Joint Inference of Demography and Selection from Genomic Temporal Data Using Approximate Bayesian Computation

Vitor A. C. Pavinato[1,2], Stéphane De Mita[3], Jean-Michel Marin[2,4], Miguel Navascués[1,4]

[1]UMR CBGP, INRA

[2]UMR IMAG, Université de Montpellier

[3]UMR IAM, INRA

[4]IBC

Postdoc InterLabex - ABCSelection

The Confounding effect of demography and selection[1,2]

The co-estimation of demographic parameters and selection is a long-standing difficulty in population genetics.

[1]Li *et al* (2012); [2]Ewing & Jensen (2015)

The Confounding effect of demography and selection[1,2]

The co-estimation of demographic parameters and selection is a long-standing difficulty in population genetics.

The common approach is to assume that selection is **LOCALIZED** in the genome and that demography would leave a **GENOME-WIDE** signature.

[1]Li *et al* (2012); [2]Ewing & Jensen (2015)

The Confounding effect of demography and selection[1,2]

The co-estimation of demographic parameters and selection is a long-standing difficulty in population genetics.

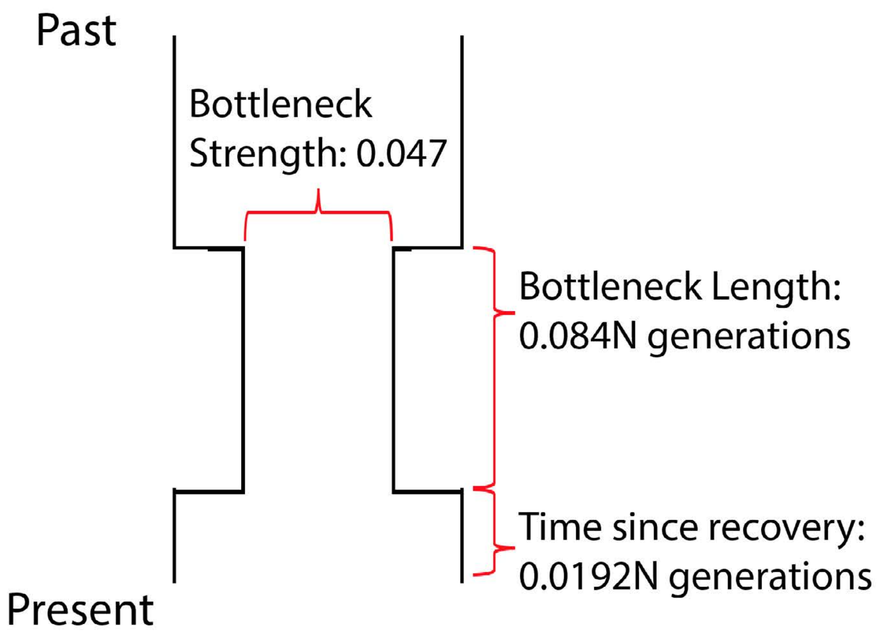
The common approach is to assume that selection is **LOCALIZED** in the genome and that demography would leave a **GENOME-WIDE** signature.

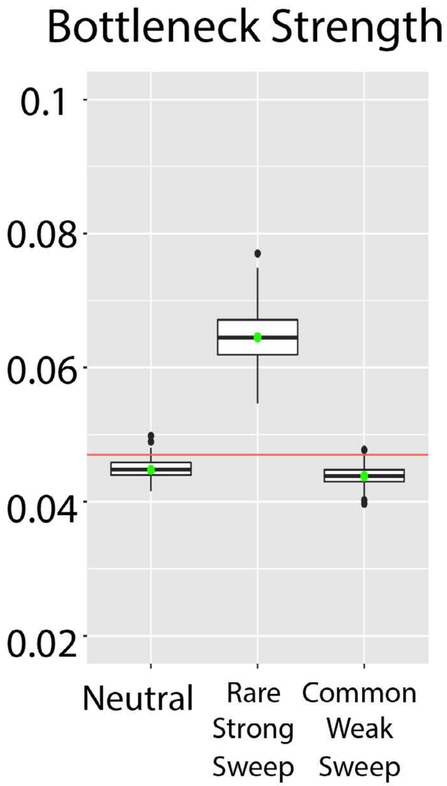
Recent works highlight the **PERVASIVE** role of selection, questioning the universal pertinence of such approach.

[1]Li *et al* (2012); [2]Ewing & Jensen (2015)

Impact of SELECTION on the Demography Inference

Recurrent Hitchhiking - RHH

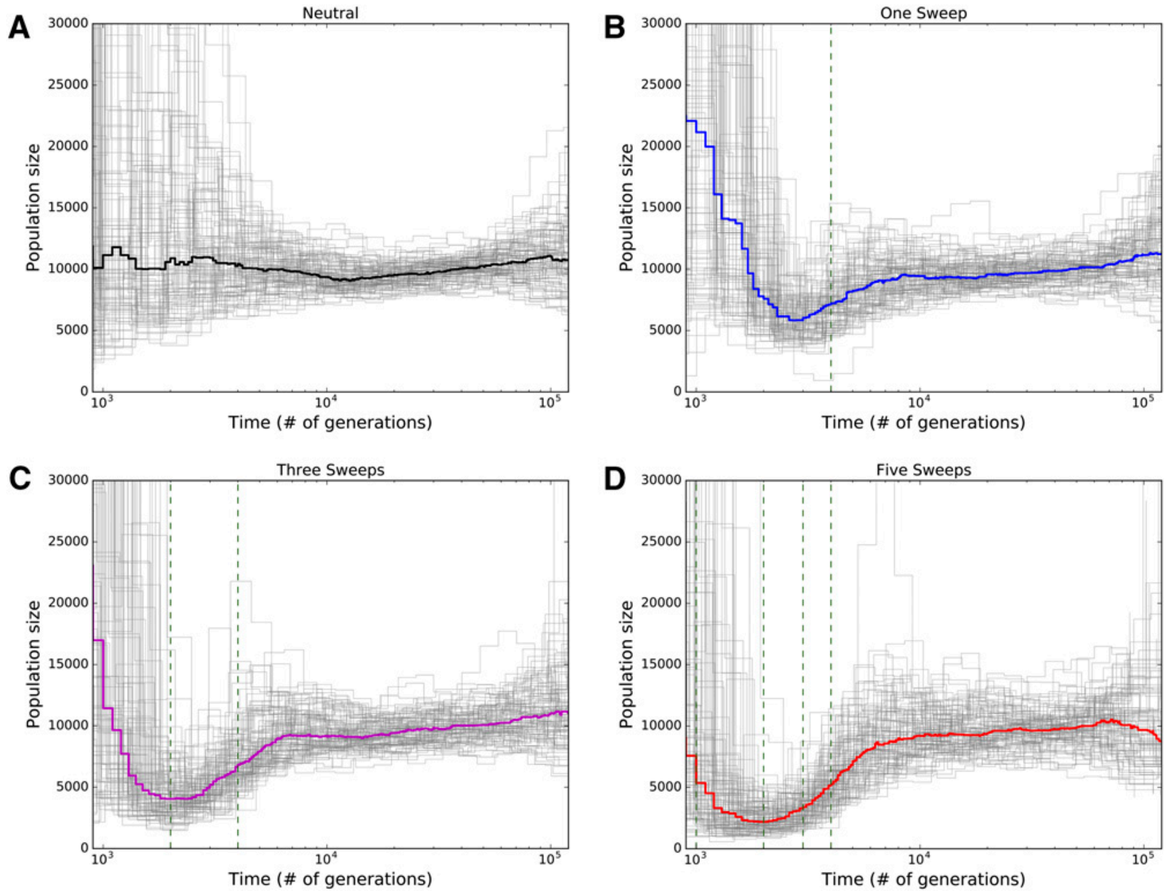




Jensen et al. (2008) Lange & Pool (2018)

Impact of SELECTION on the Demography Inference

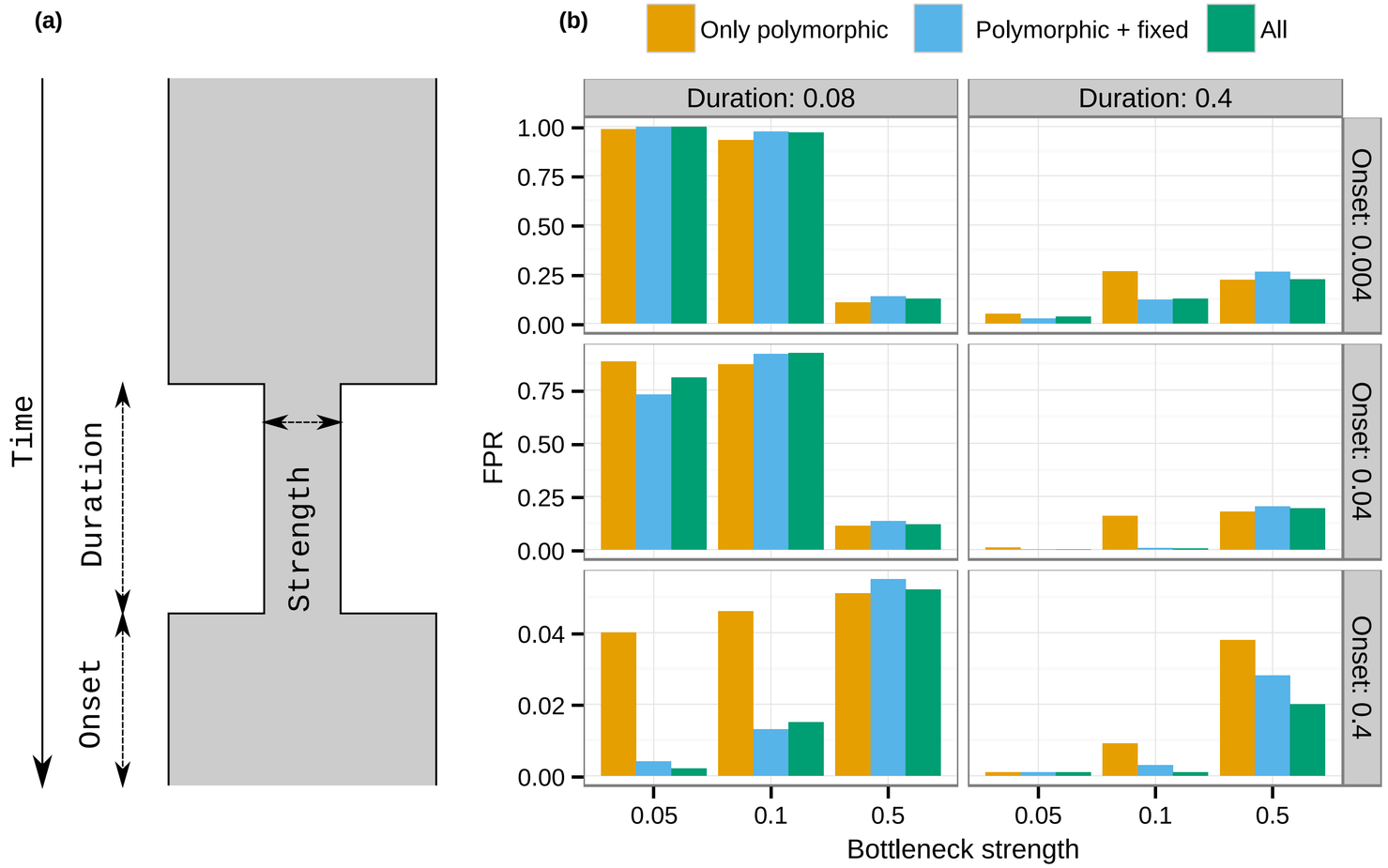
Pairwise Sequentially Markovian Coalescent (PSMC)



Schrider et al. 2016

Impact of DEMOGRAPHY on the Detection of Selection

SWEEPFINDER (SFS-based method)



Huber et al. 2016

METHODS* to jointly infer demography and selection

Table 1 Summary of the methods presented in the paper whose aim is to jointly estimate selection and demography or estimate selection while controlling for demographic effects

Methods	Strength	Weakness	References
Combining summary statistics	Ease of use	Sensitive to both demography and selection	Grossman *et al.* (2010)
Machine-learning algorithms	Decrease in the number of false positive	Same as above	Pavlidis *et al.* (2010) Lin *et al.* (2011)
Likelihood models	Optimal use of the data. Closest approach to a true joint analysis of demography and selection	Limited to simple models	Williamson *et al.* (2005) Li & Stephan (2006) Nielsen *et al.* (2009)
Approximate Bayesian computation	Easy to implement and can consider realistic models	Approximate method	Tavaré *et al.* (1997) Pritchard *et al.* (1999) Beaumont *et al.* (2002)
Unbalanced tree	Low sensitivity to demography	So far limited to completed sweeps and selection on standing variation with lowfrequency	Li (2011)

Li *et al* (2012)

METHODS to jointly infer demography and selection

Simulation of complex dynamics

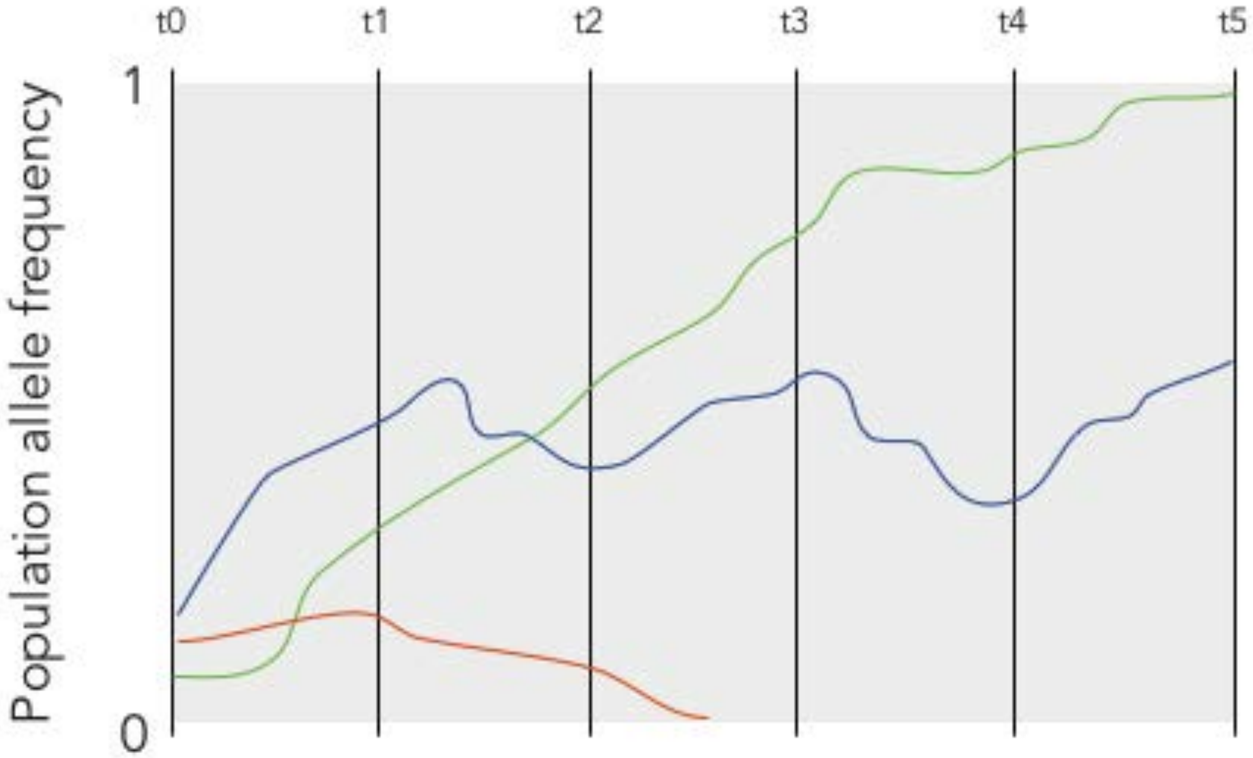
METHODS to jointly infer demography and selection

