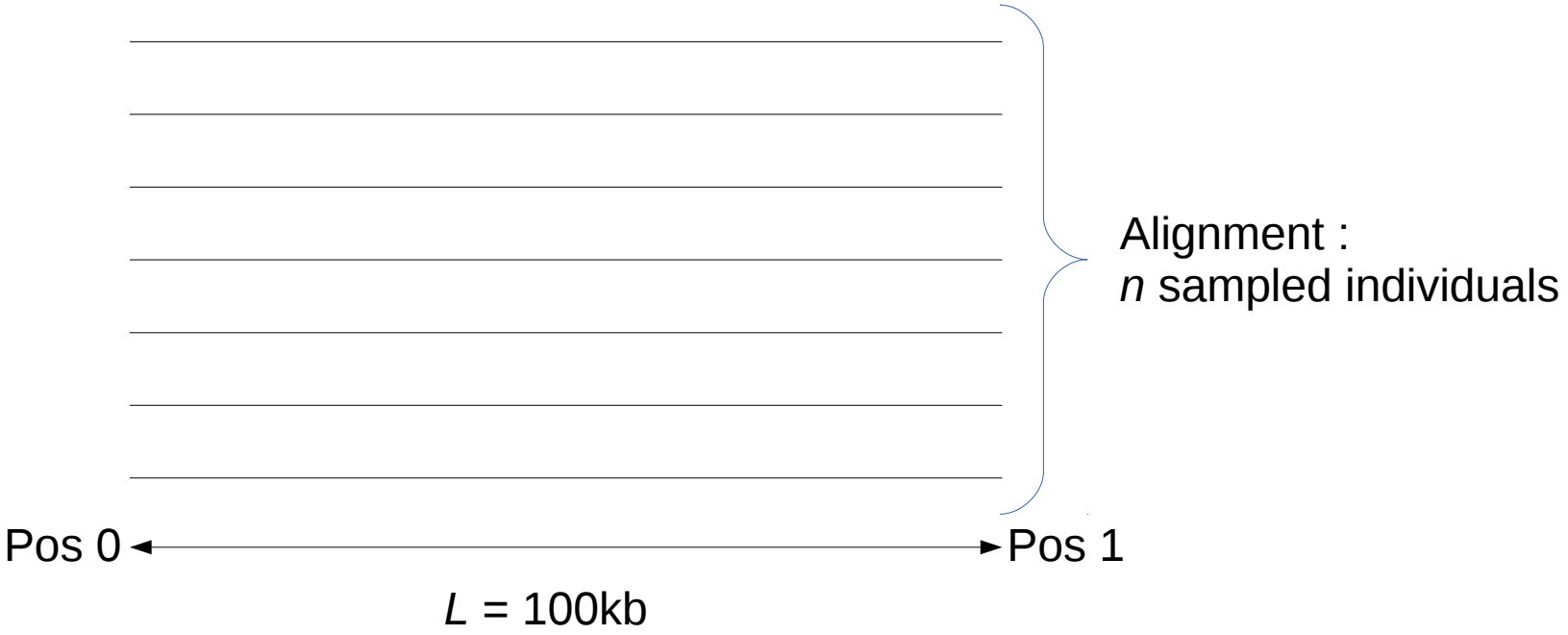
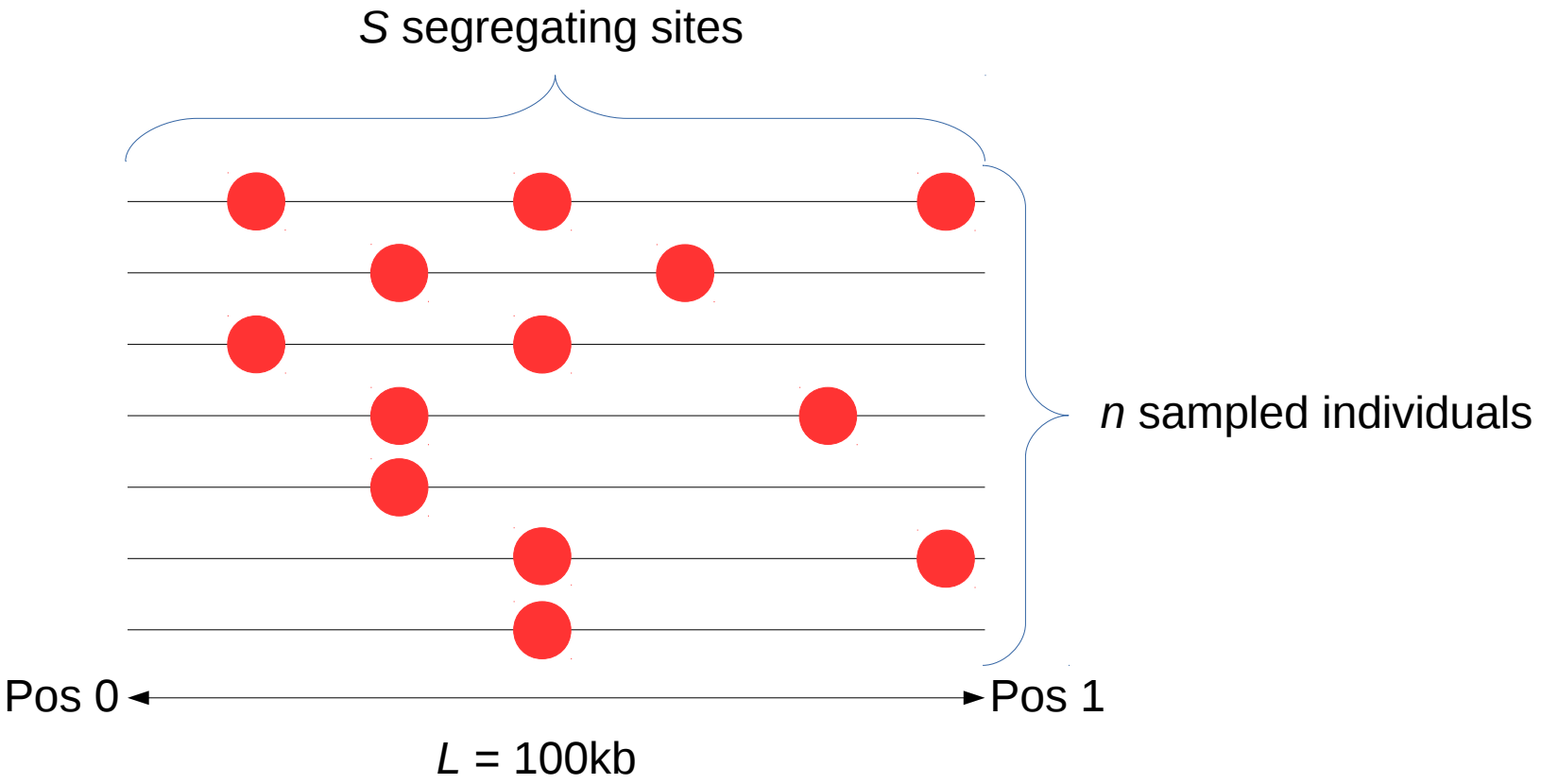
