Challenges when applying stochastic models to reconstruct the demographic history of populations.

Willy Rodríguez

Institut de Mathématiques de Toulouse

October 11, 2017

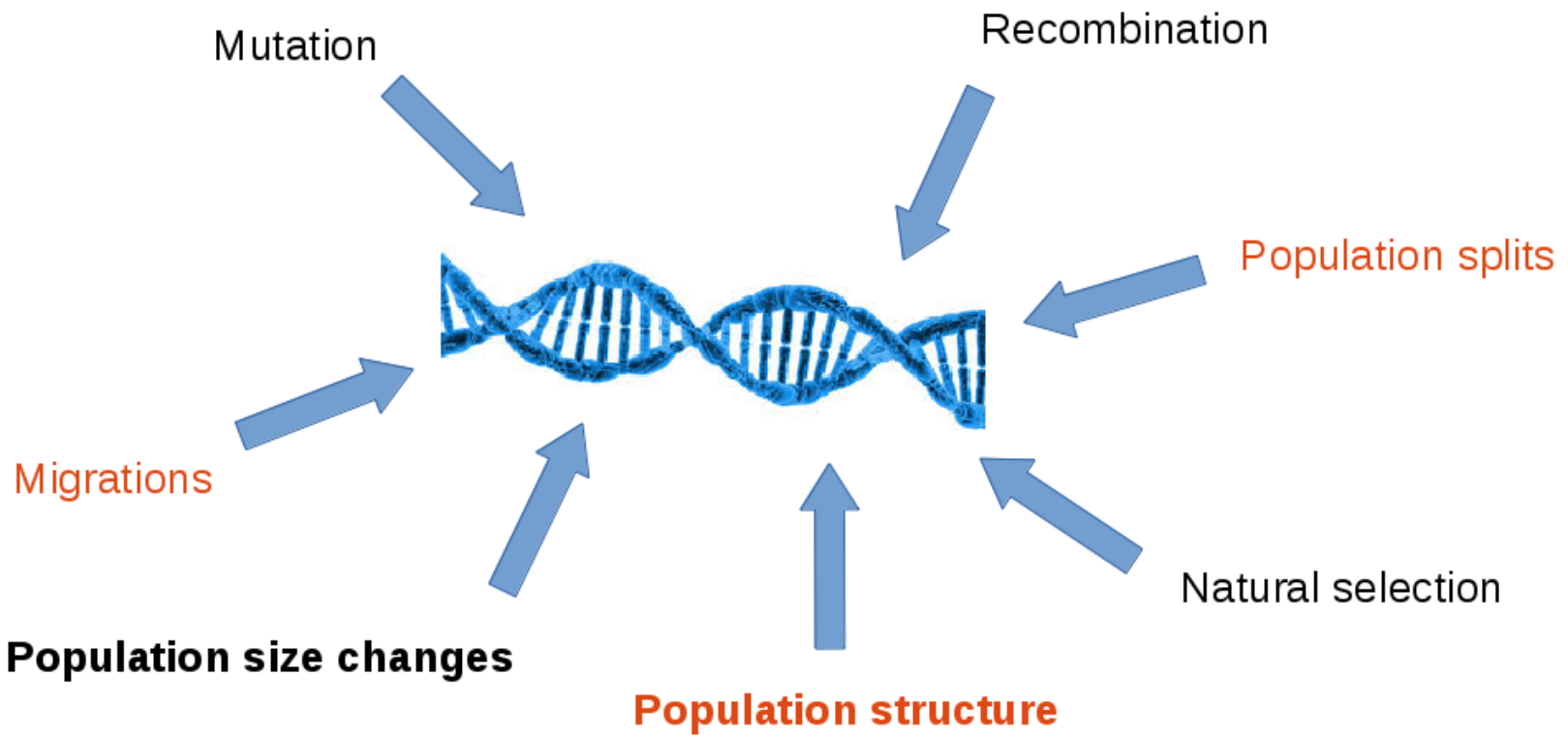
Outline

Presentation Outline

Goals of Population Genetics

Build models that help us understand the main evolutionary events that gave rise to the observed patterns of genetic variation.

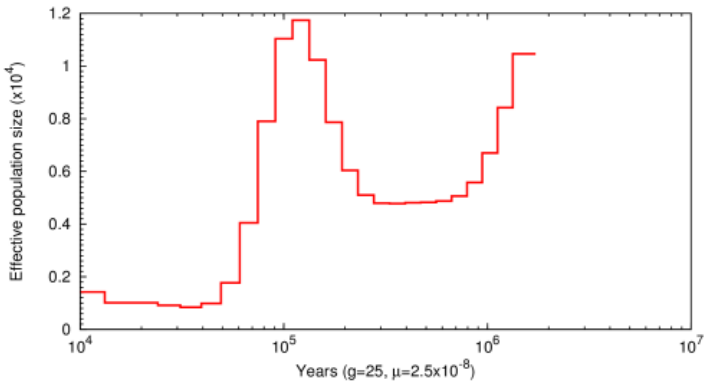
Evolutionary forces



Reconstructing demographic history from DNA

DNA sequences

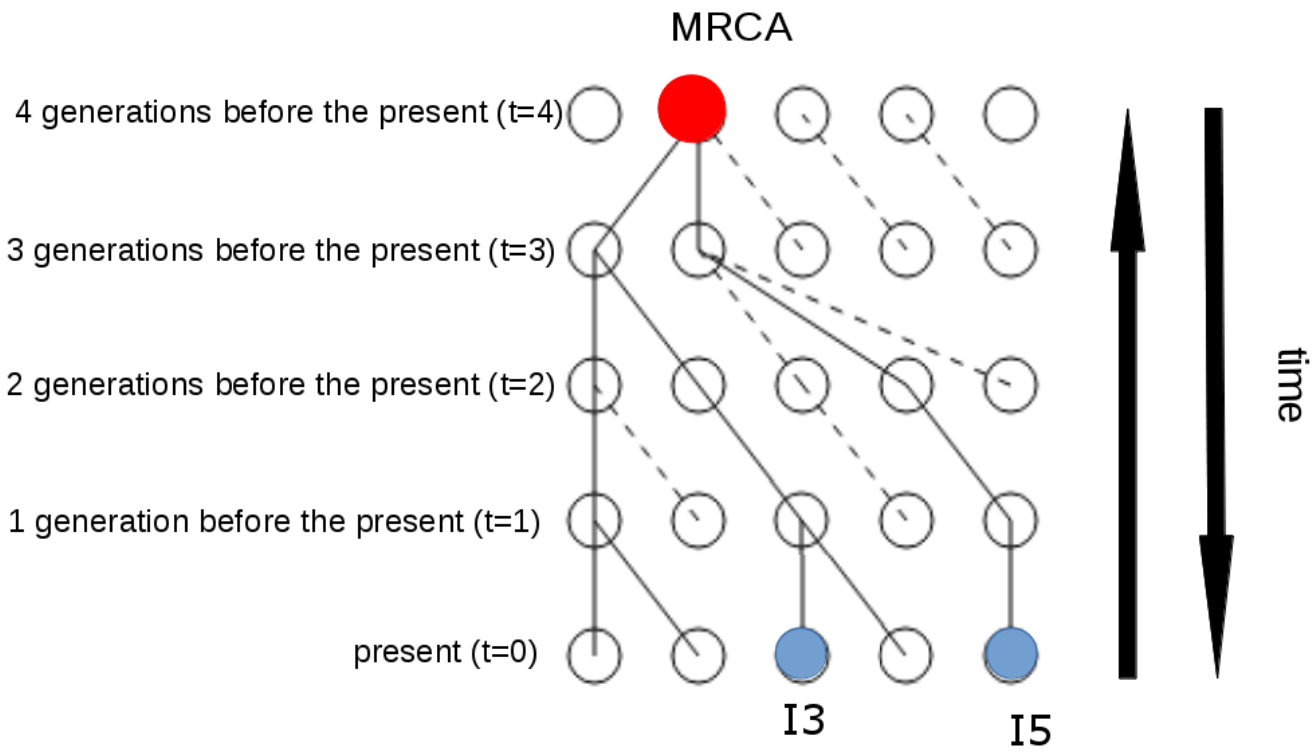
Inferred demographic history



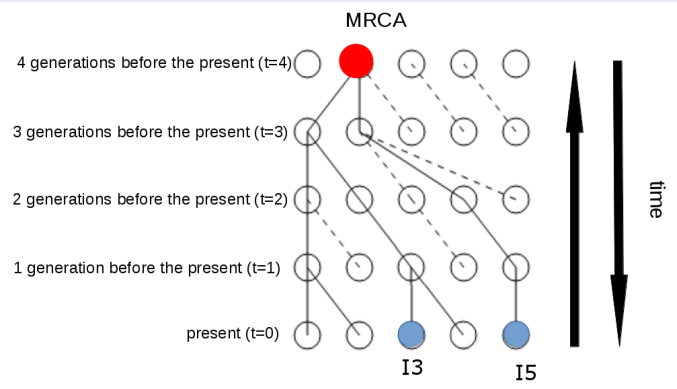
Wright-Fisher model

Wright-Fisher model (1930-1931)

- Non-overlapping generations.
- Constant population size ($2N$ genes, 1 indiv. = one gene).
- Random mating (panmixia).



Wright-Fisher model and key concepts



Key concepts

- Coalescence events: Individuals I_3 and I_5 **coalesce** (have a common ancestor) 4 generations before the present.
- Probability that two individuals coalesce in the previous generation: $\frac{1}{2N}$. Probability that they do not coalesce: $(1 - \frac{1}{2N})$.
- G_2: Number of generations to reach the ancestor of 2 genes. $\mathbb{P}(G_2 > \ell) = (1 - \frac{1}{2N})^{\ell}$.
- Let $2N$ be large, $\frac{G_2}{2N} \xrightarrow{\mathcal{D}} T_2$,with $T_2 \sim Exp(1)$.

The Coalescent and some approximations

The Coalescent and some approximations

Kingman, J., 1982. *The coalescent*.
A continuous-time Markov chain allowing to describe family relationships between individuals *backward in time*.

The Coalescent and some approximations

Kingman, J., 1982. *The coalescent*.
A continuous-time Markov chain allowing to describe family relationships between individuals *backward in time*.

Hudson, R. R., 1983. *The coalescent with recombination*.
A process allowing to trace back lineages, incorporating recombination events.

The Coalescent and some approximations

Kingman, J., 1982. *The coalescent*.
A continuous-time Markov chain allowing to describe family relationships between individuals *backward in time*.

Hudson, R. R., 1983. *The coalescent with recombination*.
A process allowing to trace back lineages, incorporating recombination events.

Herbots, H. M. J. D. 1994. *The structured coalescent*.
An extension of Kingman's coalescent to subdivided populations.