Simulation of complex dynamics

Likelihood-free approaches

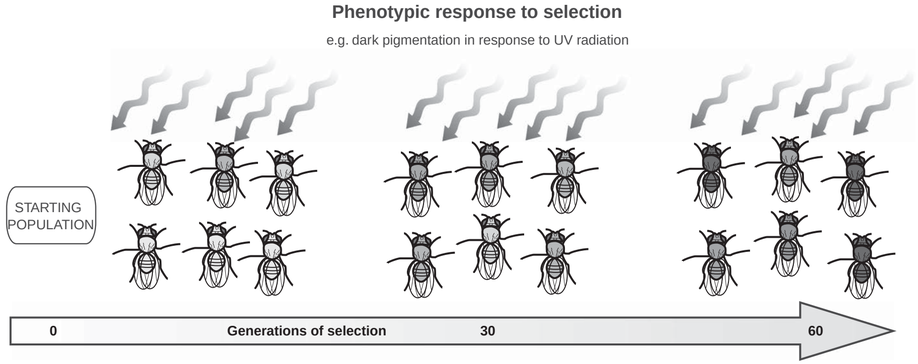
The "power" of temporal datasets

"We can see evolution in action"



Time-series datasets

Stringer C (2016)



Schlötterer et al (2015)

Traditional ABC Framework	ABC-RF[1,2] Framework
Requires a large number of simulations	**Reference table 10-100x less simulations**
Requires the choice of informative summary statistics	**Automatically find the more informative SSs**
Requires to define a tolerance level for acceptance	**Not dependent of tolerance level**

[1]Pudlo *et al* (2016), [2]Raynal *et al* (2017)

ABC-RF[1,2] Framework: Joint Inference of **Demography** and **Selection** in Temporal Data

Genome-wide pattern of

DEMOGRAPHY

and

SELECTION

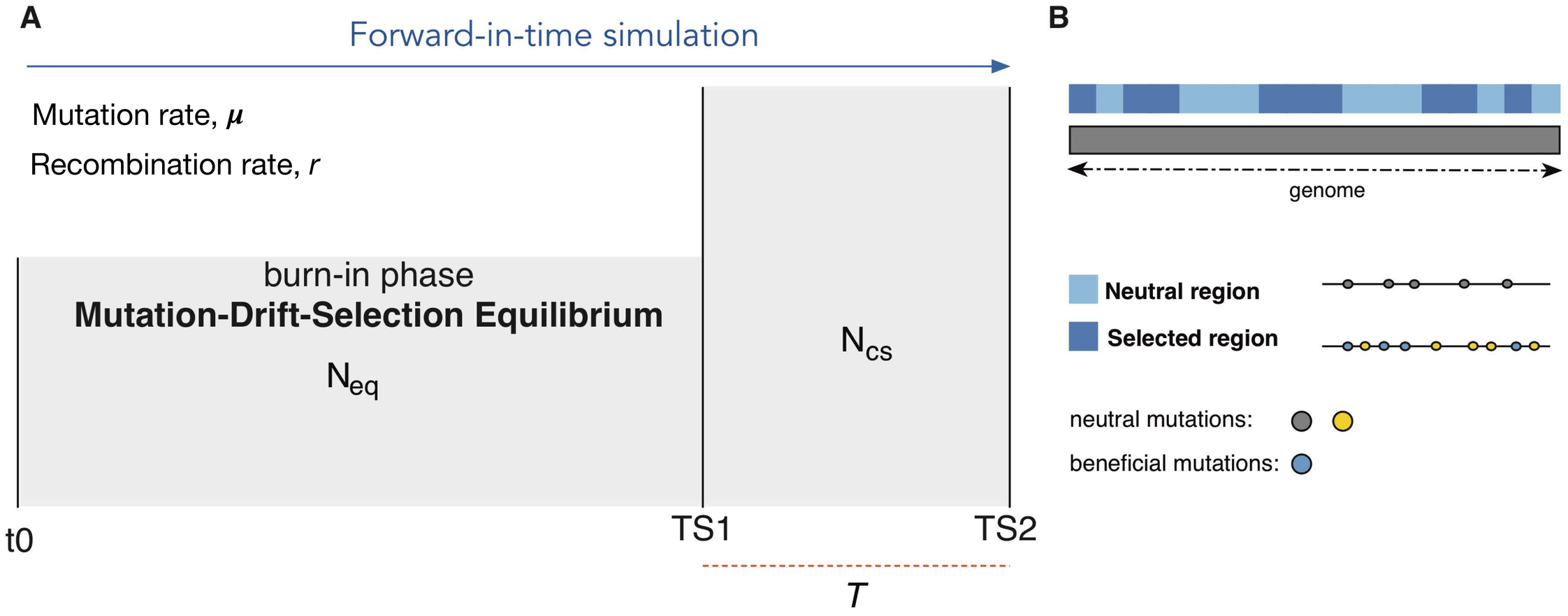
[1]Pudlo *et al* (2016), [2]Raynal *et al* (2017)

Model

Mean for the DFE ~Γ ($\kappa\Theta$, Θ)

