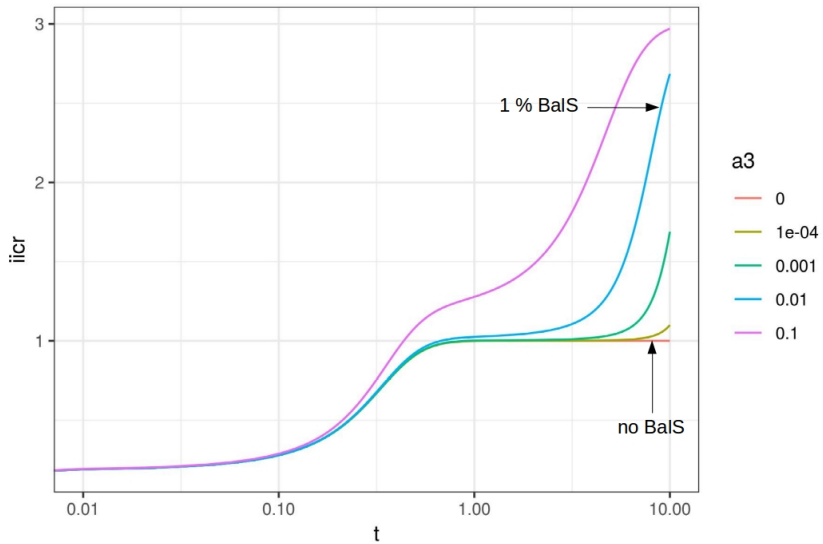
Three classes of N_e

$\lambda_1 = 0.1$ (BGS/SW), $\lambda_2 = 1$ (neutral), $\lambda_3 = 3$ (BalS).

$a_3 = 0.01$



$a_1 = 0.5$

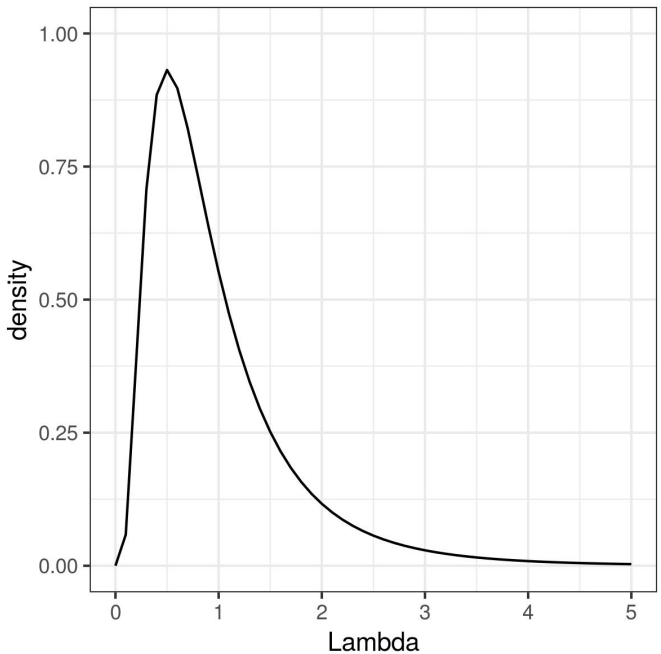


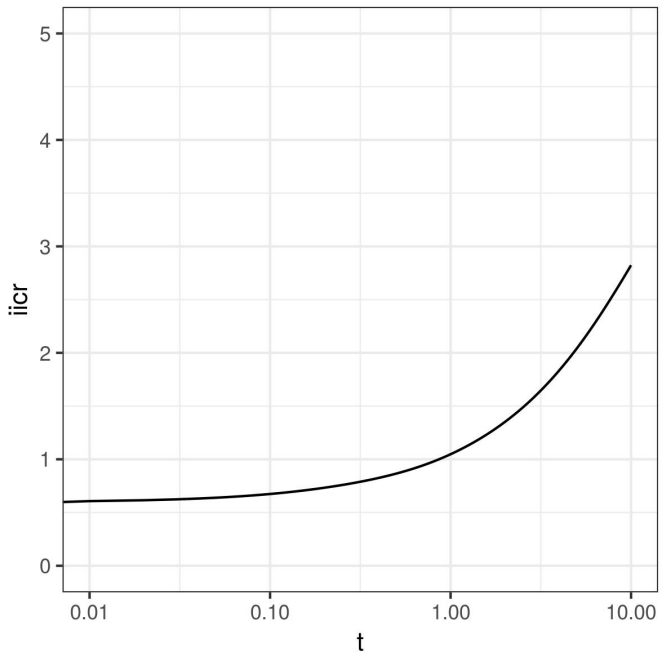
$\rightarrow$ stronger effect of BalS vs BGS/sweeps for the same proportion.

$\rightarrow$ intermediate plateau depending on a_1 and a_3

Realistic N_e distributions: *humans*

Gossmann *et al* (2011): N_e distribution assumed log-normal and estimated from polymorphism and divergence data.





$\rightarrow$ long term 6-fold decline from $t = 10$.