The Coalescent and some approximations

Kingman, J., 1982. *The coalescent*.
A continuous-time Markov chain allowing to describe family relationships between individuals *backward in time*.

Hudson, R. R., 1983. *The coalescent with recombination*.
A process allowing to trace back lineages, incorporating recombination events.

Herbots, H. M. J. D. 1994. *The structured coalescent*.
An extension of Kingman's coalescent to subdivided populations.

McVean & Cardin, 2005. *The Sequentially Markovian Coalescent*.
An approximation to the coalescent with recombination using a Markov chain along the genome.

Variable population size. Relation with T_2

Griffiths & Tavaré. 1994.

Distribution of T_2, function of population size change.

- Population evolving with deterministically varying size.
- $\lambda(t) = \frac{N(t)}{2N}$, with $2N$: present population size (genes).
- $\Lambda(t) = \int_0^t \frac{1}{\lambda(u)} du$.
- Distribution of the coalescence times of two genes (T_2).
- $\mathbb{P}(T_2 > t) = 1 - F_{T_2}(t) = e^{-\Lambda(t)}$
- $f_{T_2}(t) = \left(F_{T_2}(t)\right)' = \frac{1}{\lambda(t)} e^{-\Lambda(t)}$.

Objective: Reconstruct the function λ.

Methods for estimating past population size changes

- *MSVAR* (1999).
- *Skyline plot* (2000).
- *Bayesian skyline plot* (2005).
- dadi (2009)
- *PSMC* (2011).
- *DiCal* (2013).
- *VarEff* (2014).
- *MSMC* (2014).
- stairway plot (2015)
- *PopSizeABC* (2016).
- SMC++ (2017)

Methods for estimating past population size changes

- *MSVAR* (1999).
- *Skyline plot* (2000).
- *Bayesian skyline plot* (2005).
- dadi (2009)
- *PSMC* (2011).
- *DiCal* (2013).
- *VarEff* (2014).
- *MSMC* (2014).
- stairway plot (2015)
- *PopSizeABC* (2016).
- SMC++ (2017)

Methods for estimating past population size changes

- *PSMC* (2011).

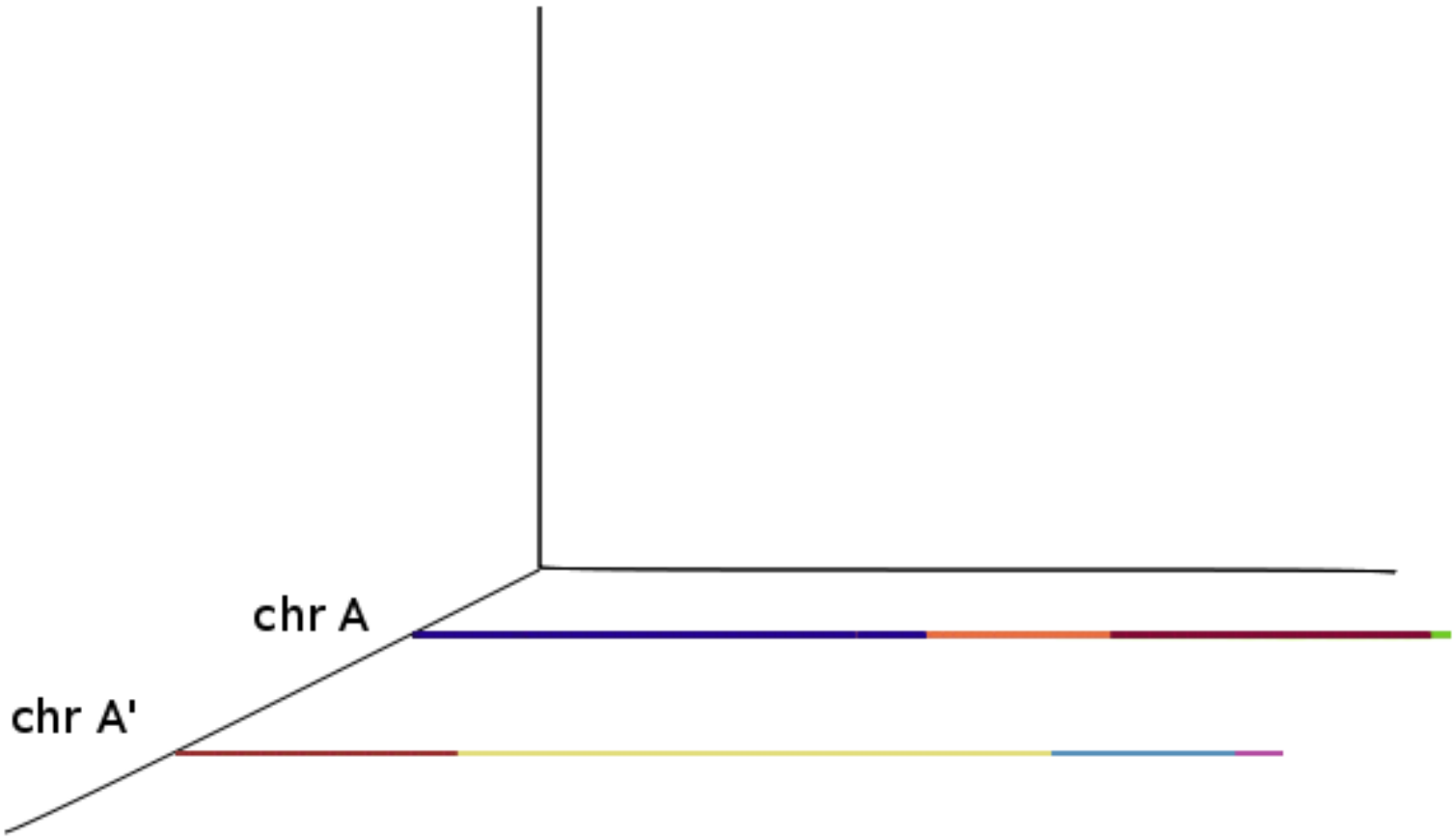
The PSMC. Markov Chain along the genome

Li and Durbin, 2011. *Pairwise Sequentially Markovian Coalescent.*

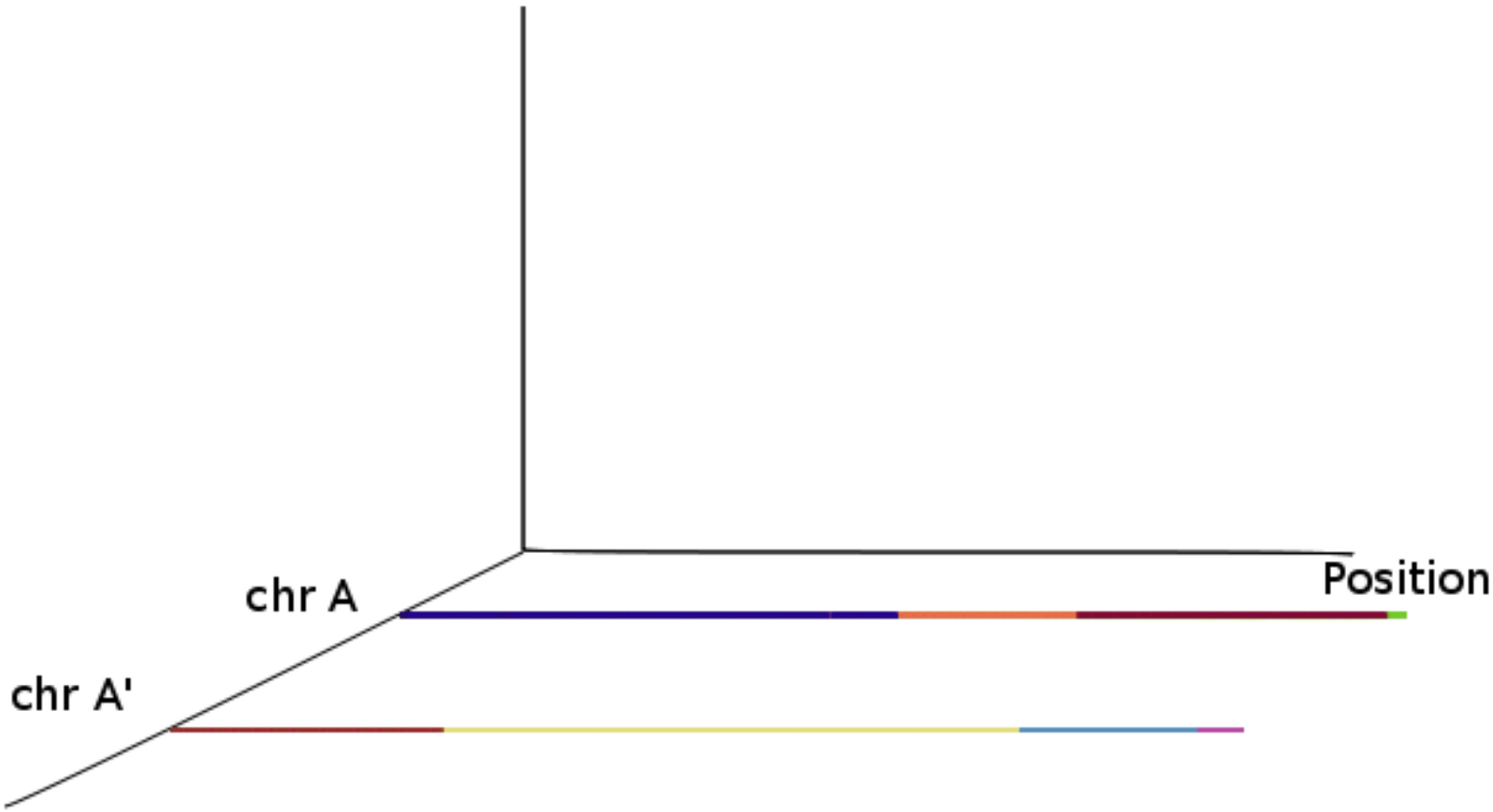
The PSMC method

- Input data: one diploid genome (one diploid individual).
- Based on the Sequentially Markovian Coalescent (SMC. McVean & Cardin, 2005).
- Uses a relation between recombinations, mutations and λ_k in a model of variable population size.
- Describes a Hidden Markov Model along the genome, allowing to estimate values of population size in the past.
- Developed to be applied on long DNA sequences (*e.g.*: one chromosome.)