Pr Selected regions, P_R

Pr Beneficial Mutations, P_S



Summary Statistics

Locus-specific:: single site

$$H_E, D_j, WCF_{ST}$$

Locus-specific:: windowed

$$S, \pi, \theta_W, Tajimas'D, Da, ZZ, Z_{ns}$$

Global

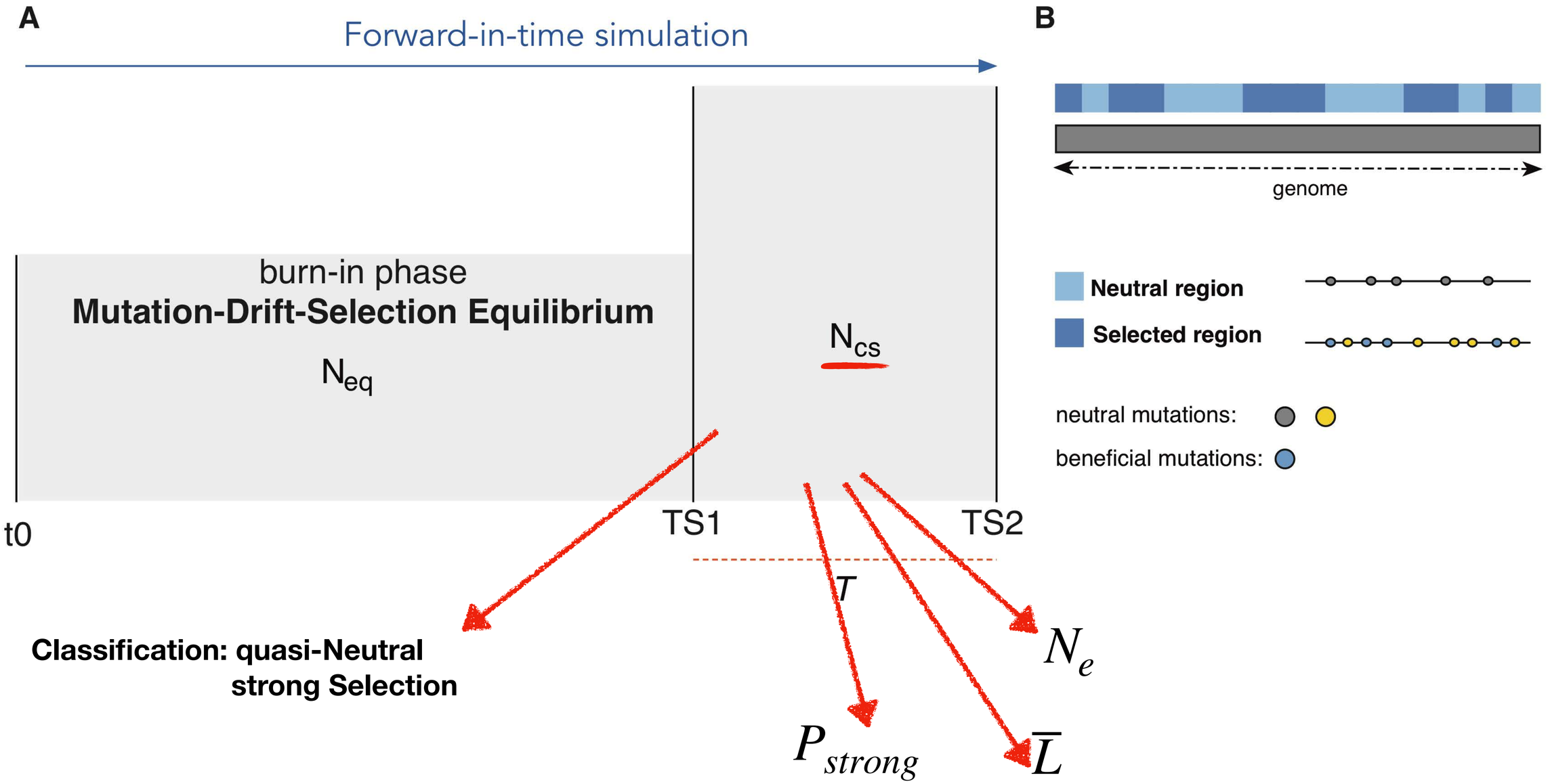
$$H_E, D_j, WCF_{ST}$$

$$S, \pi, \theta_W, Tajimas'D, Da, ZZ, Z_{ns}$$

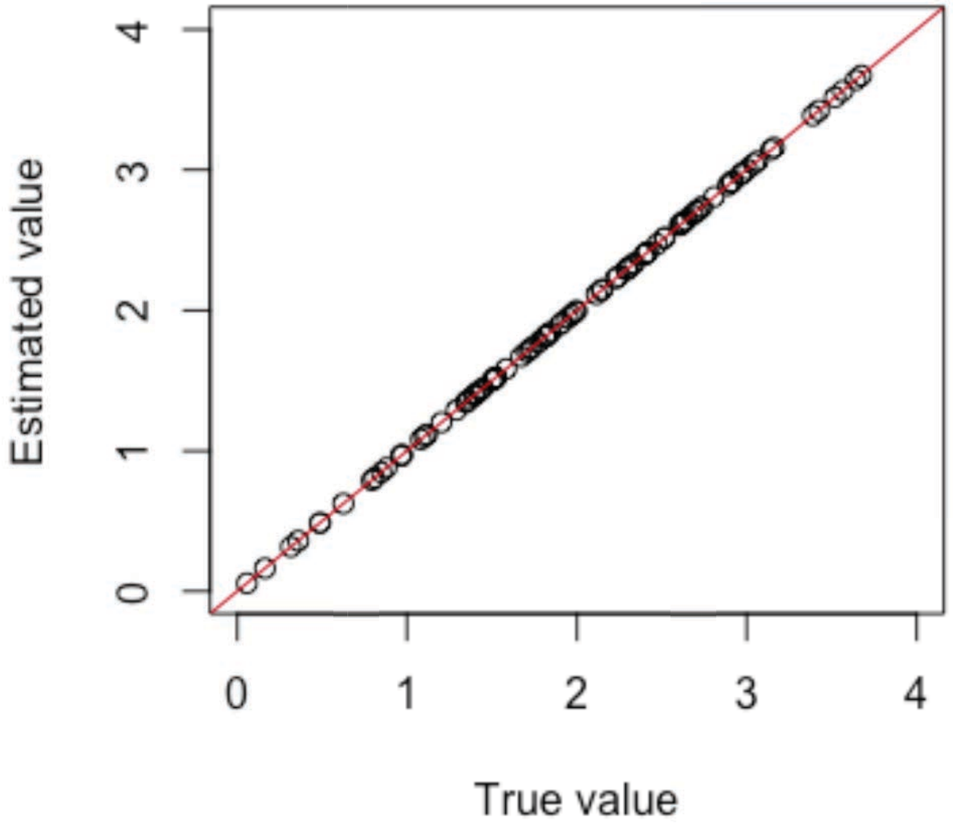
$$SFS$$

$$Mean, Var, Kurtosis, Skewness, 5\,\%\,, 95\,\%\, quantiles$$

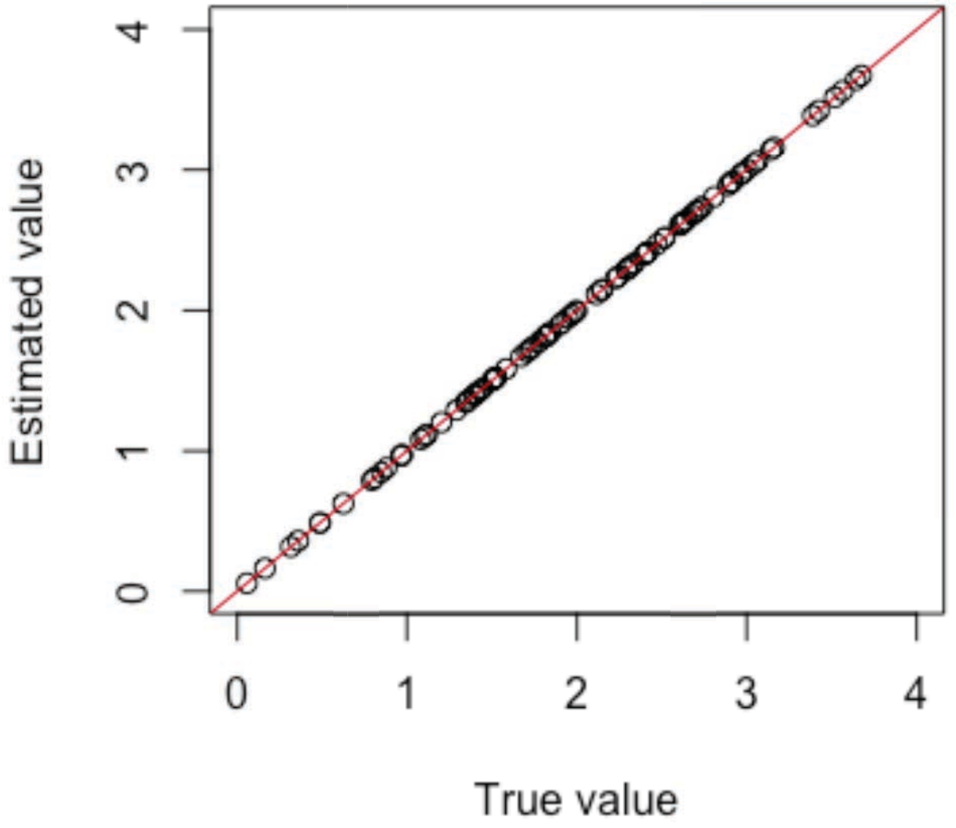
Posterior Estimates and Inference

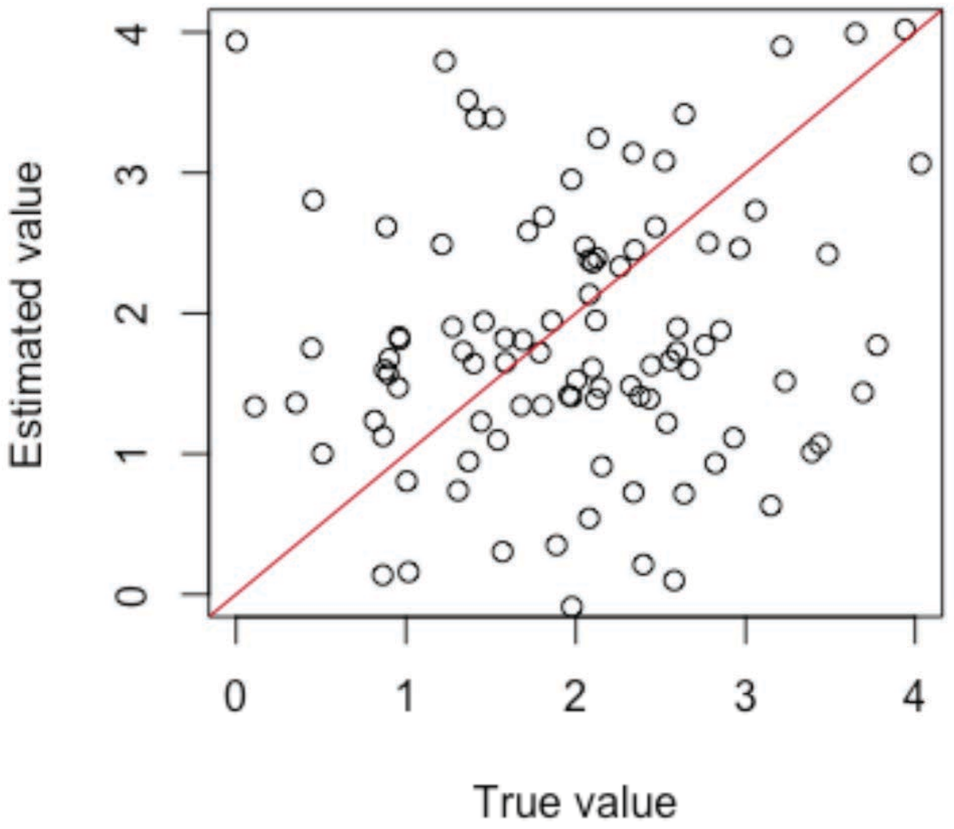


Evaluating ABC-RF Performance



Evaluating ABC-RF Performance





Characterizing Demography

Effective Population Size N_e

Census Size N_{cs}

Demography: Effective Population Size

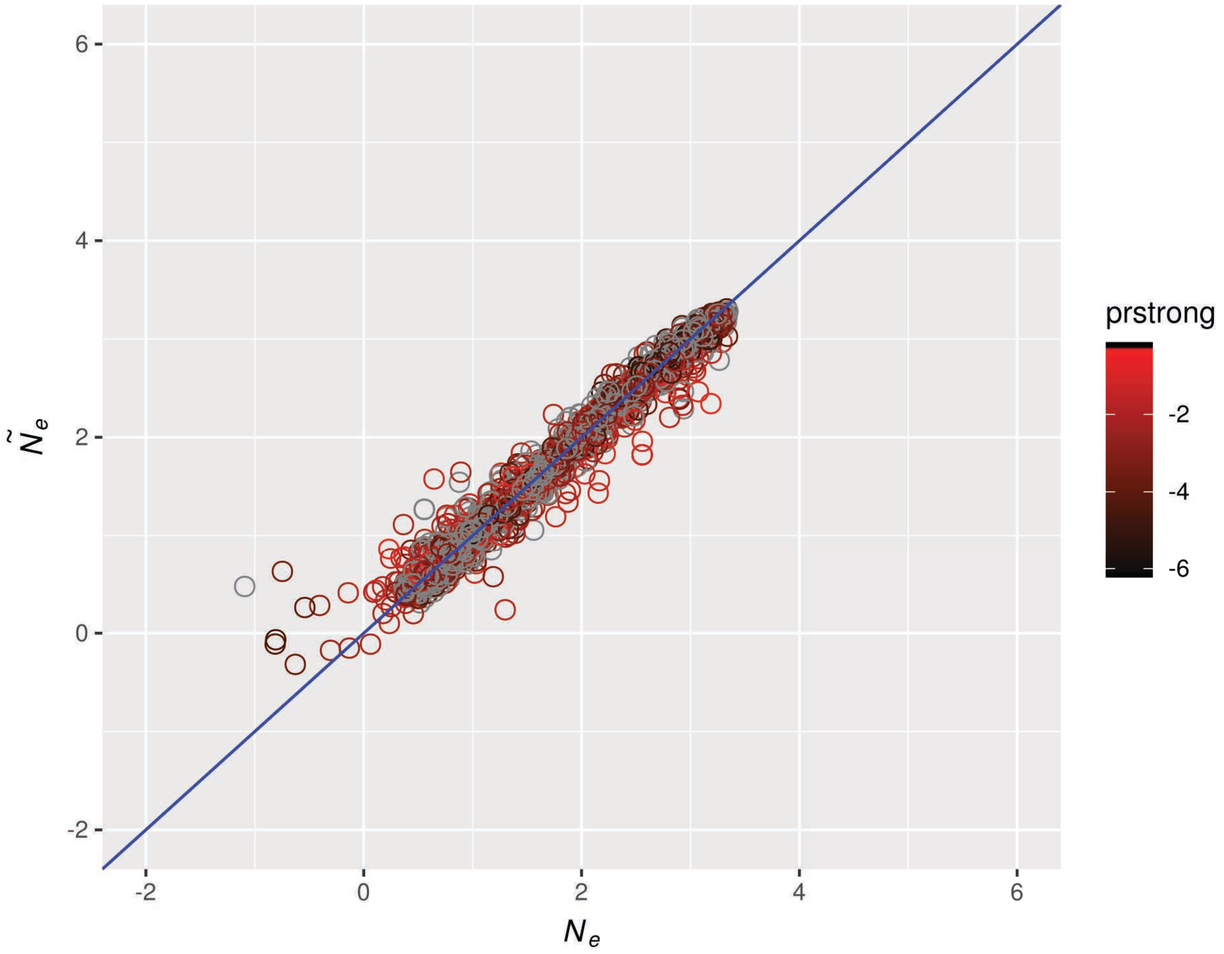
$$N_e = \frac{4N_{cs}}{2 + var(gametes)}$$

Gametes are the contribution of each individual in the generation g_i - 1 (parents)

$$\frac{1}{N_e} = \frac{1}{N_1} + \frac{1}{N_2} + \frac{1}{N_n}$$

Demography: Effective Population Size

ABC-RF



WFABC

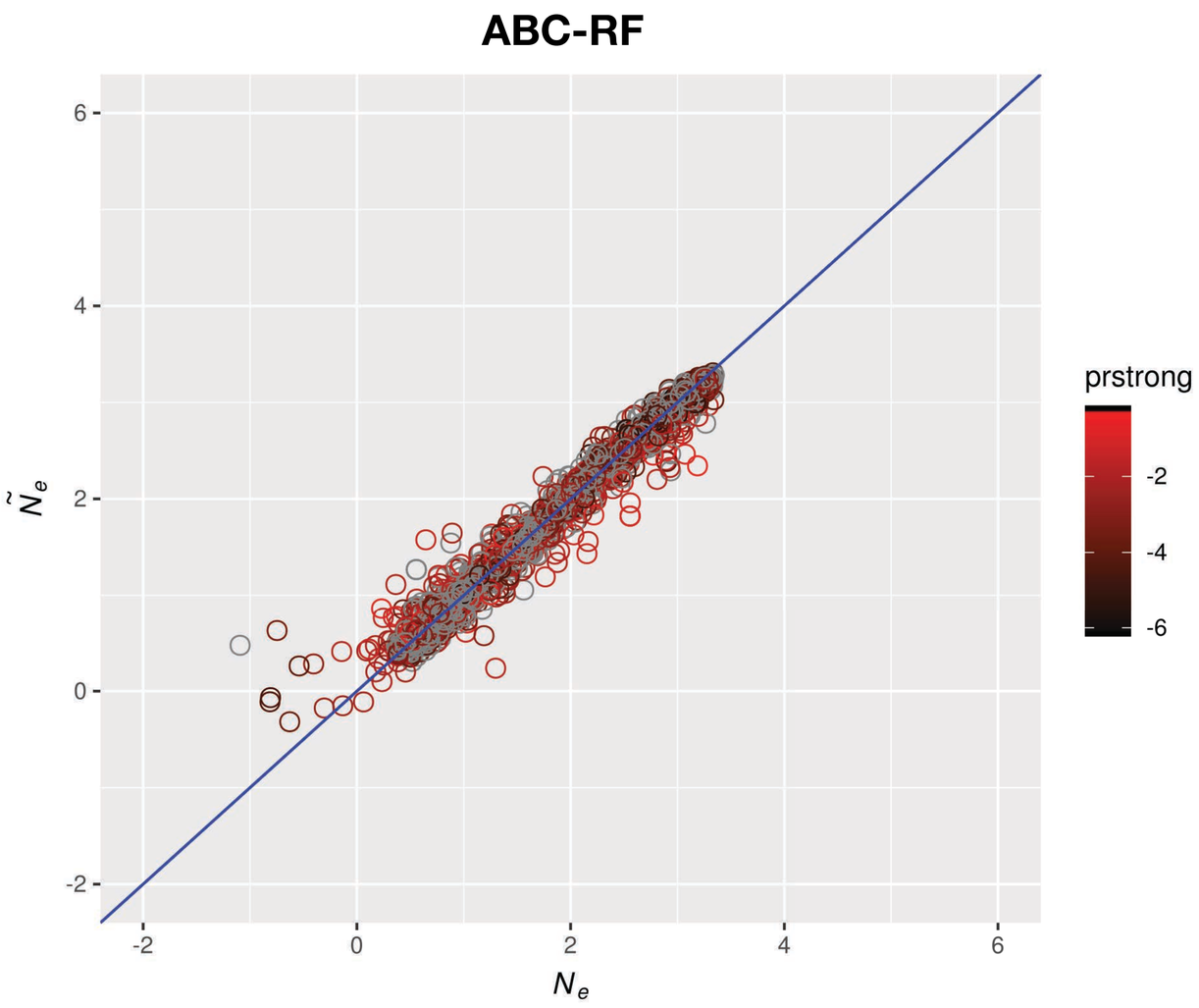
Implementation of 2-steps ABC (Bazin, Dawson & Beaumont 2010)