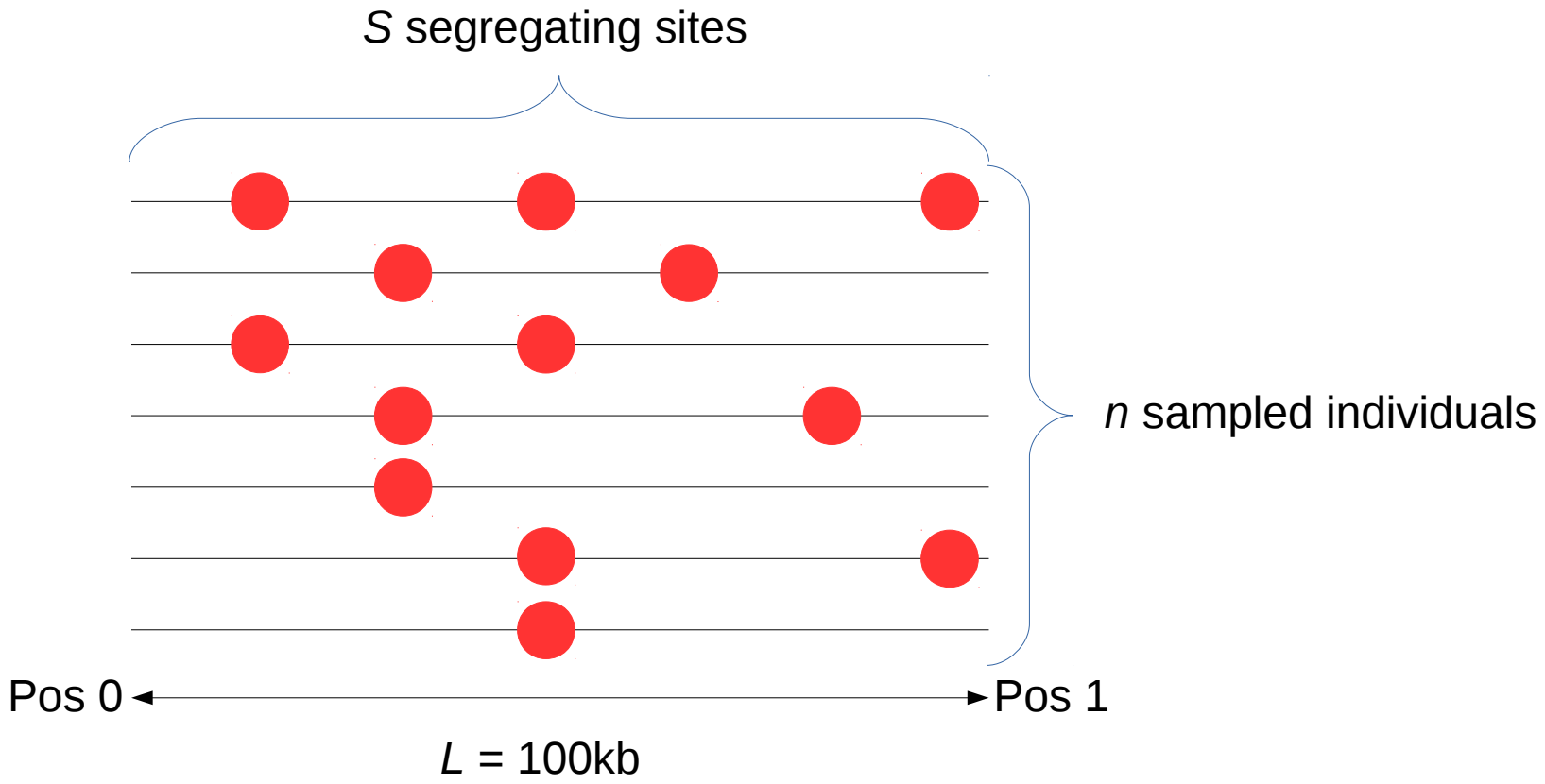
Can we detect a selective sweep within a genomic region using ABC ?