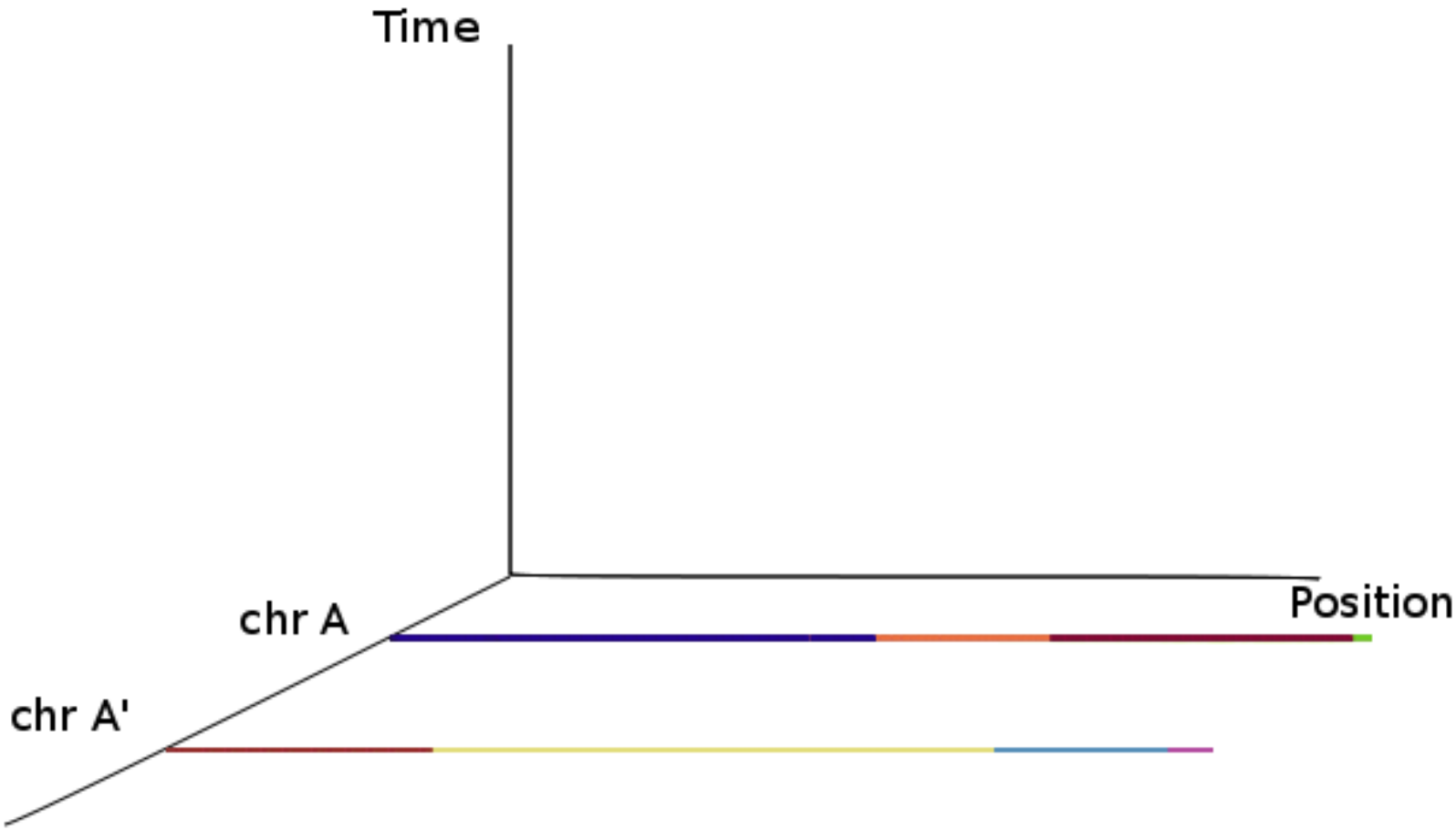
PSMC. Markov Chain along the genome



PSMC. Markov Chain along the genome

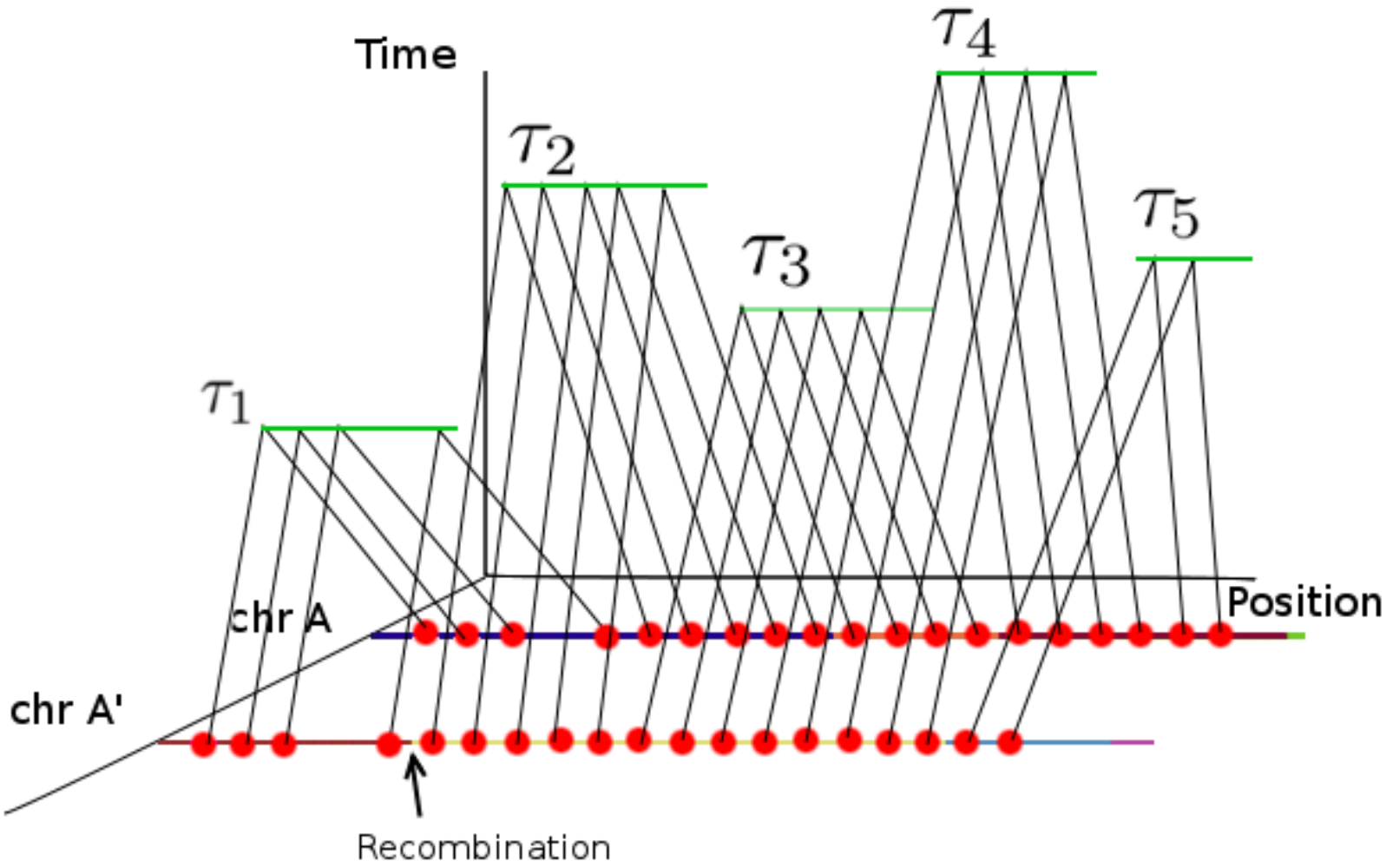


PSMC. Markov Chain along the genome

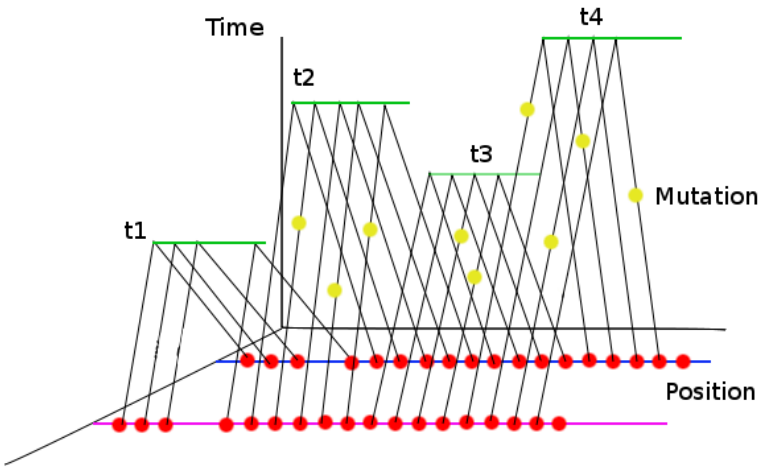


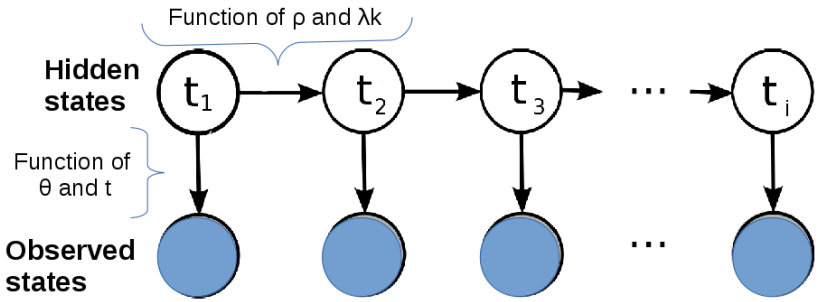
PSMC. Markov Chain along the genome

PSMC. Markov Chain along the genome



PSMC. Hidden Markov Chain along the genome





Discrete state space, discrete-time HMM.

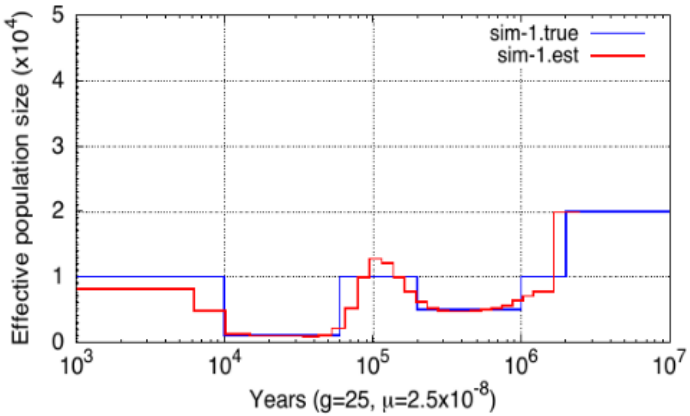
Hidden states: coalescence times (T_2) at position a ($T_2 \in [t_k, t_{k+1}]$).