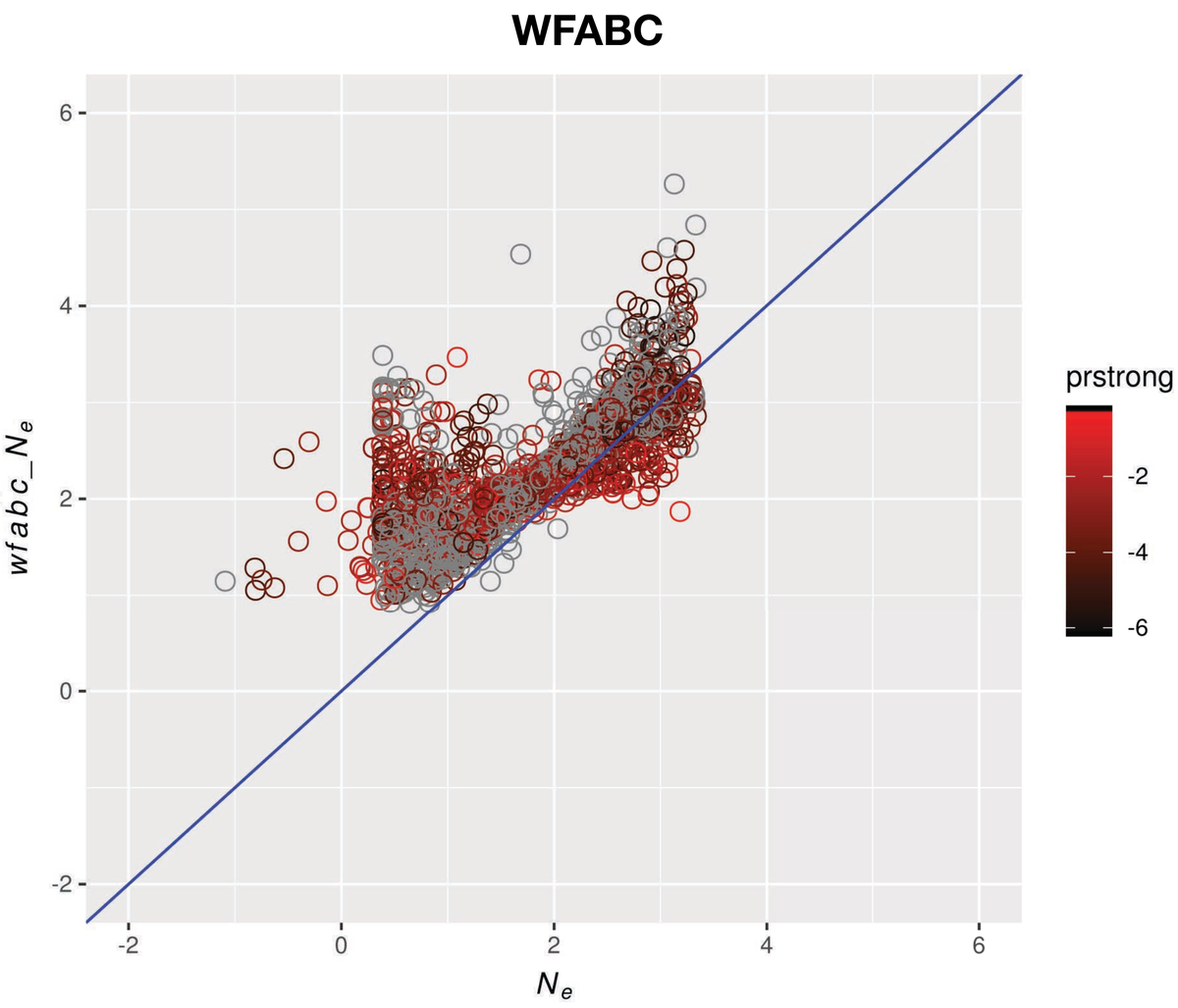
First step - Infer demography - N_e

Second step - Infer selection coefficients
Foll et al. 2014, 2015

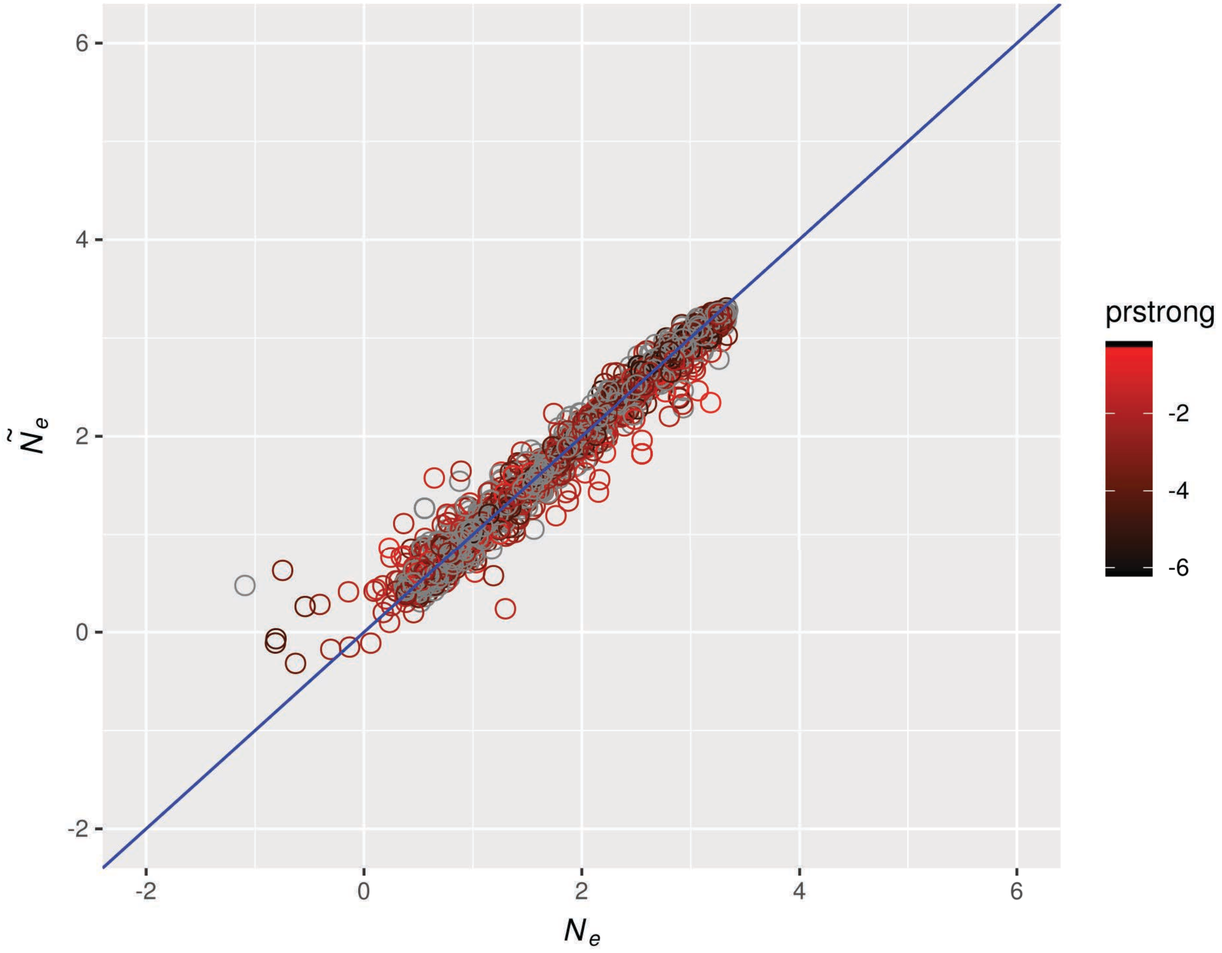
Calculation of a summary statistics *Fs'* (Jordan & Rayman 2007)