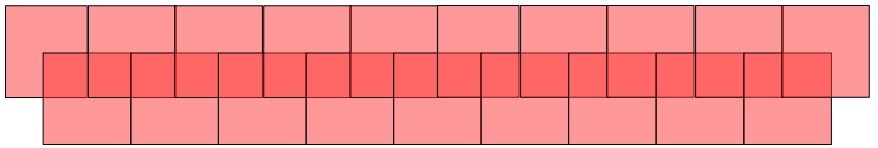
Can we detect a selective sweep within a genomic region using ABC ?



Can we detect a selective sweep within a genomic region using ABC ?



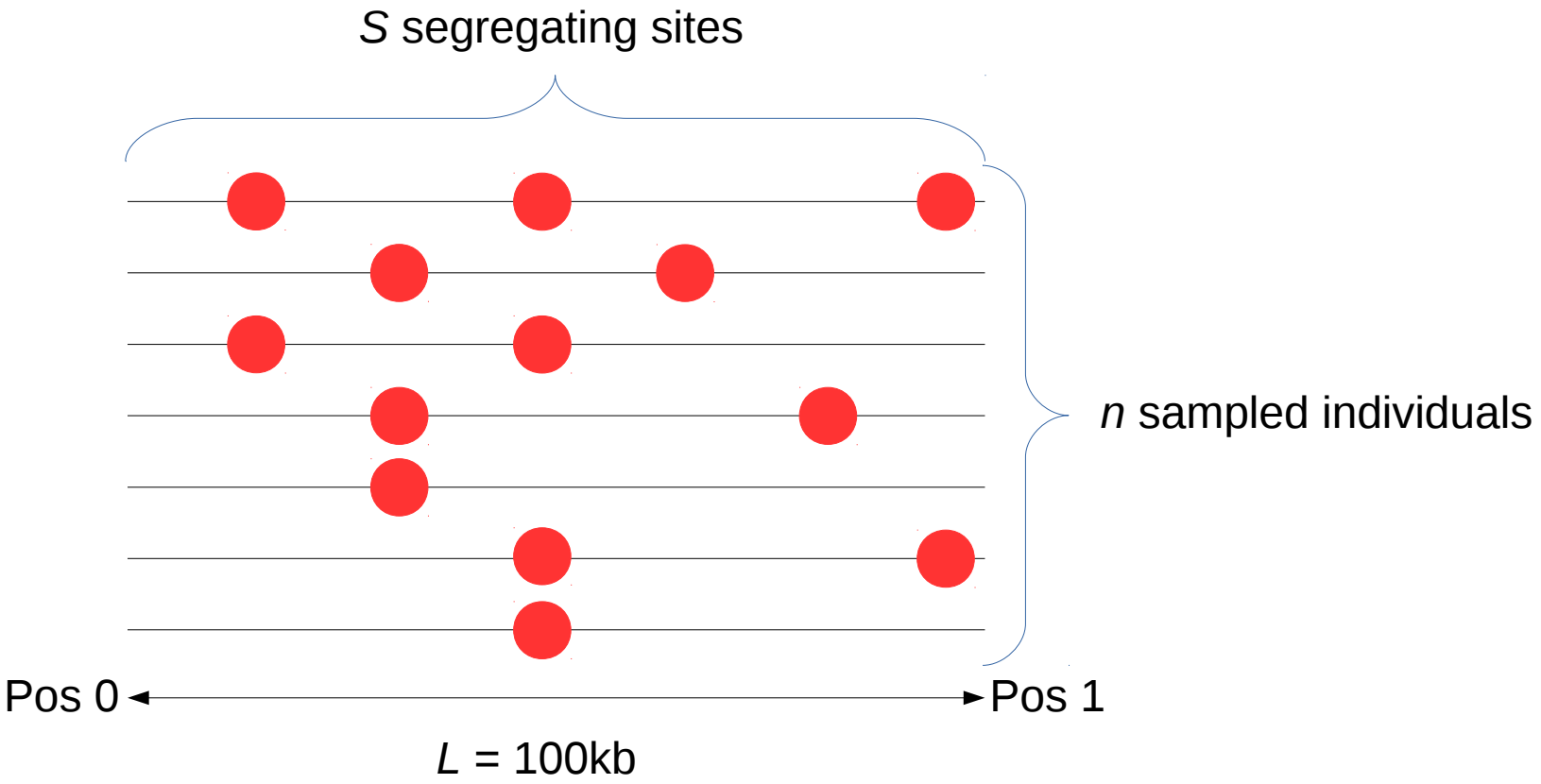
Scanning the surveyed region using a sliding window :



Overlapping windows :
Width = 0.1 (10kb)*
Step = 0.05 (5kb)

*values are not fixed
Can be setted by users.

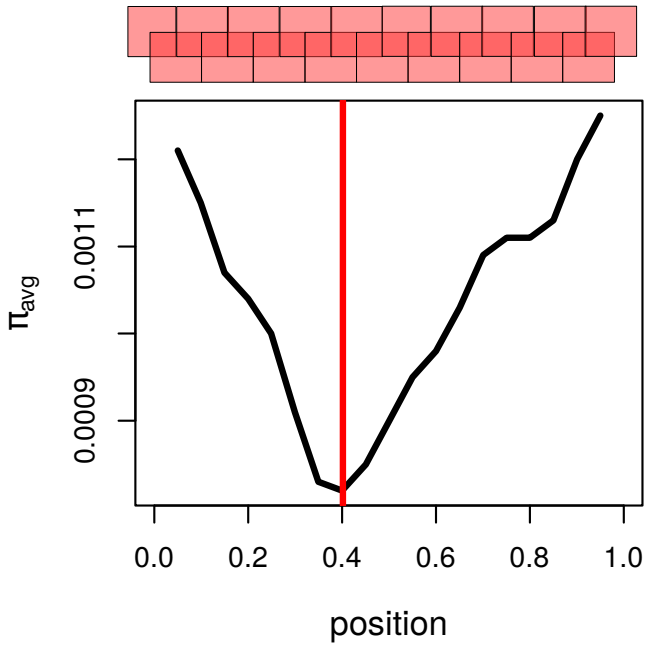
Can we detect a selective sweep within a genomic region using ABC ?