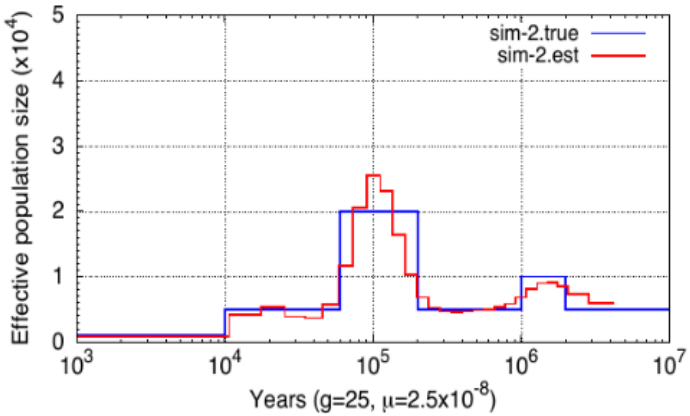
Observed states: homozygous or heterozygous at position a.

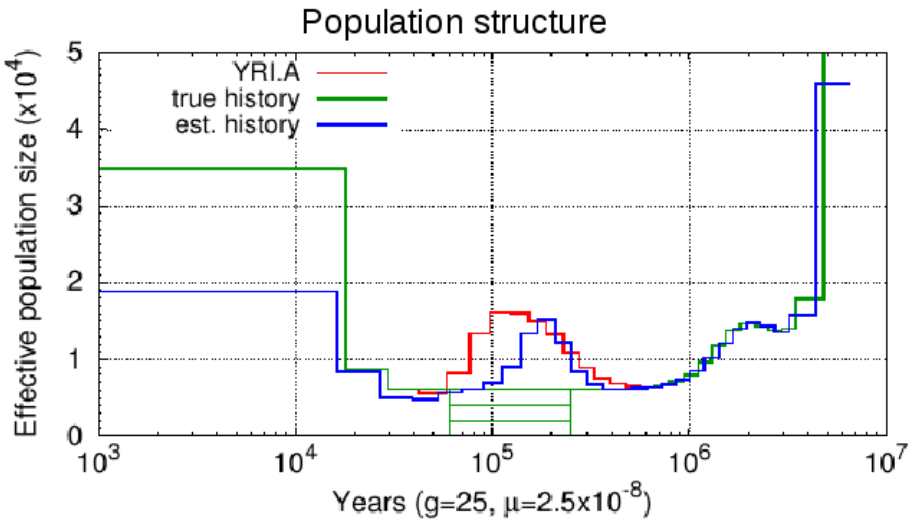
Parameters: scaled mutation rate (θ), scaled recombination rate (ρ), demographic history (λ_k).

One example (the psmc)

The psmc method (Heng Li & Richard Durbin, 2011) can be affected by population structure.

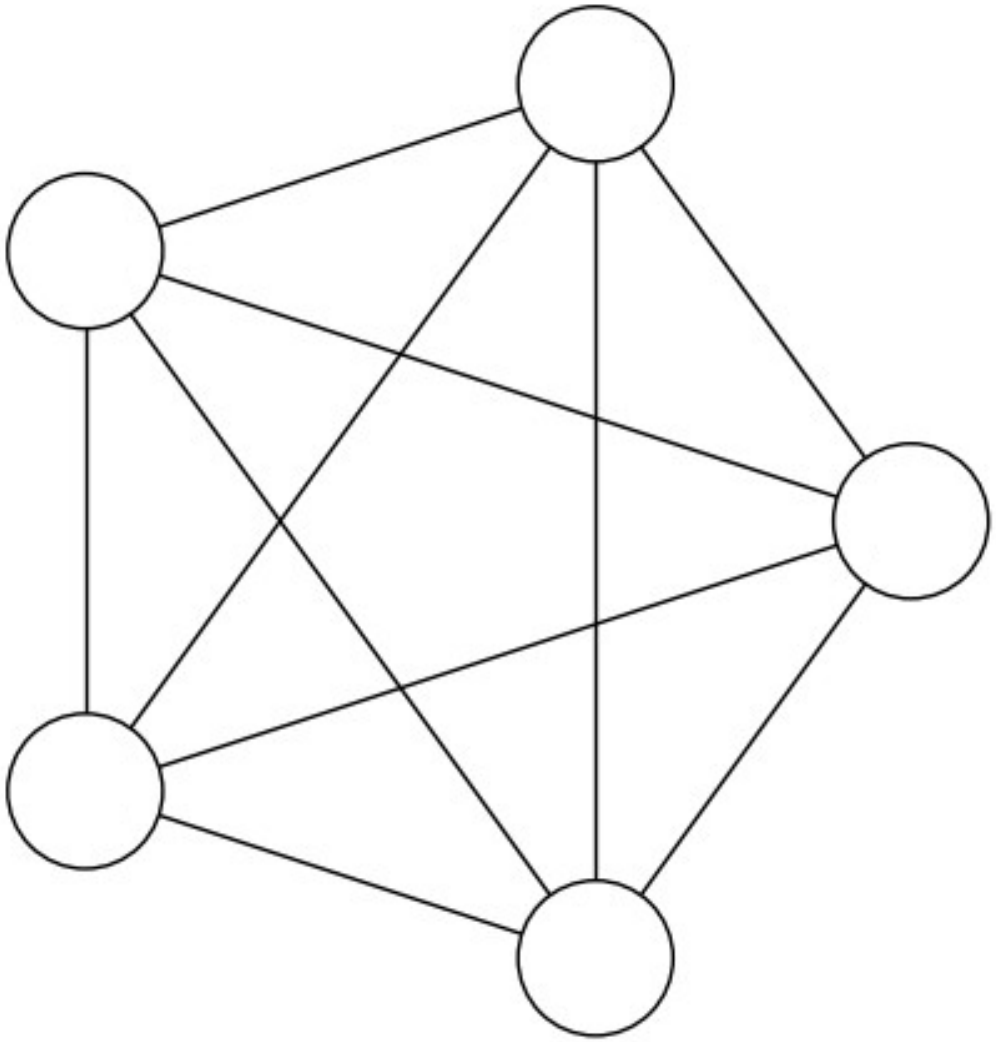




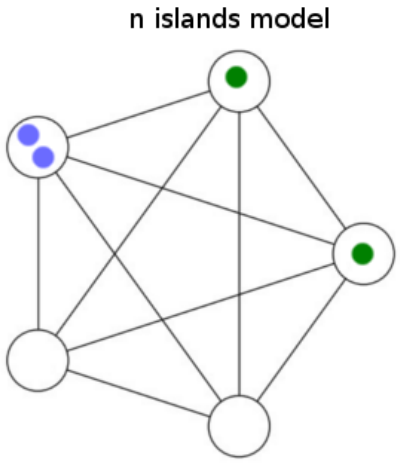


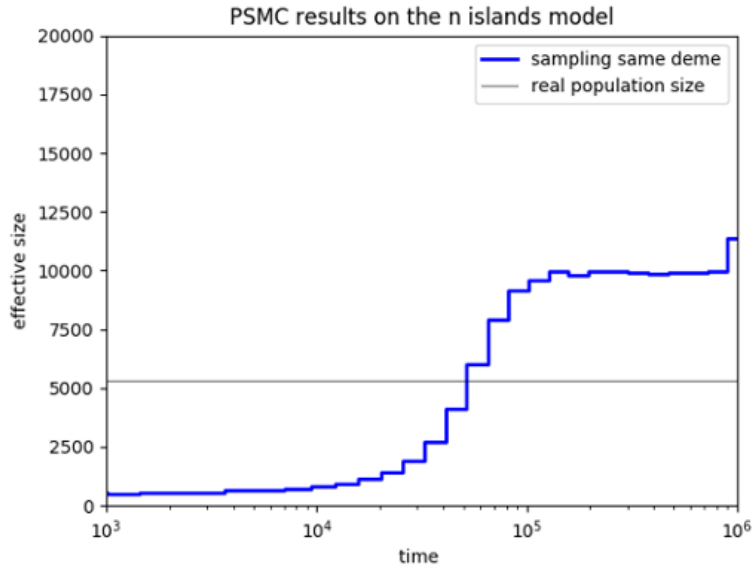
One example (the psmc). A simple structure model