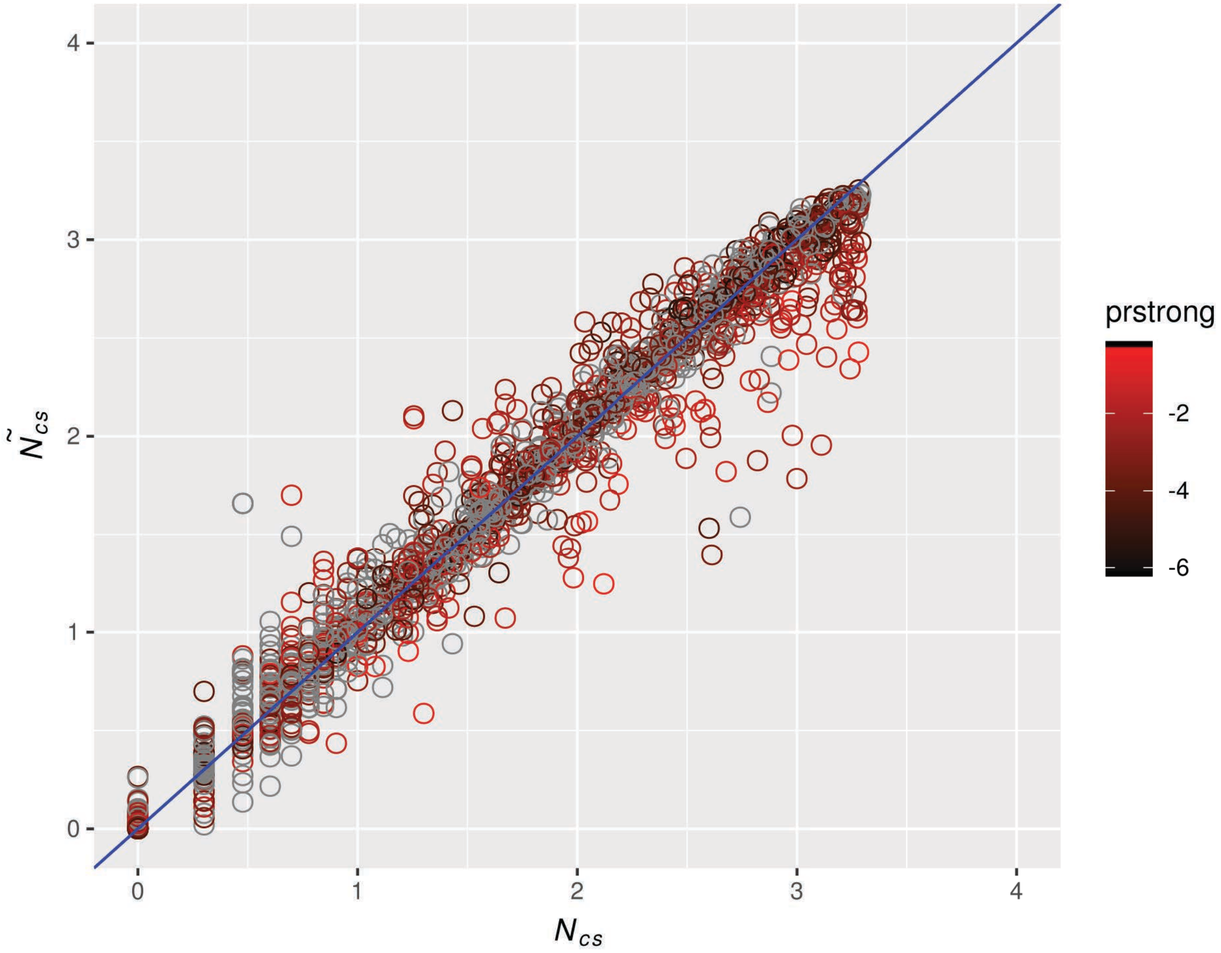
Demography: Effective Population Size



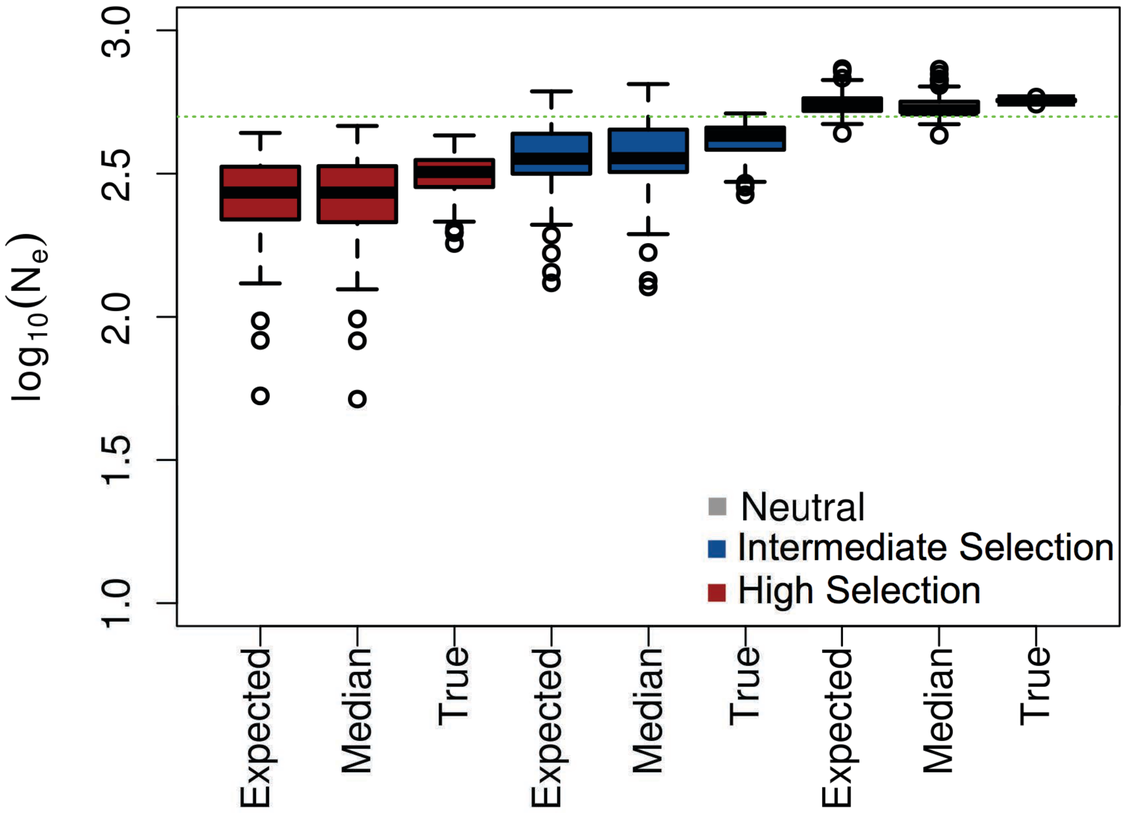


Demography: Census Size

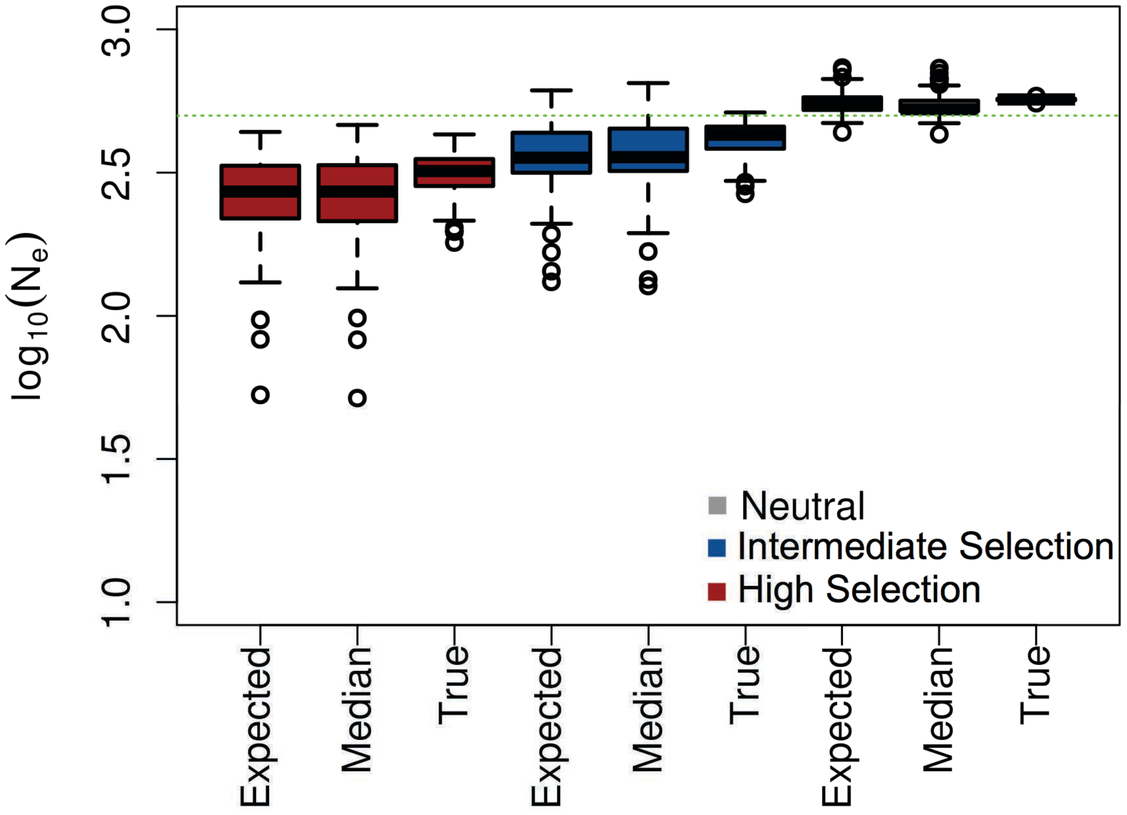


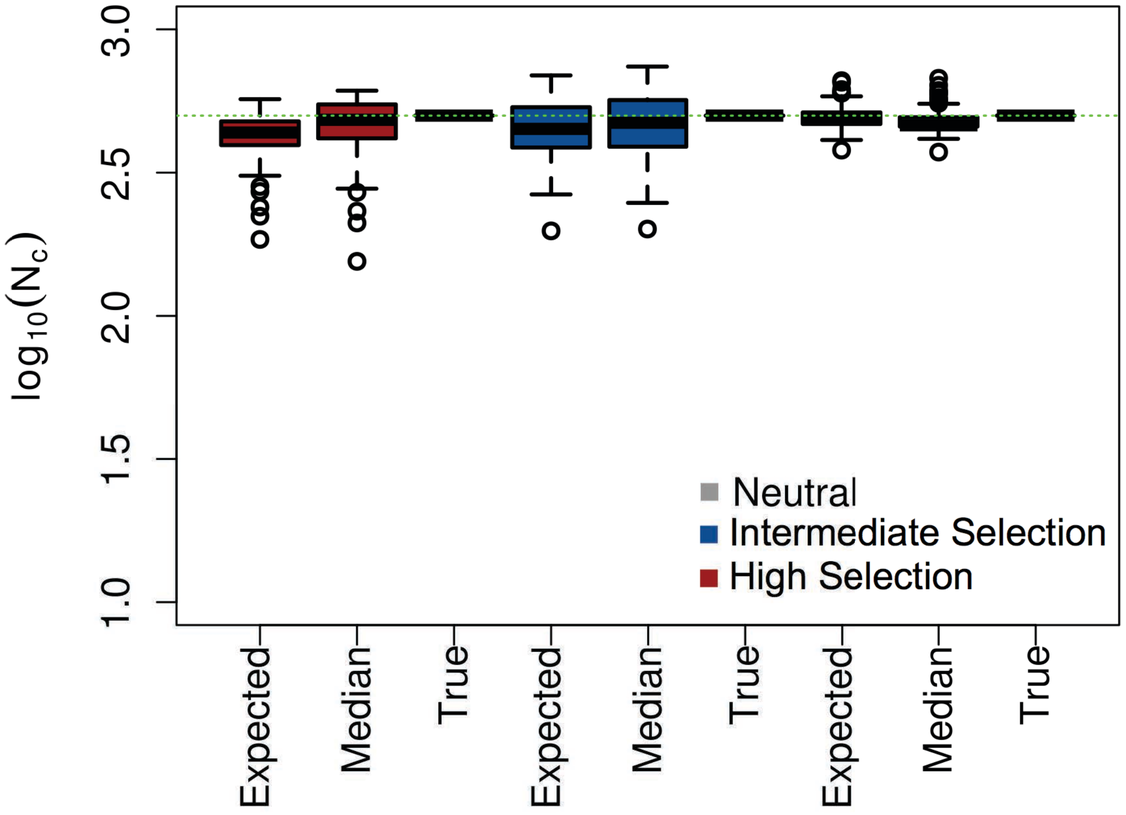


Demography: Effective Population vs Census Size



Demography: Effective Population vs Census Size





Characterizing Selection

Classification: quasi-Neutral and strong Selection

Average Genetic Load $\overline{L}$

$$L = \frac{w_{max} - \overline{w}}{w_{max}}$$

Substitution Load, "the cost of natural selection"[1]

Proportion of Strongly Selected Mutations

$$P_{strong} = \frac{Mutations[Ns > 1]}{Mutations}$$

[1]Haldane (1957)

Classification: “Neutral" vs “Selection"

Proportion of Strongly Selected Mutations

$$P_{strong} = 0$$

quasi-Neutral

$$P_{strong} > 0$$

strong Selection

Classification: “Neutral" vs “Selection"

	qNeutral	sSelection	error
qNeutral	709	164	0,188
sSelection	201	803	0,200

Classification: “Neutral" vs “Selection"

“Neutral” and **“Selection”** dynamics

Neutral **Selection**

It is a continuum determined by:

$$\theta_s = 4N_e\mu_b$$

$$\mu_b = P_R P_S \mu$$

$$w_i = 1 + s_i$$

Selection Dynamics

$$\theta_s = 4N_e\mu_b$$

Rate at which beneficial mutations enter the simulation

"Controls **how long** the population **must wait** to produce a beneficial mutation"

Adaptation

"Mutation Limited"

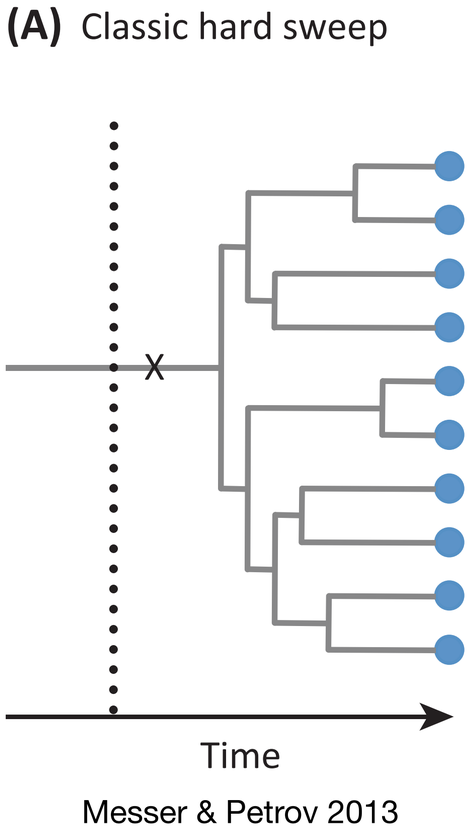
$$\theta_s < 1$$

Adaptation

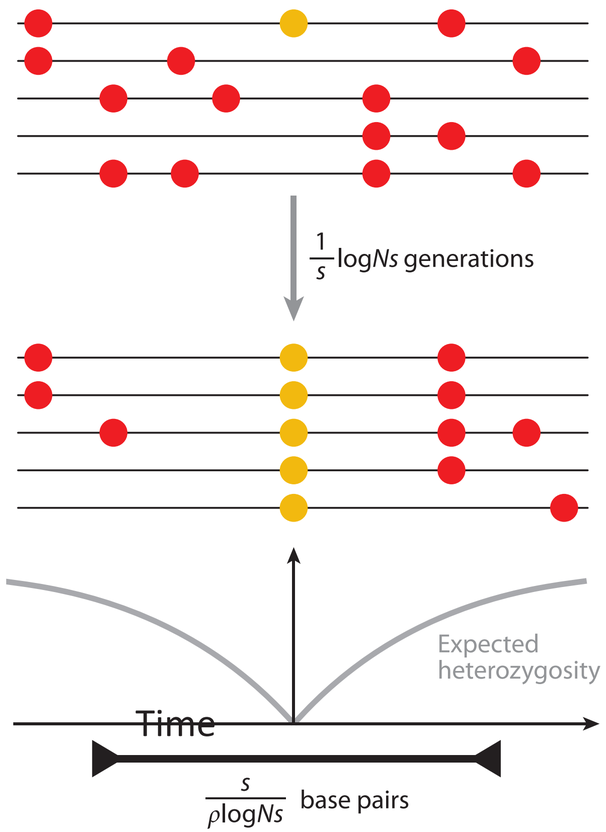
"Mutation Unlimited"

$$\theta_s > 1$$

Classic hard sweep



Messer & Petrov 2013



Neher 2013

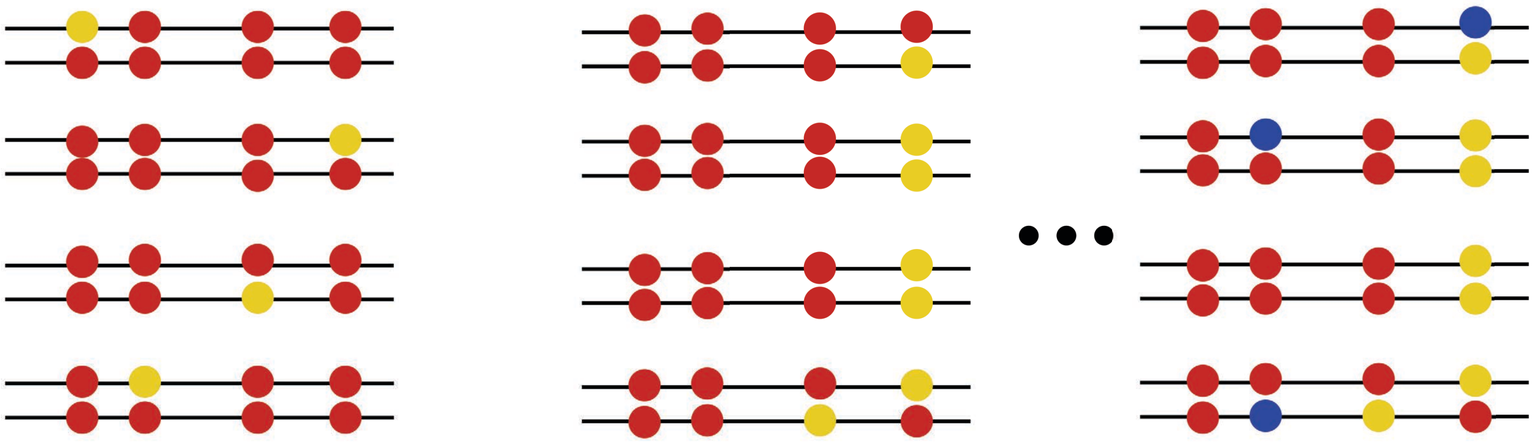
A single adaptive allele rises to high frequency **hitchhiking genetic neighbors that also fix** in the population.

The r**atio of selection strength and recombination rate** governs the distance on the chromosome from the adaptive site with depressed diversity following a sweep.