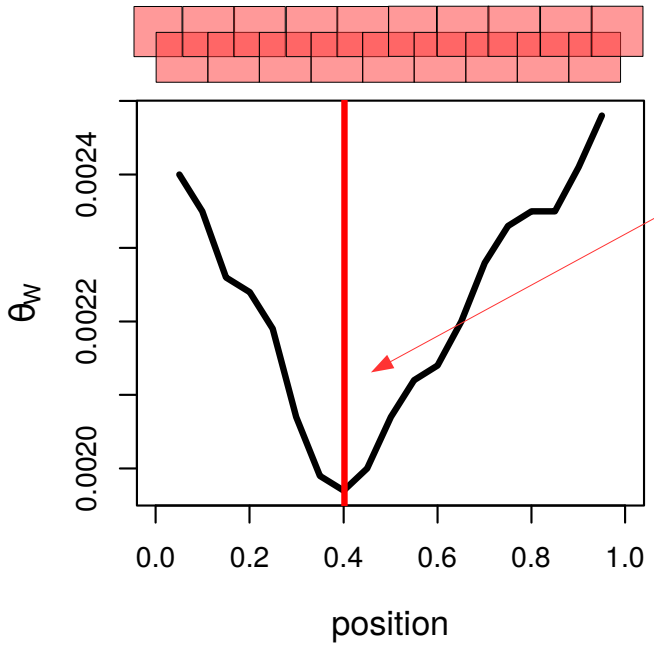


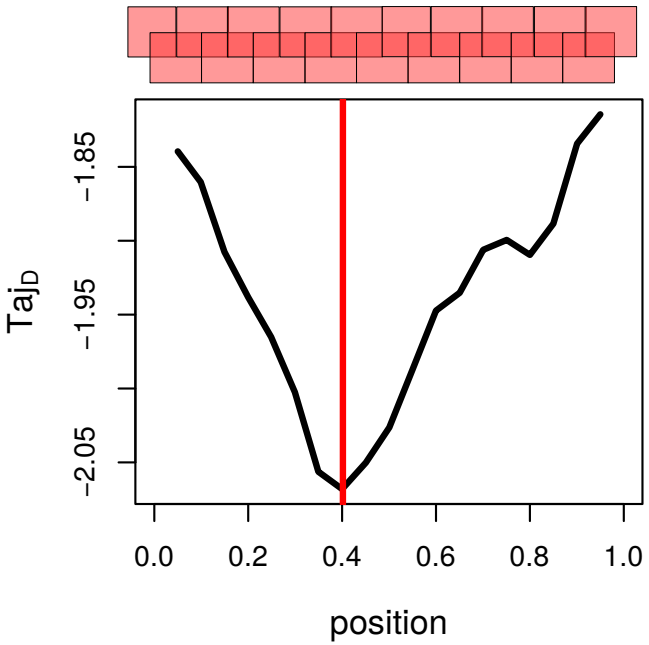
7 computed statistics in different windows :

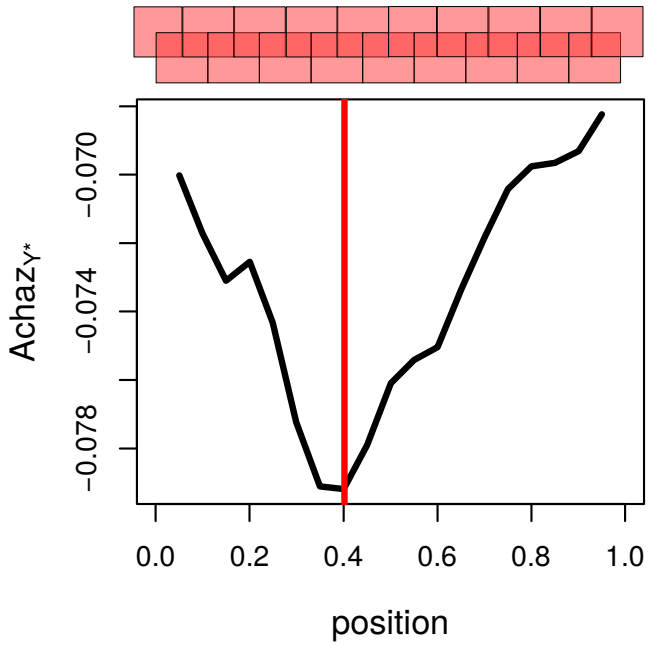
- Average π over SNPs
- Variation for π over SNPs
- Watterson's θ
- Tajima's *D*
- Achaz's *Y**
- Pearson's r (position ~ pi)
- P-value of Pearson's r