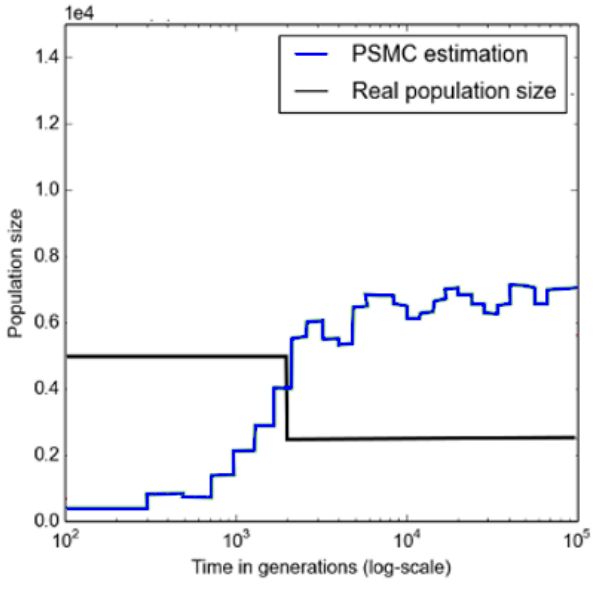
n islands model

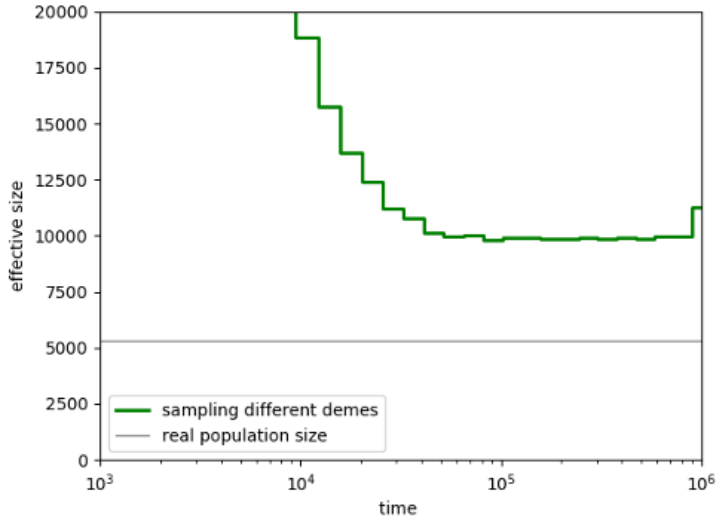


One example (the psmc). A simple structure model





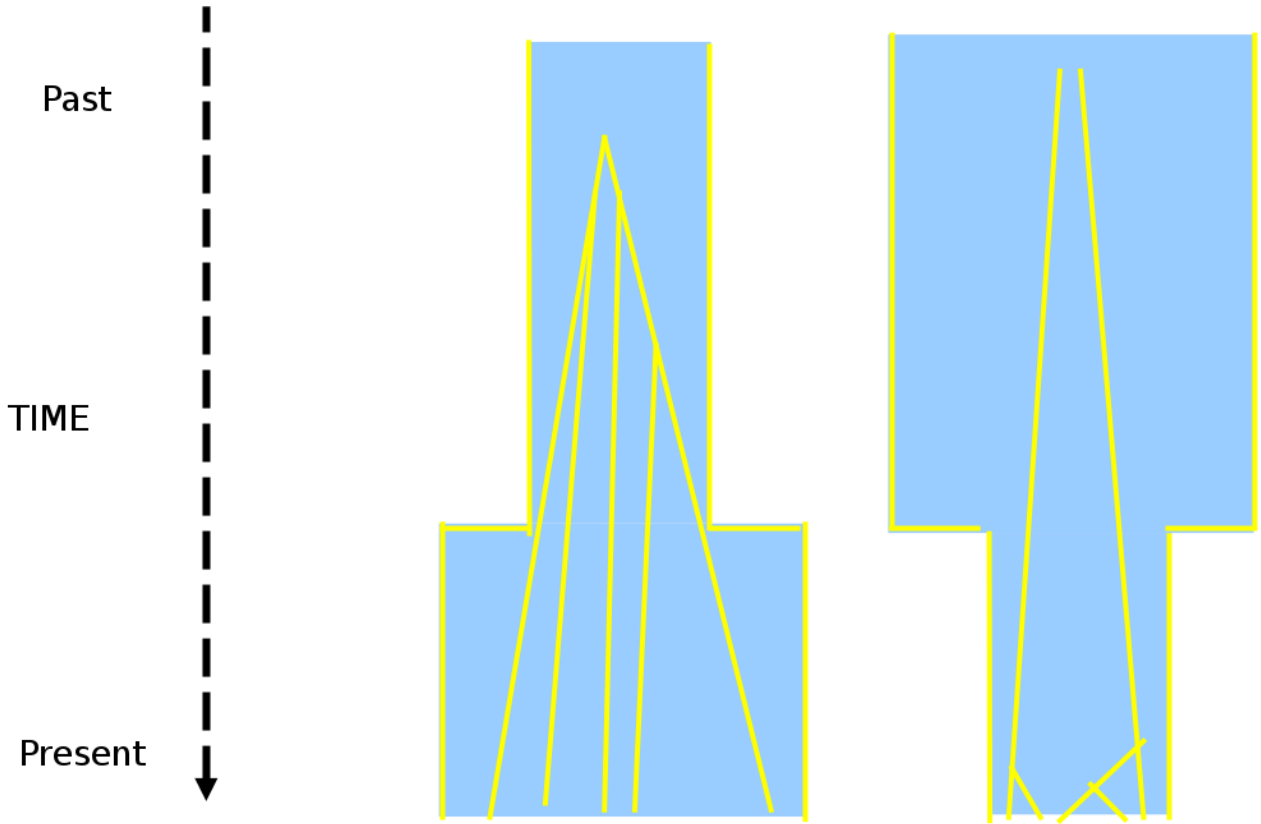




What is going on here ?

Presentation Outline

Coalescence times and population size changes. The **panmixia** hypothesis (random mating).



What if we do not have random mating ?

What if we do not have random mating ?

Instantaneous Coalescence Rate

Griffiths & Tavaré (1994). Variable population size (N_e)

$$\begin{aligned} N_e(t) &= N_0\lambda(t) \\ \mathbb{P}(T_2 > t) &= e^{-\int_0^t \frac{1}{\lambda(u)}du} \end{aligned}$$

Inverse Instantaneous Coalescence Rate (IICR). *Mazet et al. 2015*

- $\log(\mathbb{P}(T_2 > t))' = -\frac{1}{\lambda(t)}$
- $\lambda(t) = \frac{\mathbb{P}(T_2>t)}{f_{T_2}(t)} = \frac{1-F_{T_2}(t)}{f_{T_2}(t)}$.
- λ can be evaluated at any t using only the T_2 distribution. Valid for **any model**.
- T_2 can be interpreted as a *lifetime*. $\frac{1}{\lambda(t)}$: instantaneous coalescence rate (hazard function of failure rate).
- $\lambda(t)$ may be disconnected with size changes.

Based on T_2, it is **not possible** to distinguish *structure* from a *panmictic* model with *population size change*.

Based on T_2, it is **not possible** to distinguish *structure* from a *panmictic* model with *population size change*.

We can estimate the IICR using any method based on the panmixia hypothesis (ex: psmc).

Based on T_2, it is **not possible** to distinguish *structure* from a *panmictic* model with *population size change*.

We can estimate the IICR using any method based on the panmixia hypothesis (ex: psmc).

We can predict the IICR for any model by doing simulations
Take a vector of independent values of T_2. Then:

$$\hat{\lambda}(t) = \frac{1 - \hat{F}_{T_2}(t)}{\hat{f}_{T_2}(t)}.$$

Based on T_2, it is **not possible** to distinguish *structure* from a *panmictic* model with *population size change*.

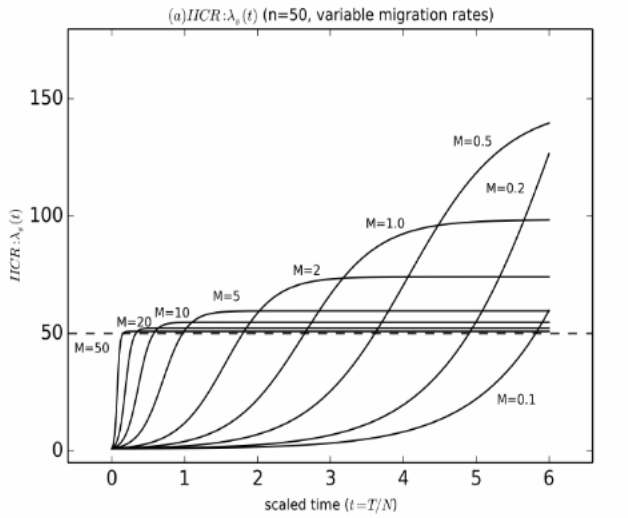
We can estimate the IICR using any method based on the panmixia hypothesis (ex: psmc).

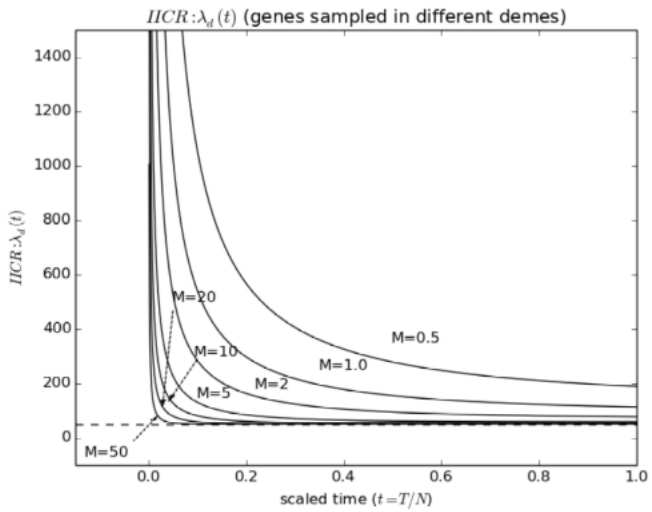
We can predict the IICR for any model by doing simulations
Take a vector of independent values of T_2. Then:

$$\hat{\lambda}(t) = \frac{1 - \hat{F}_{T_2}(t)}{\hat{f}_{T_2}(t)}.$$

https://github.com/willyrv/IICREstimator

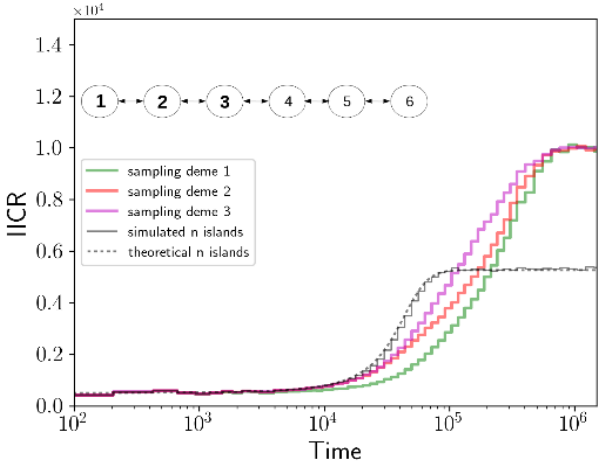
Explicit expression for the IICR under the n islands model.