Time

Hard sweep

Many beneficial mutations

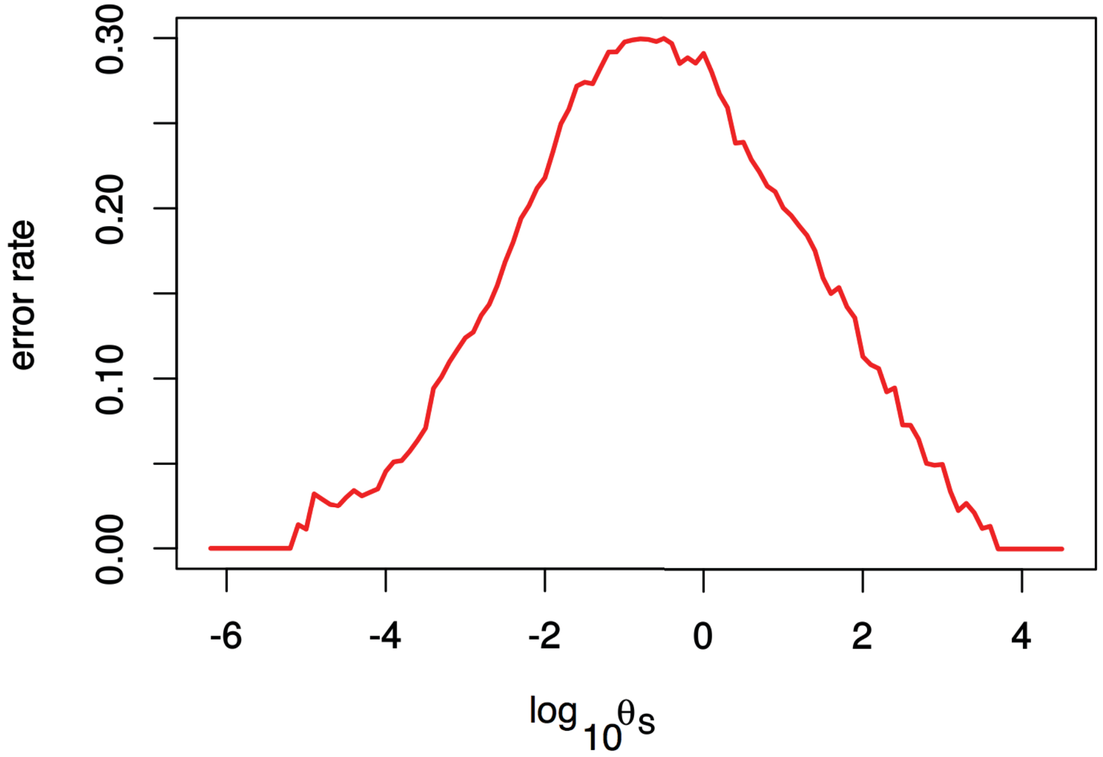


Classes: windows of θ_s

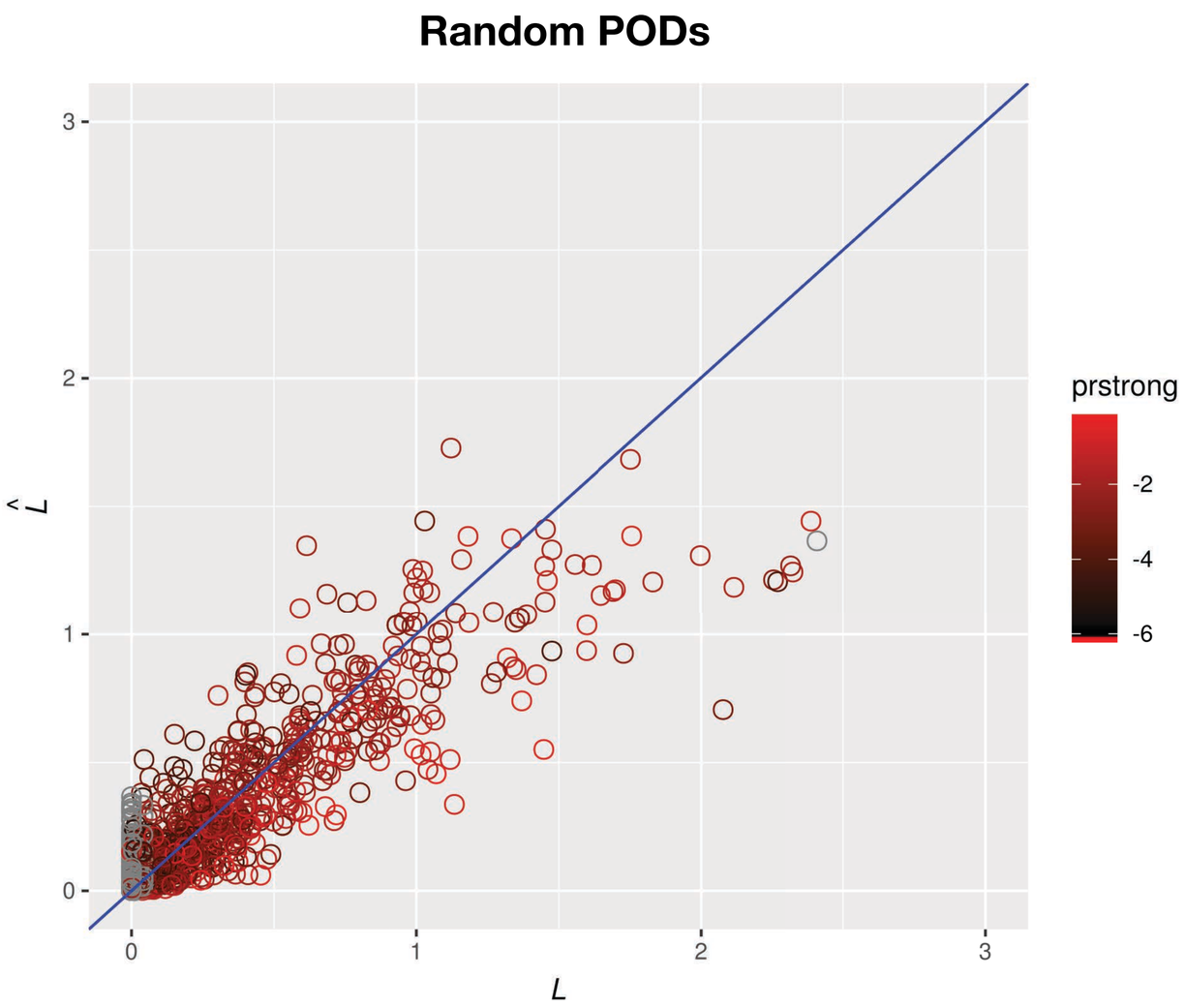
θ_s

Classes: windows of θ_s

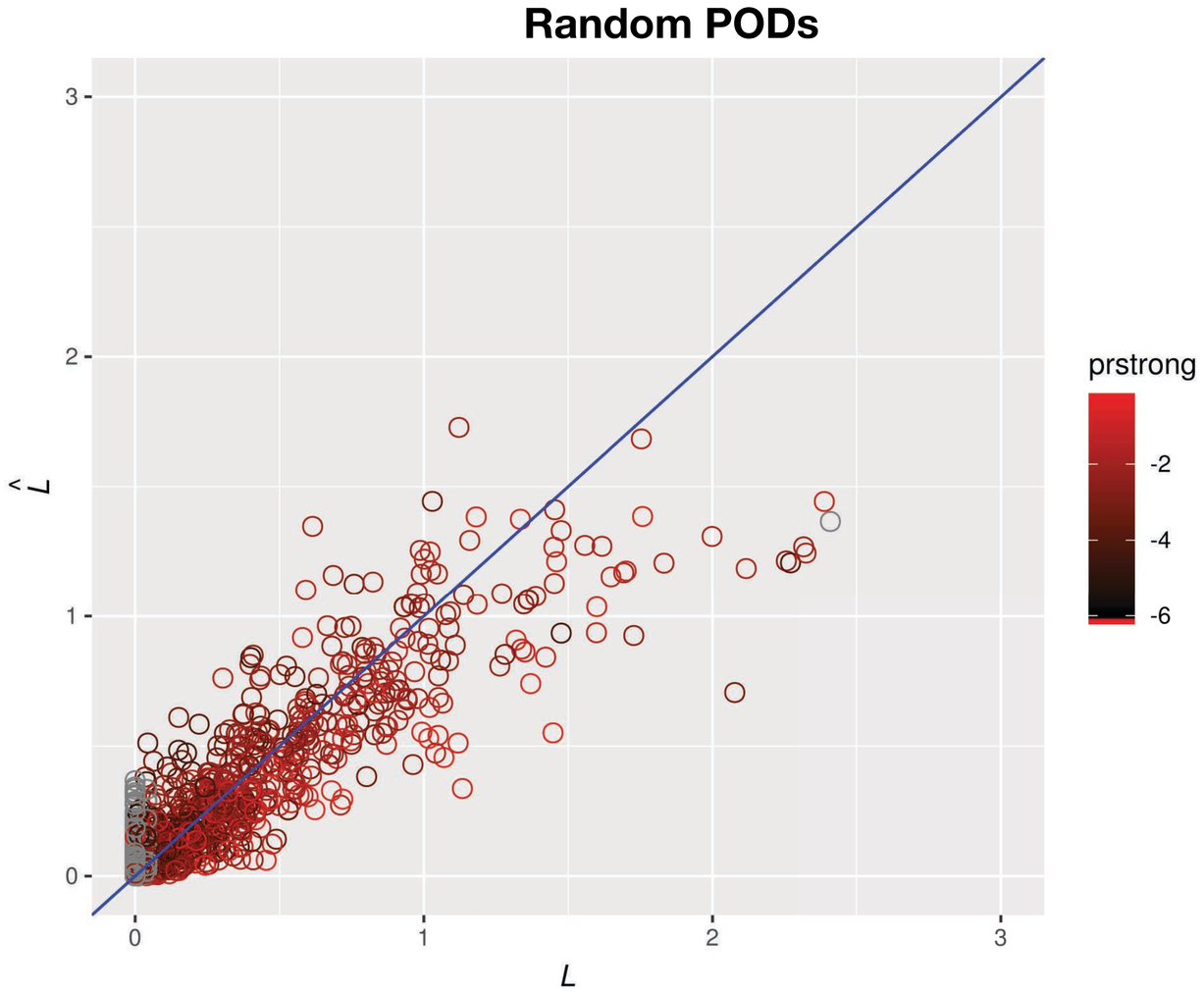
θ_s



Selection: Genetic Load



Selection: Genetic Load



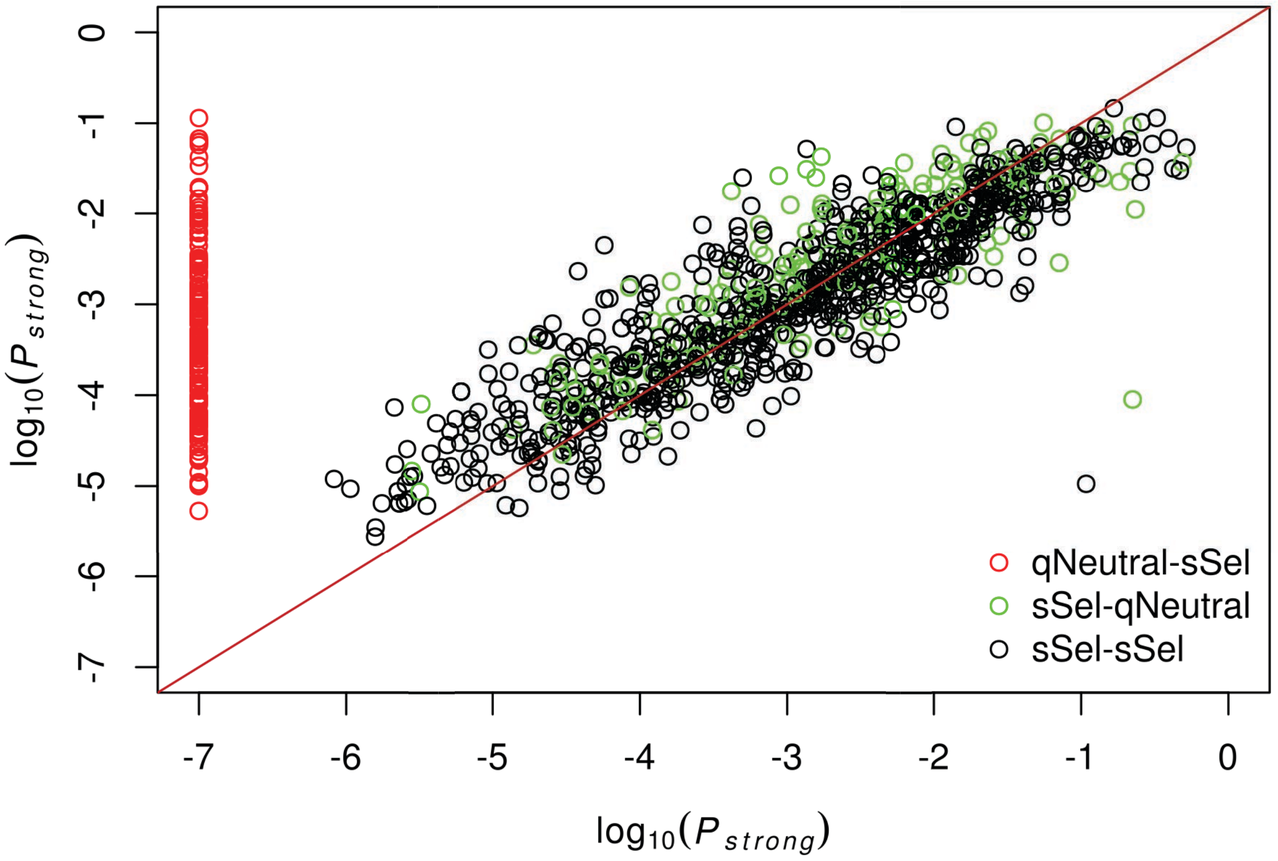
SMALL *L* but BIG P_{strong}

Simulation with lots of small effect mutations

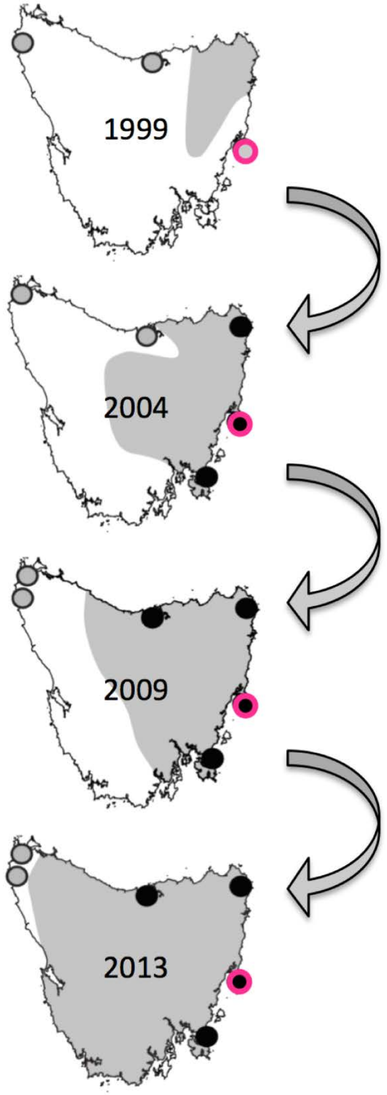
BIG *L* but SMALL P_{strong}

Simulation with lots of big effect mutations
It behaves as a Neutral/near-Neutral scenario

Selection: Proportion of Strongly Selected Mutations



Application



Temporal population genomics data of the Tasmanian Devil (*Sarcophilus harrisii)*

Samples before and after the emergence and spread of Devil Facial Tumor Disease (DFTD)

Low-coverage RADseq data

Adaptation is mutation limited

Soft sweep from Standing Variation (SV)

http://www.utas.edu.au/news/2016/2/18/41-securing-the-future-of-our-tasmanian-devil/

CONCLUSION

ABC-RF is able to jointly characterize **DEMOGRAPHY** and **SELECTION.**

HIGHLIGHTS

1) Characterize selection without additional information:

- mutation within genes;
- synonymous / non-synonymous information;
- without the position in the genome (scaffold or RADtag position)
  - **Can be applied in non-model organisms**

2) See the impact of selection on estimates of effective population size

3) Allow separating estimates of effective population and census size

PERSPECTIVES 1

For the moment, the model is very simple:

- *de novo* mutations - hard sweep;

PERSPECTIVES 1

For the moment, the model is very simple:

- *de novo* mutations - hard sweep;

Things to think about …

- *What is going to happen if we include background selection?*
- *How about selection on standing variation?*

PERSPECTIVES 2

“Dichotomy” between speed and accuracy

Small Genome: 100 Mb took 3 weeks to produce the reference table with 50,000 simulations for a scenario with *de novo* mutations

PERSPECTIVES 3

For the moment, the model is very simple:

- define two genomic regions: neutral and under selection is too simplistic;

PERSPECTIVES 3

For the moment, the model is very simple:

- define two genomic regions: neutral and under selection is too simplistic;

Things to think about …

- *How about more complex genomic backgrounds?*

PERSPECTIVES 4

The power of temporal data:

Allows us to use the information of the allele frequency changes to characterize selection.

This framework could be used in different settings? **Local adaptation**

ACKNOWLEDGMENTS

- Alexander Dehne-Garcia - UMR CBGP, INRA

- CBGP cluster

- Genotoul

- Génomique Statistique et Évolutive des Populations

THANK YOU!