Simple visualisation of some statistics across a surveyed region of 100kb









Example :

.Position **Sp** = 0.4

.Coefficient S_{AA} = 3.5 %

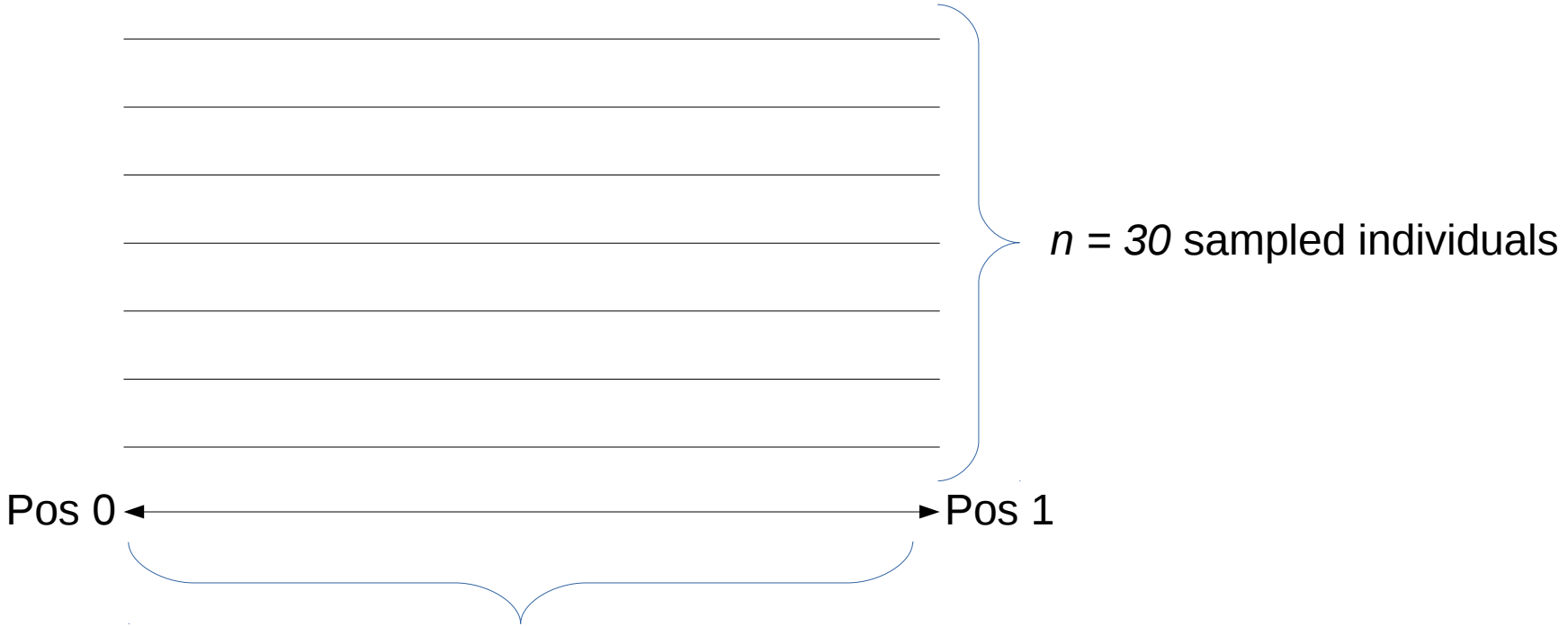
.Fixation time **SF** = 30,000 gen

.L = 100kb

.Recomb rate = $2e^{-8}$/bp

Can we use [ABC + random forests] to estimate Sp, SF and S_{AA} ?