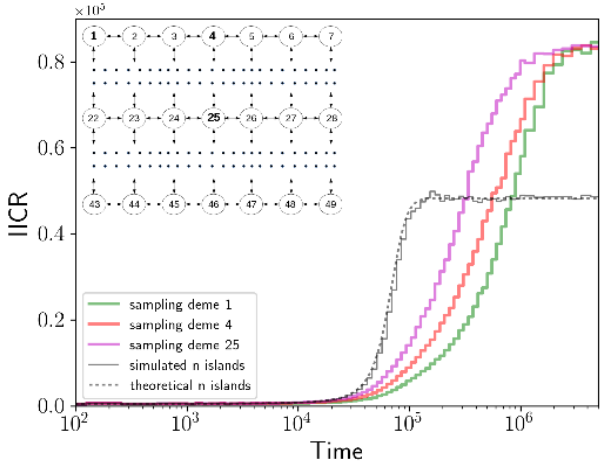


Presentation Outline

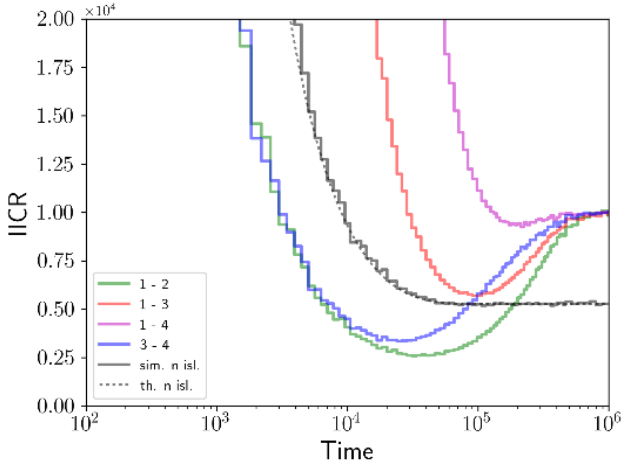
Trace the IICR for non panmictic models



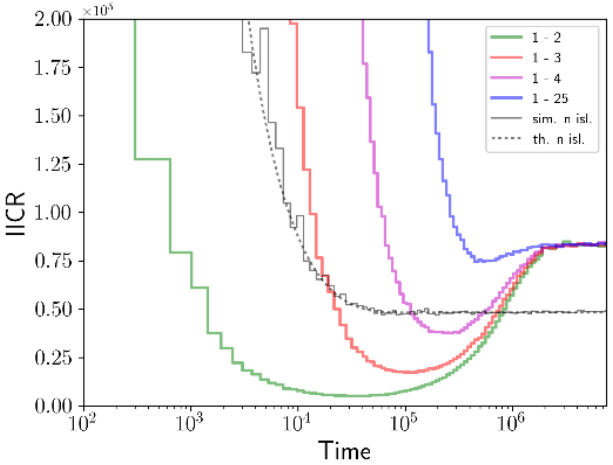
(a) $IICR_s$ 1D, $n = 6$, $M = 1$



(b) $IICR_s$ 2D, $n = 49$, $M = 1$



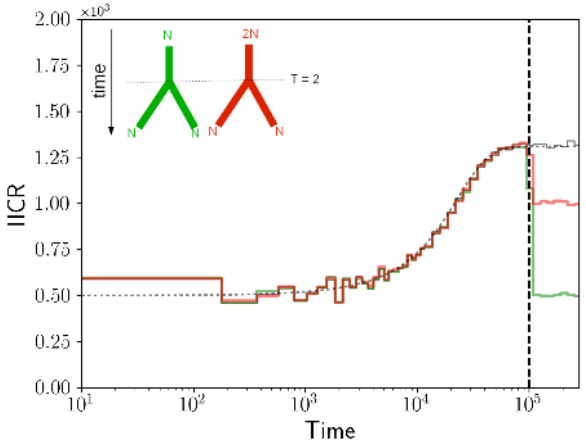
(c) $IICR_d$ 1D, $n = 6$, $M = 1$



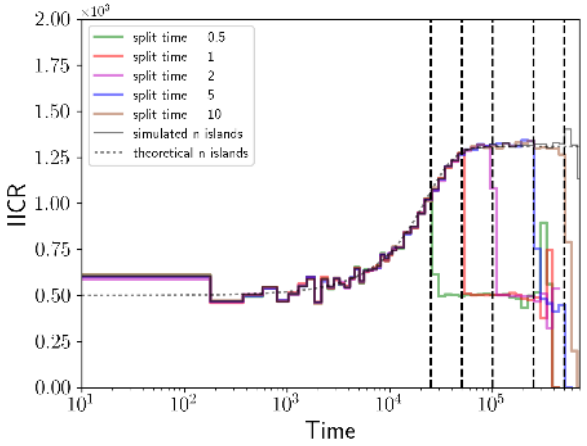
(d) $IICR_d$ 2D, $n = 49$, $M = 1$

L. Chikhi et al. *The IICR (inverse instantaneous coalescence rate) as a summary of genomic diversity: insights into demographic inference and model choice* (to appear HDY)

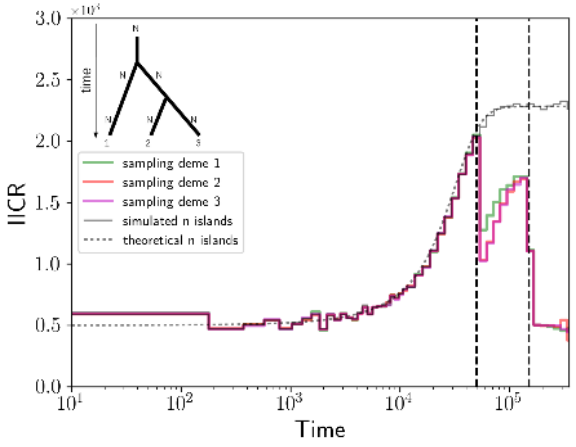
Trace the IICR for non panmictic models



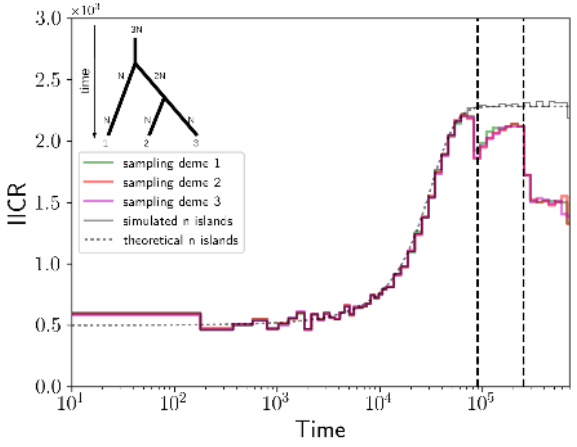
(a) Ancestral size N (green) or $2N$ (red)



(b) Ancestral size N and different splitting times



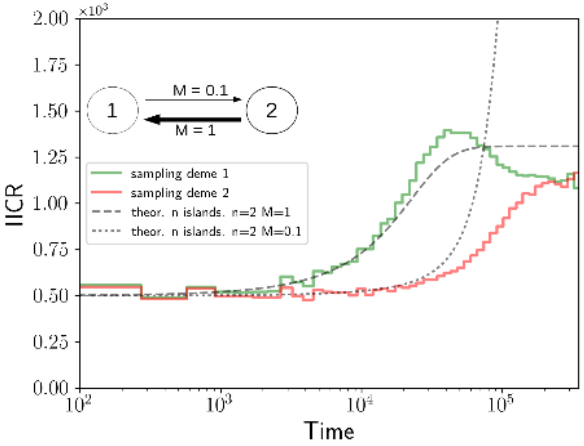
(c) Split with 3-island model and ancestral size N



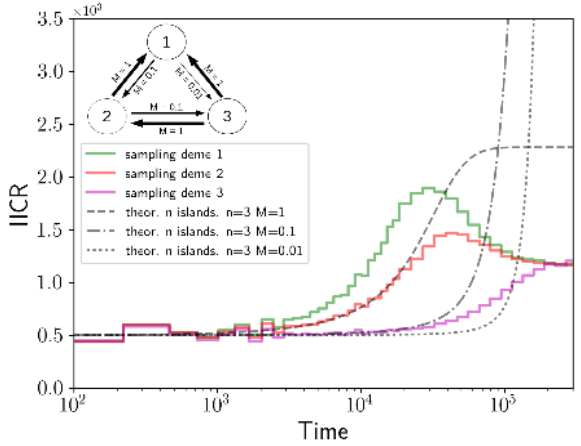
(d) Split with 3-island model and ancestral size $3N$

L. Chikhi et al. *The IICR (inverse instantaneous coalescence rate) as a summary of genomic diversity: insights into demographic inference and model choice* (to appear HDY)

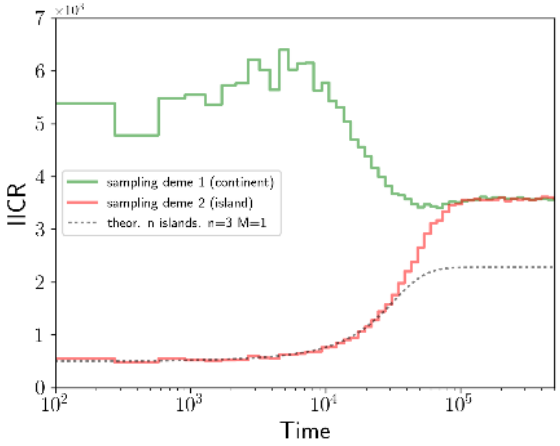
Trace the IICR for non panmictic models



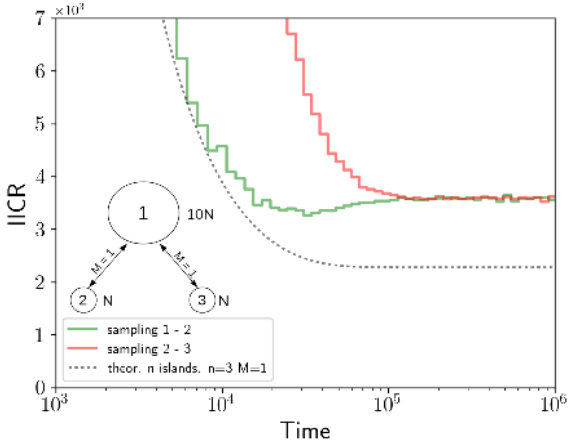
(a) Asymmetrical gene flow: ratio 1 : 10



(b) Asymmetrical gene flow: ratios 1 : 10 and 1 : 100



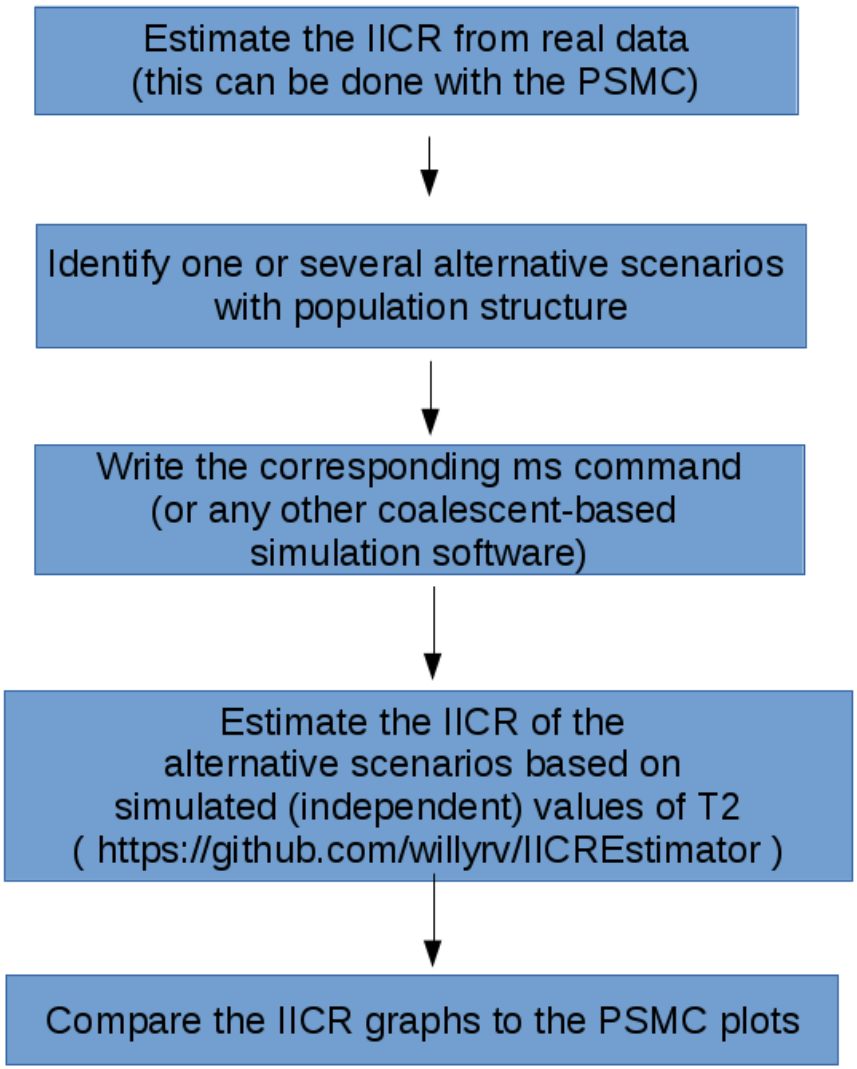
(c) Continent-island model, size ratio 1 : 10