Vitor Pavinato
vitor.pavi@gmail.com

Census Size

Effective Population Size

Genetic Load

Proportion of Strongly Selected Mutations

Priors

Table 1: Simulation parameters and their prior distribution

Parameter	Prior probability distribution
Mutation rate, μ	$\mu \sim log_{10}(Uniform)$
Recombination rate, r	$r \sim log_{10}(Uniform)$
Population size for the equilibrium phase, N_{eq}	$N_{eq} \sim log_{10}(Uniform)$
Population size for the interval, N_{cs}	$N_{cs} \sim log_{10}(Uniform)$
Mean for the DFE $\sim \Gamma(mean = \kappa\theta, shape = \theta)$	$\kappa\theta \sim log_{10}(Uniform)$
Proportion of the genome under selection:	
1) Proportion of regions under selection, P_R	$P_R \sim Uniform$
2) Probability of beneficial mutation , P_S	$P_S \sim log_{10}(Uniform)$

Evaluating ABC-RF Performance

1) “RANDOM” pseudo-observed data (PODs) from prior

2) “FIXED” PODs

Table 2: Simulation parameters for the PODs

Parameter	Neutral	Intermediate Selection	High Selection
μ	$1e-7$	$1e-7$	$1e-7$
ρ	$5.0e-7$	$5.0e-7$	$5.0e-7$
N_{eq}	500	500	500
N	500	500	500
DFE $mean = \kappa\theta$	NA	0.1	0.1
PrGWSel	NA	0.1	0.25
PrMSel	0	0.1	0.1

ABC Random Forests

Random Decision Forests

Ensemble methods to build ***predictive models*** for both **CLASSIFICATION** and **REGRESSION**

RANDOM FORESTS creates an entire **“FOREST"** of ***uncorrelated decision trees***