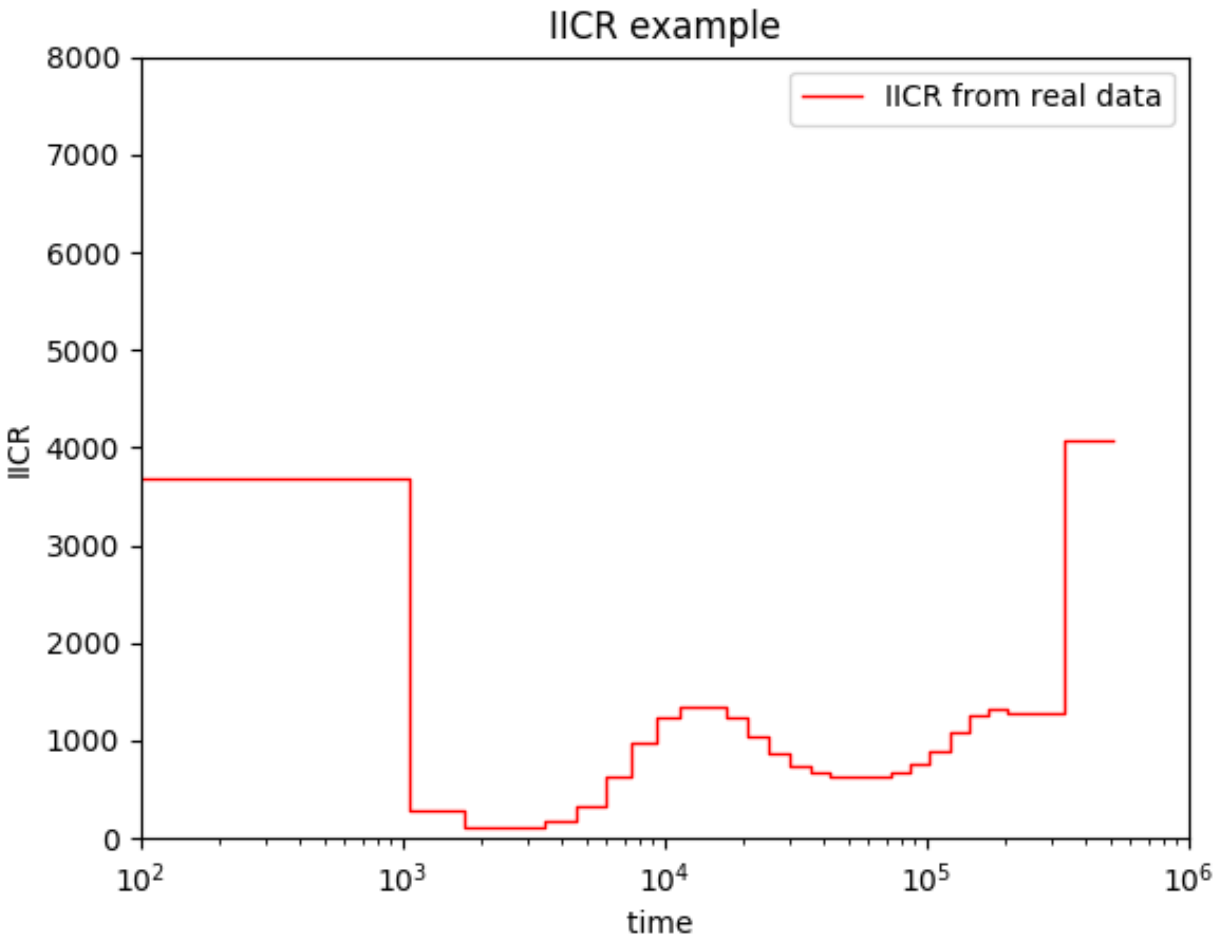
(d) Continent-island model, size ratio 1 : 10

L. Chikhi et al. *The IICR (inverse instantaneous coalescence rate) as a summary of genomic diversity: insights into demographic inference and model choice* (to appear HDY)

Considering alternative scenarios

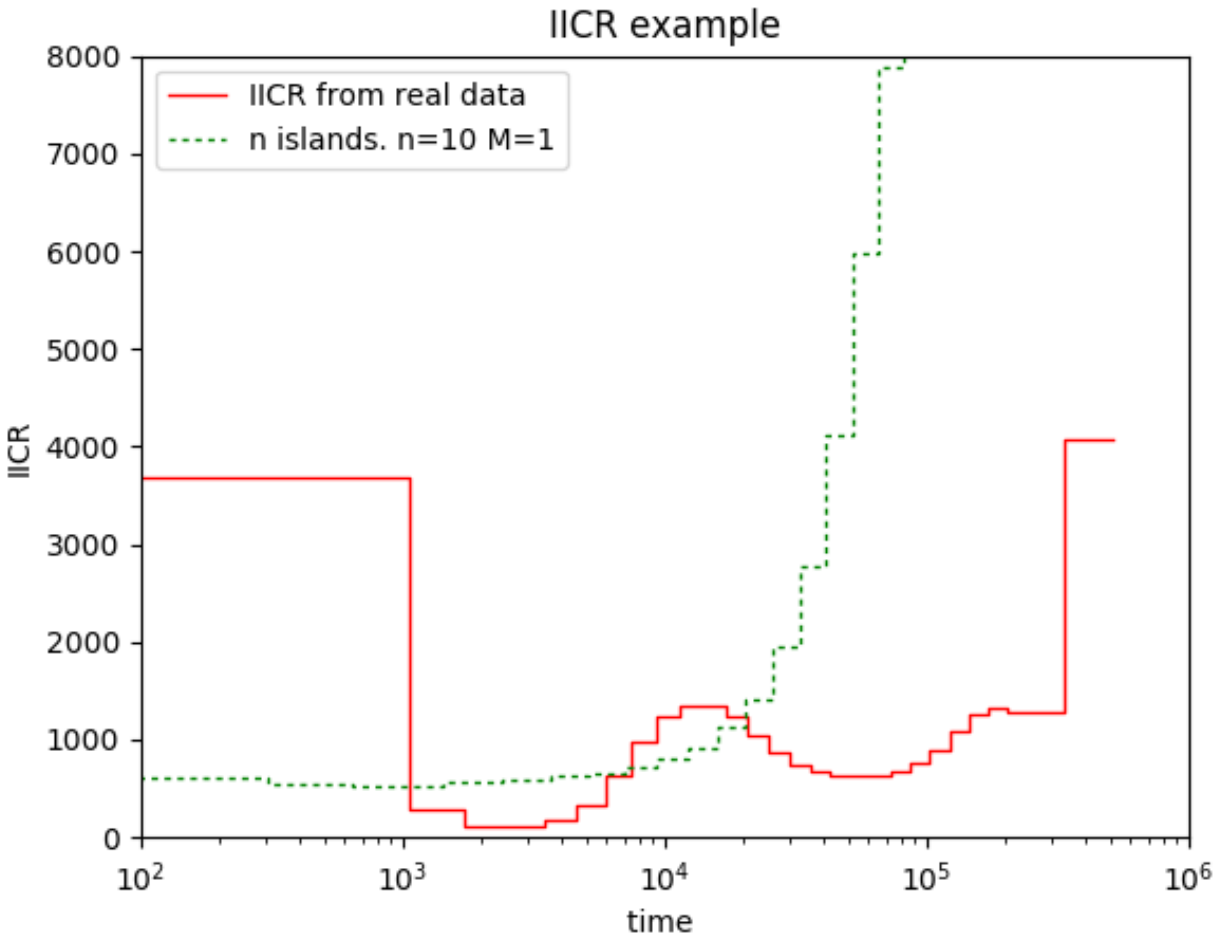


Considering alternative scenarios



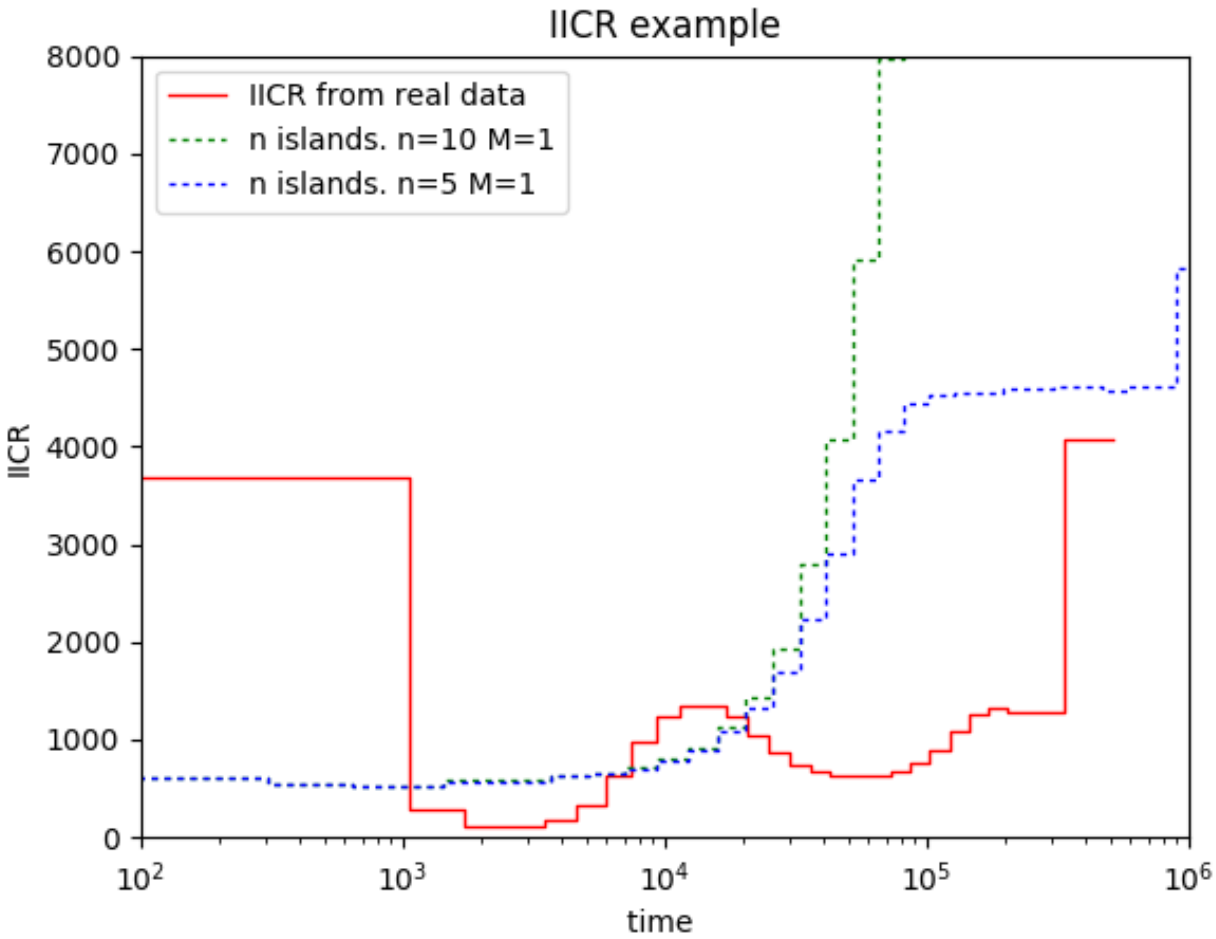
Courve fitting (https://github.com/MaxHalford/stsicmr-inference)

Considering alternative scenarios



Courve fitting (https://github.com/MaxHalford/stsicmr-inference)

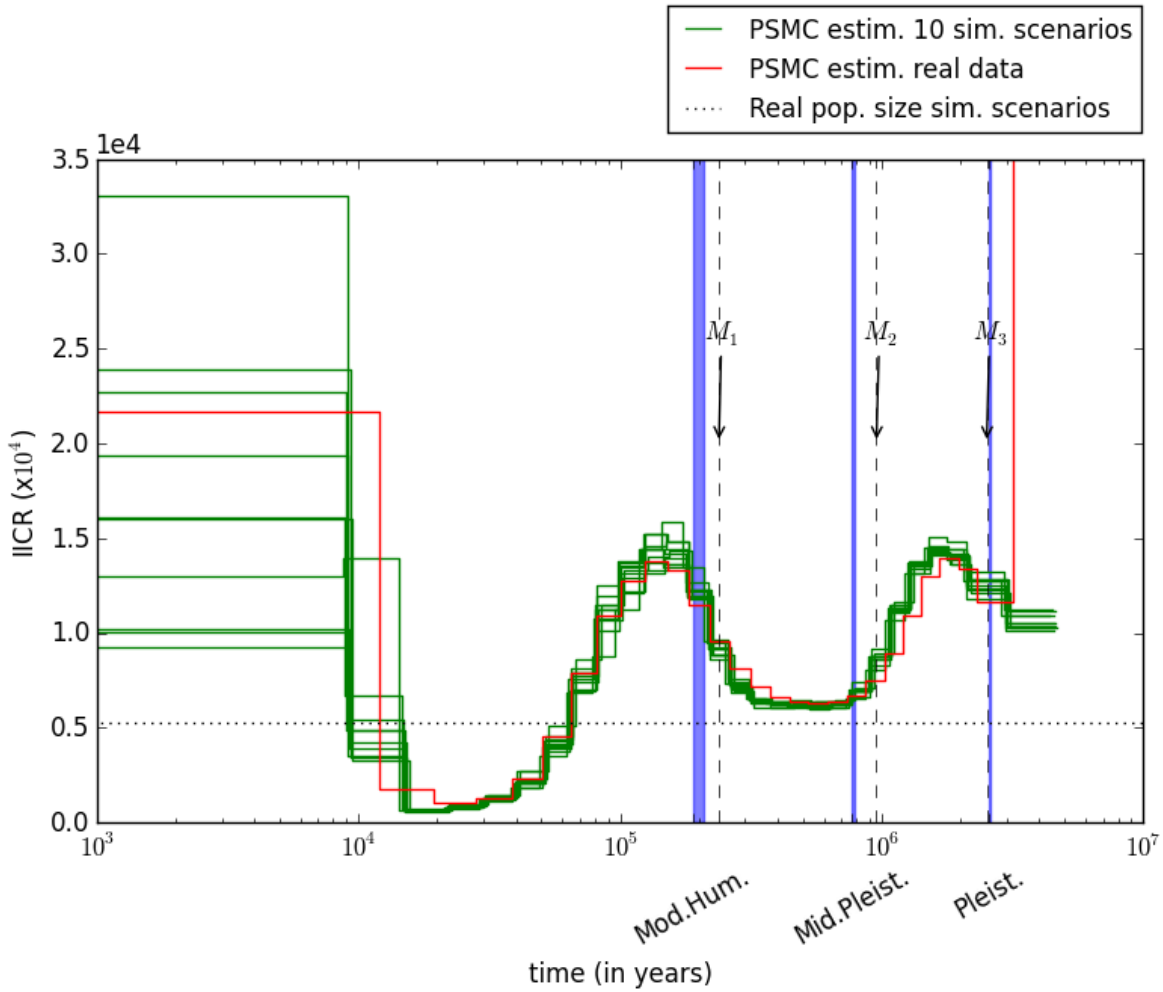
Considering alternative scenarios



Courve fitting (https://github.com/MaxHalford/stsicmr-inference)

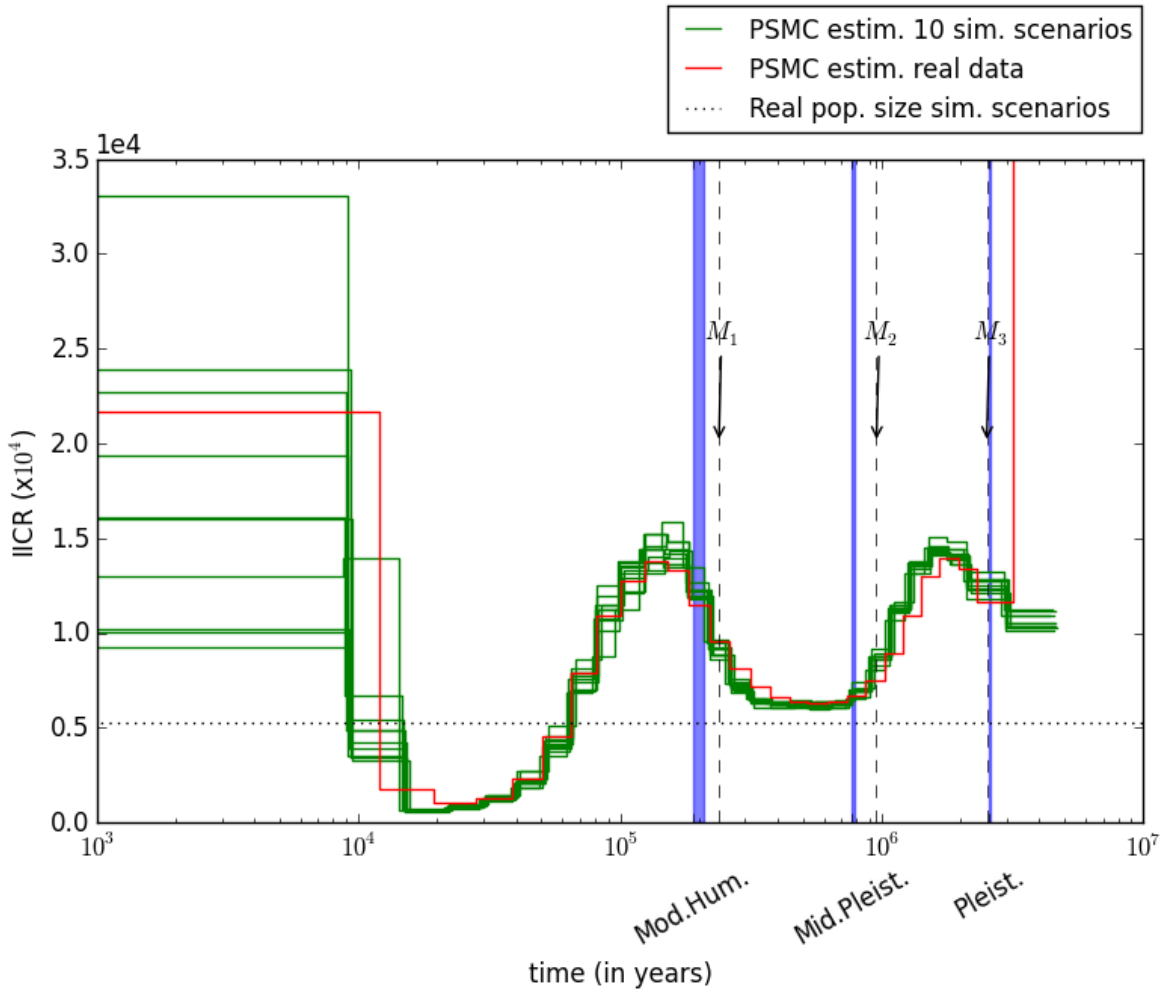
Considering alternative scenarios

PSMC in human data. Alternative scenario: n islands model with changes in gene flow and recent grow



Considering alternative scenarios

PSMC in human data. Alternative scenario: n islands model with changes in gene flow and recent grow



Summary statistic in an ABC framework

Presentation Outline

Conclusions

Identifiability problem when using T_2 (T_k) based methods.

Reconstructed skyline plots are not meaningless but should be interpreted carefully (as an IICR) in a general case.

For some models it is possible to have an explicit expression of the IICR.

The IICR can be used as a summary statistic to explain the data under any model, provided we can simulate coalescence times.

The IICR can be used as a summary statistic in an ABC framework.

Collaborators

Olivier Mazet
INSA - Toulouse

Lounès Chikhi
EDB - Toulouse

Simona Grusea
INSA - Toulouse

Simon Boitard
INRA - Toulouse

Patricia Santos
IGC - Lisbon

Didier Pinchon
Université Toulouse III

Acknowledgement

FUNDING

BiodiVERsA
INFRAGECO Project

The organizers

Robin Aguilée
Manon Costa
Grégory Faye
Sylvain Gandon
Sepideh Mirrahimi

Challenges when applying stochastic models to reconstruct the demographic history of populations.

Willy Rodríguez

Institut de Mathématiques de Toulouse

October 11, 2017

Challenges when applying stochastic models to reconstruct the demographic history of populations.

Willy Rodríguez

Institut de Mathématiques de Toulouse

October 11, 2017