Random Forests

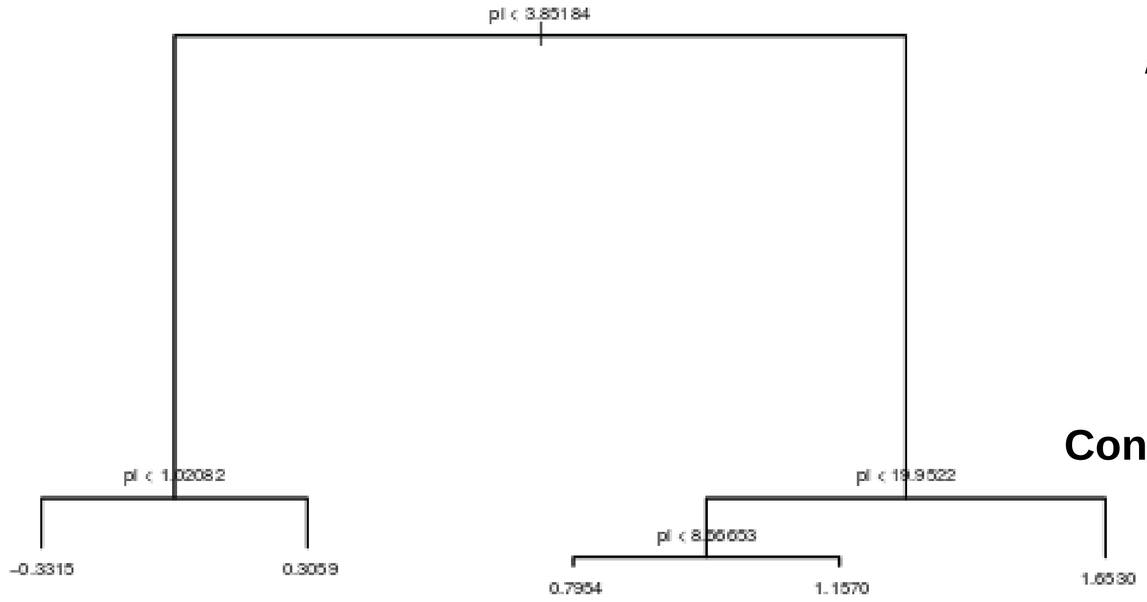
$$model \sim \pi$$

Classification

Constant size
C

Population decline
D

Random Forests



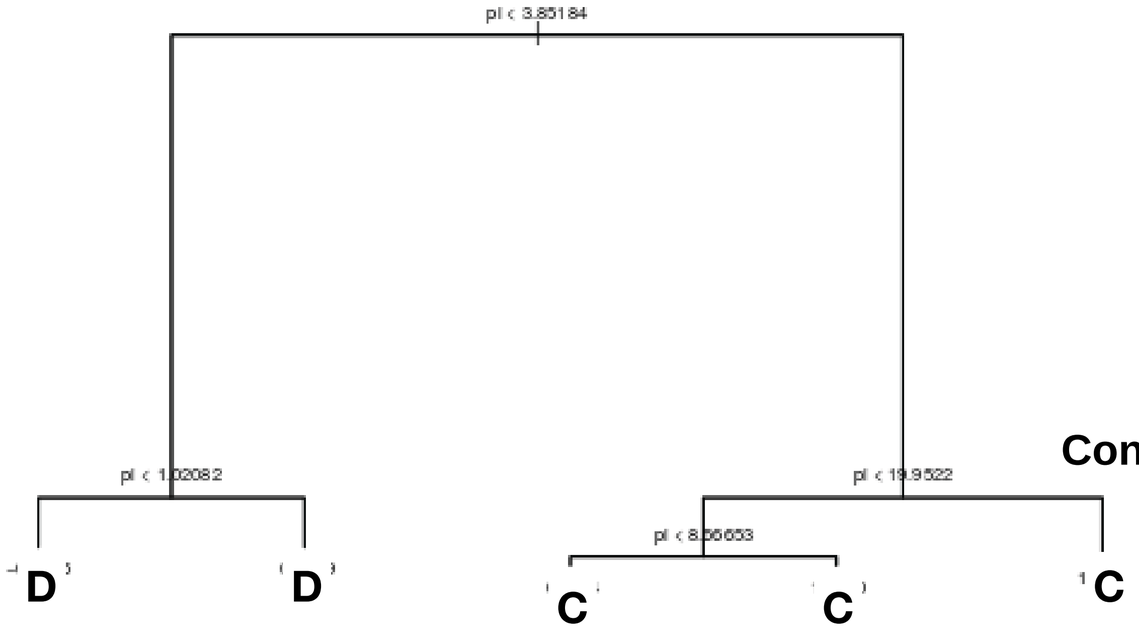
$$model \sim \pi$$

Classification

Constant size
C

Population decline
D

Random Forests



$$model \sim \pi$$

Classification

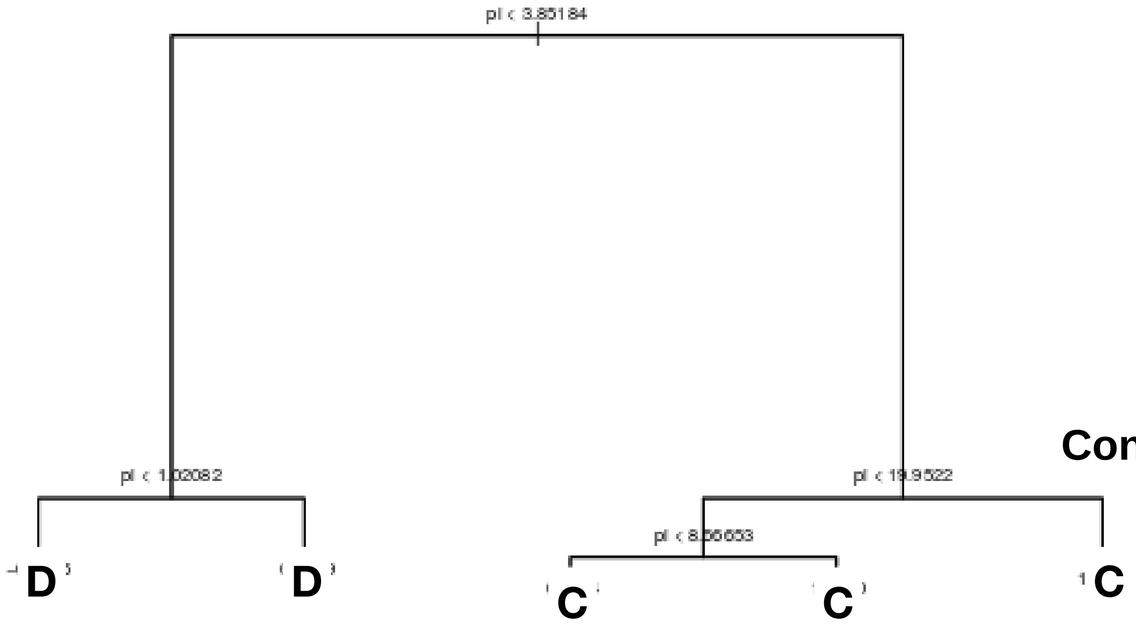
Constant size
C

Population decline
D

Random Forests

$$log_{10}(\theta) \sim \pi$$

Regression



$$model \sim \pi$$

Classification

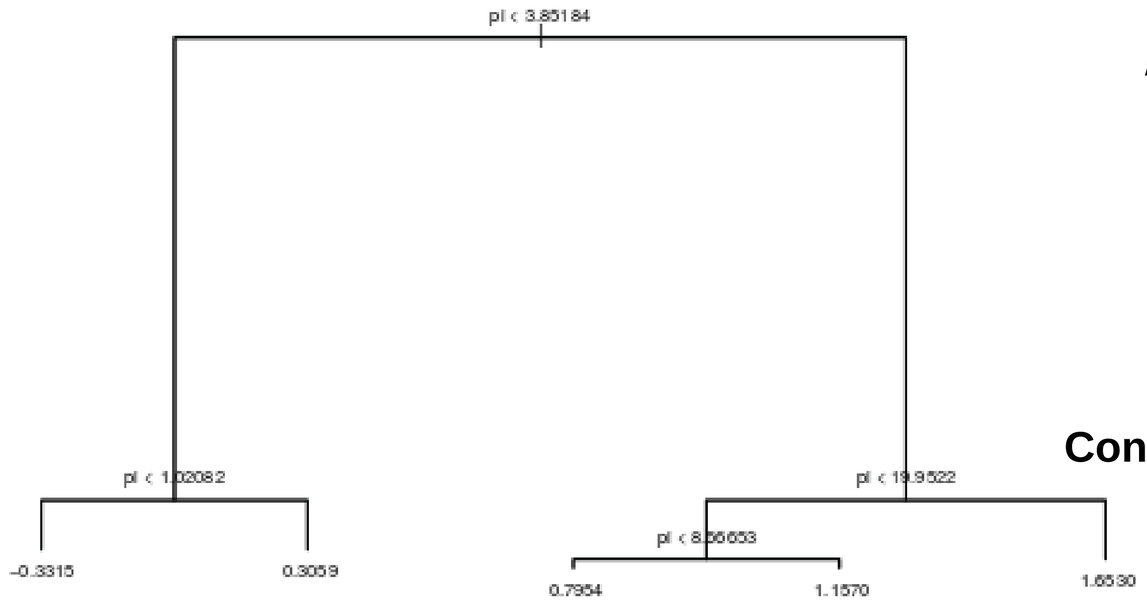
Constant size
C

Population decline
D

Random Forests

$$log_{10}(\theta) \sim \pi$$

Regression



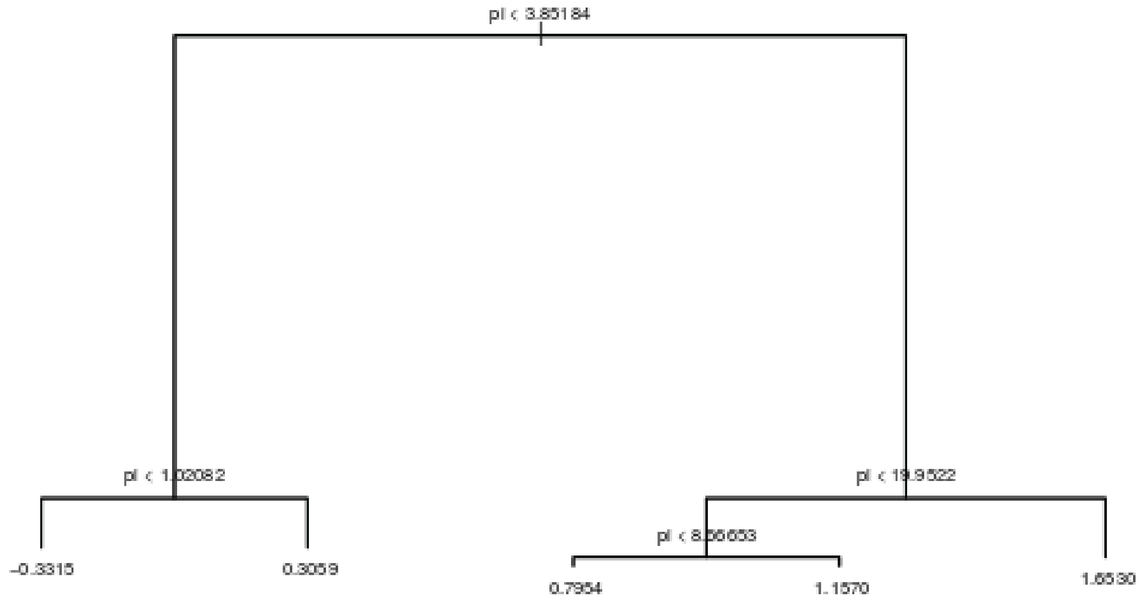
$$model \sim \pi$$

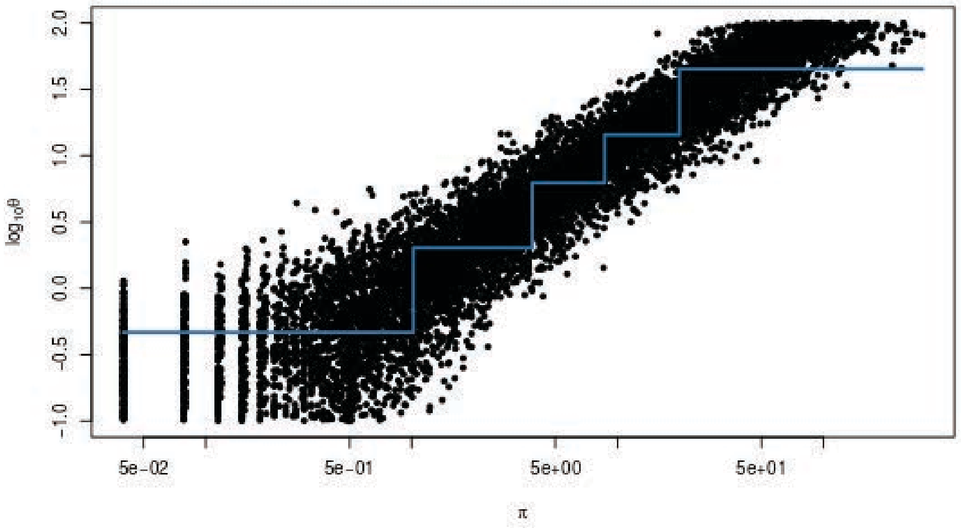
Classification

Constant size
C

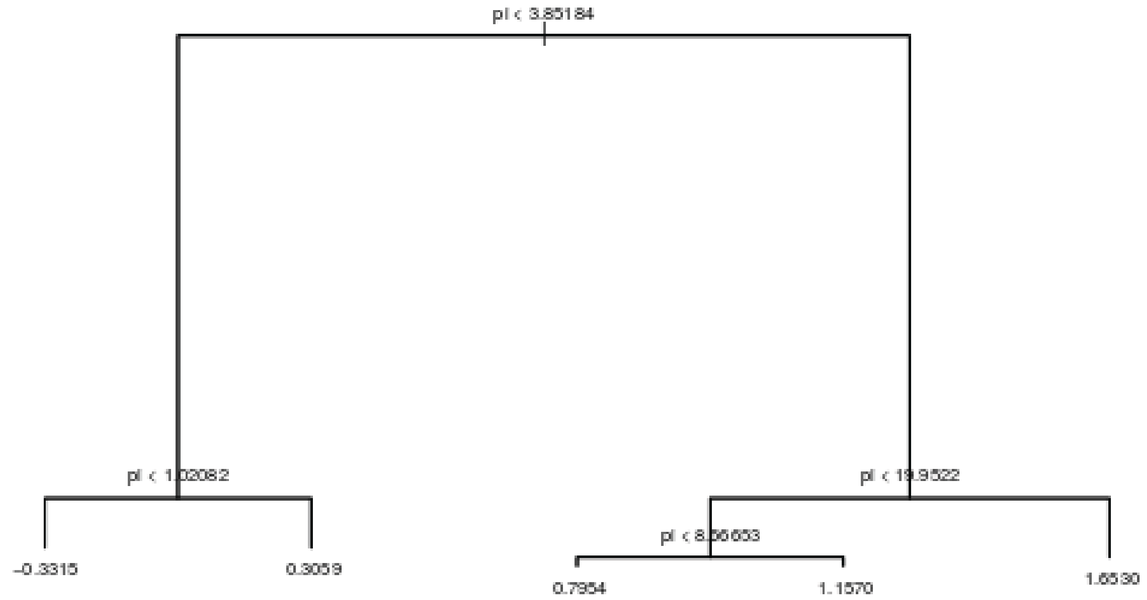
Population decline
D

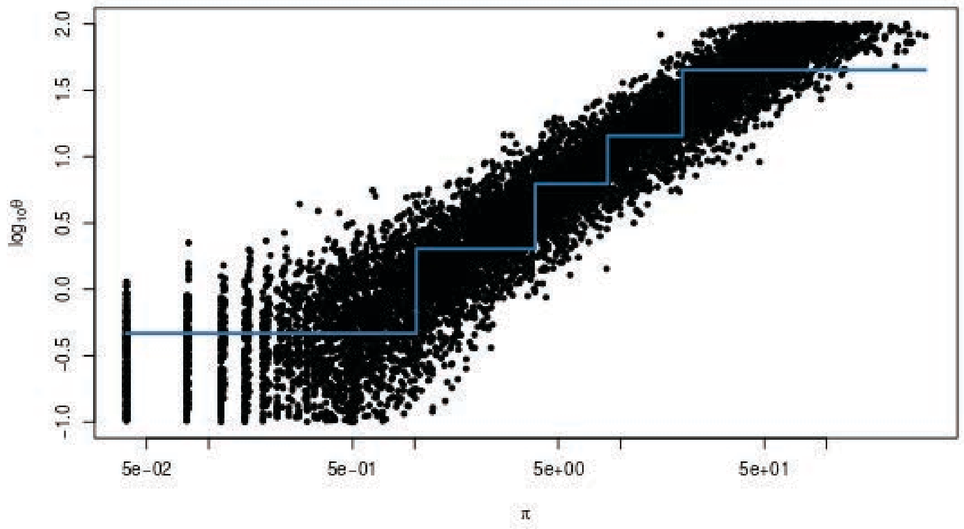
Random Forests

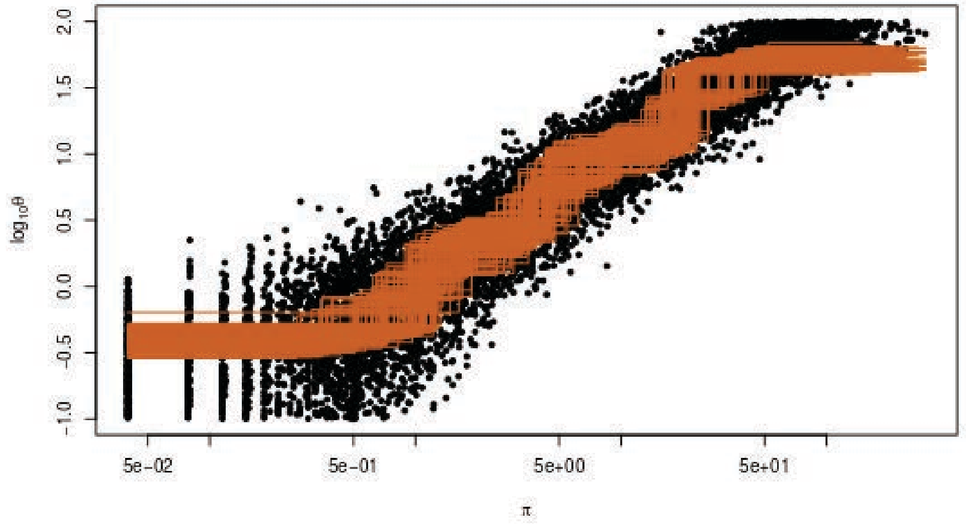




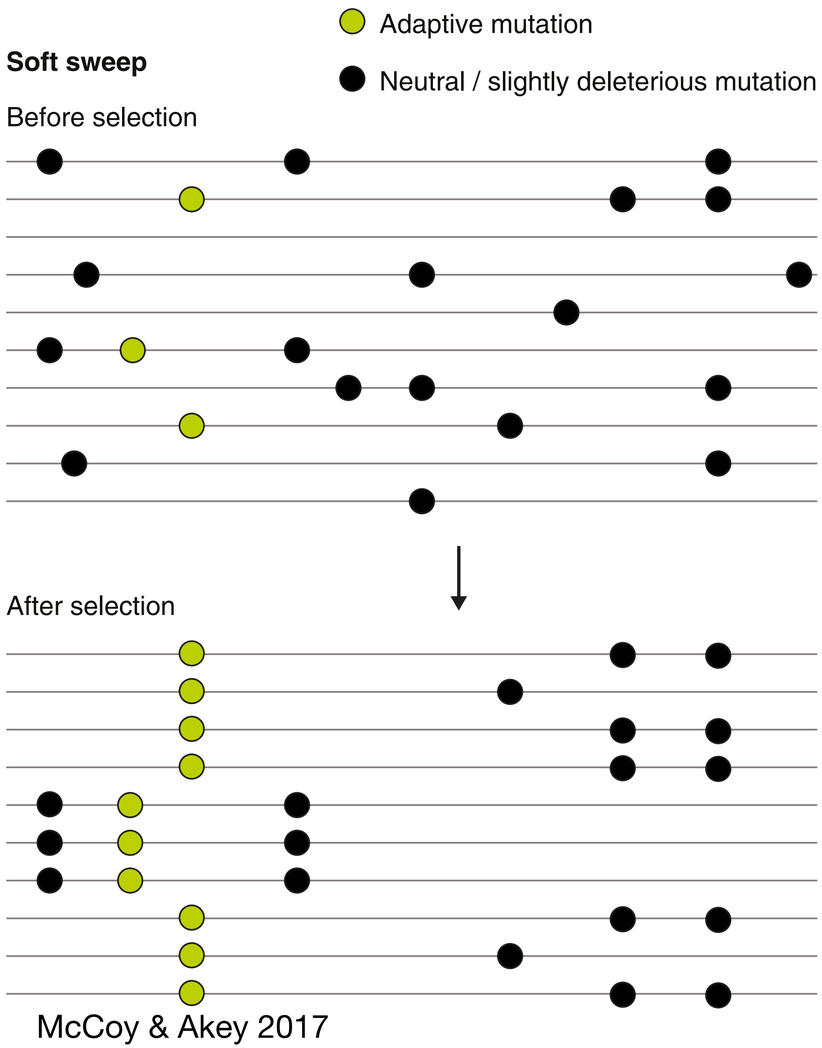
Random Forests







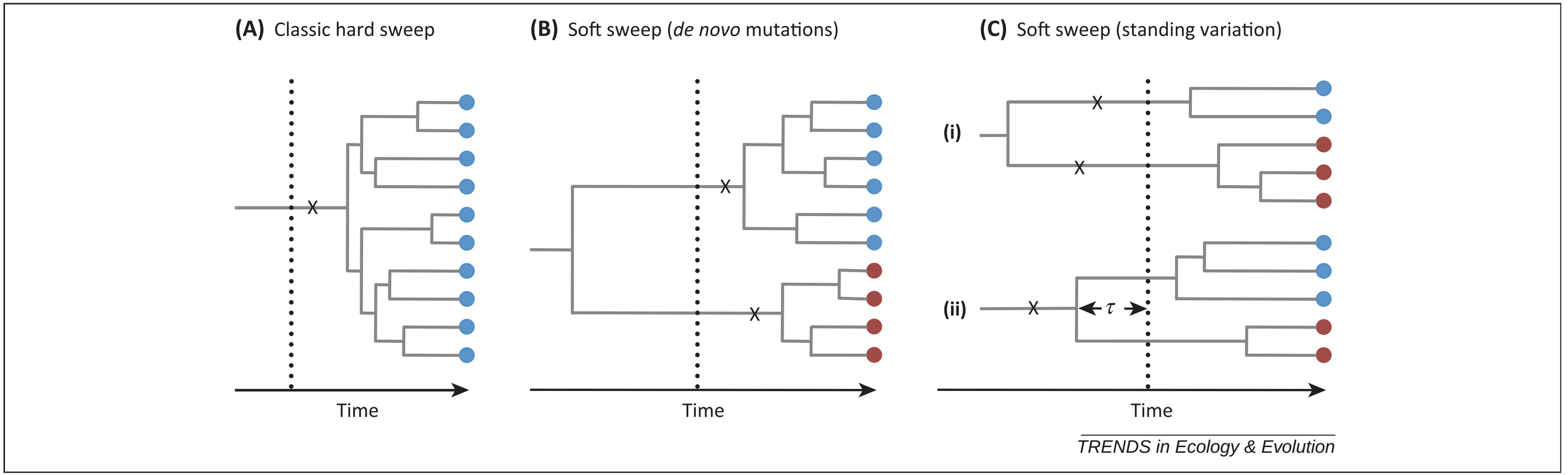
Soft sweep



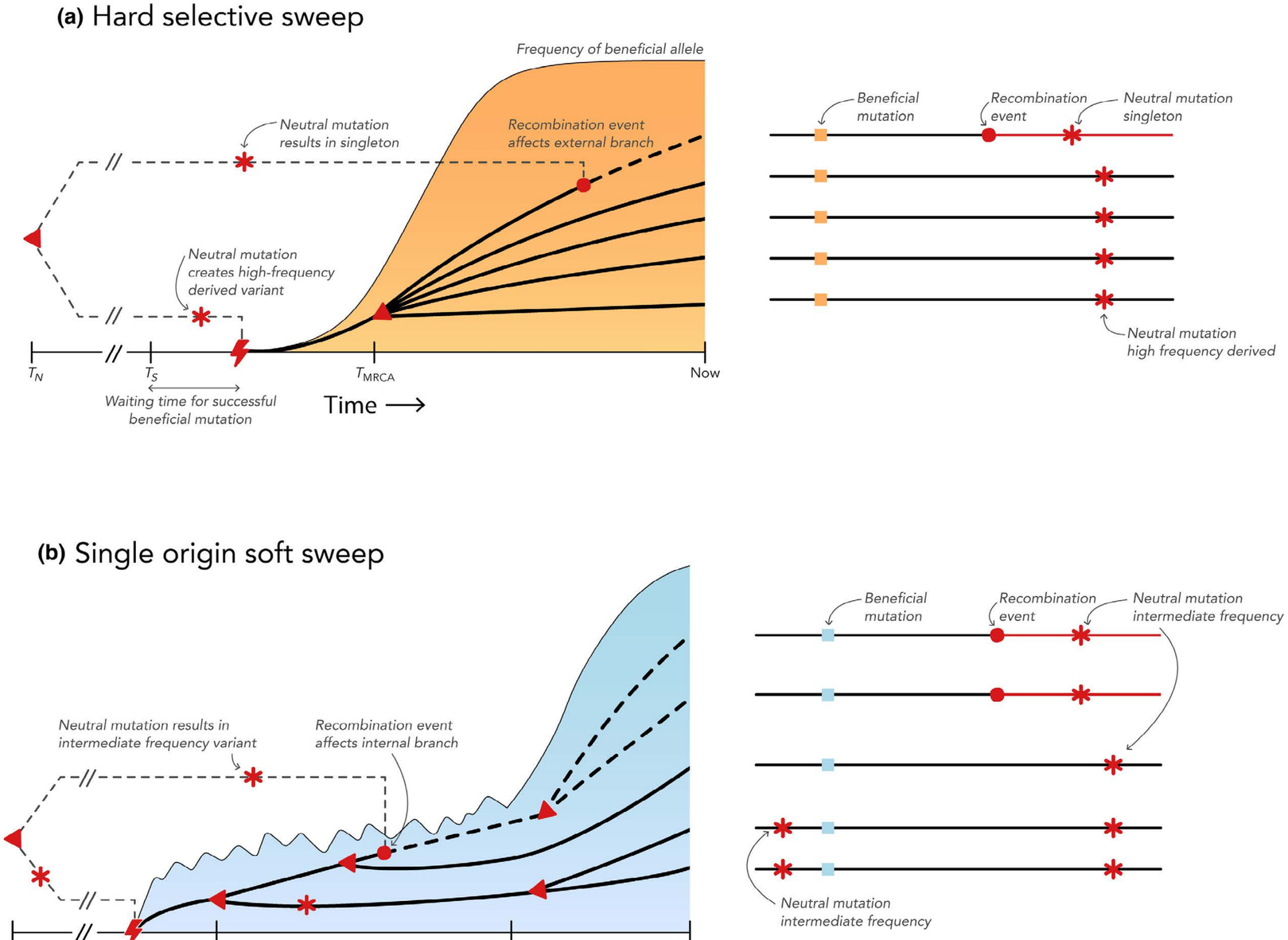
Beneficial mutation arise on different genetic backgrounds before any single background can sweep, the backgrounds carrying the beneficial mutation will spread concurrently.

More **genetic diversity** will be retained following the fixation of the beneficial mutation, because diverse genetic background linked with each beneficial mutations arose in frequency.

Hard and Soft sweeps



Messer & Petrov 2013



Hermisson & Pennings 2017