Sp, **SF** and S_{AA} are estimated for 10,000 simulated datasets with random combinations of {**Sp ; SF;** S_{AA}}

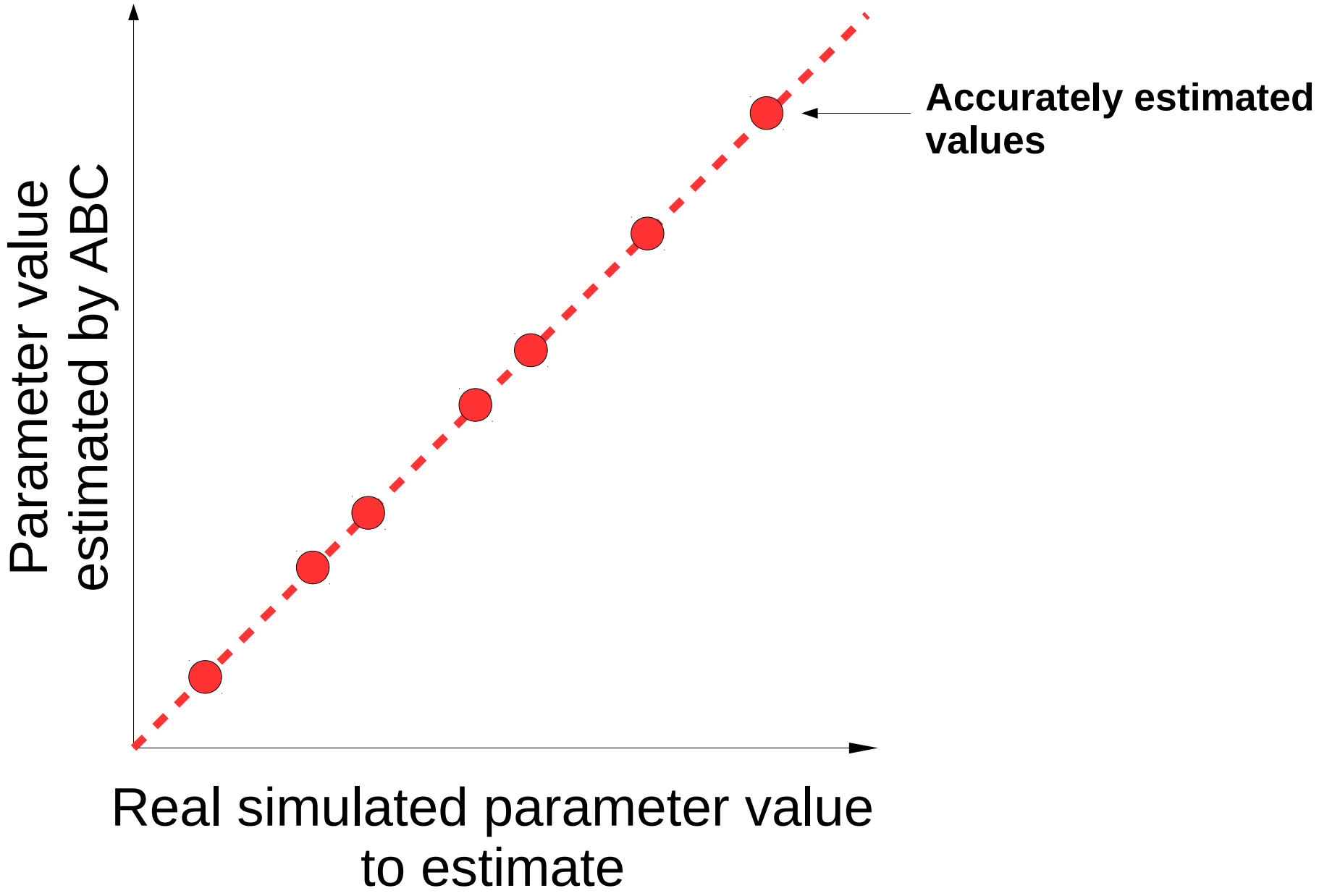


Sp : randomly sampled from uniform ***U*(0, 1)**

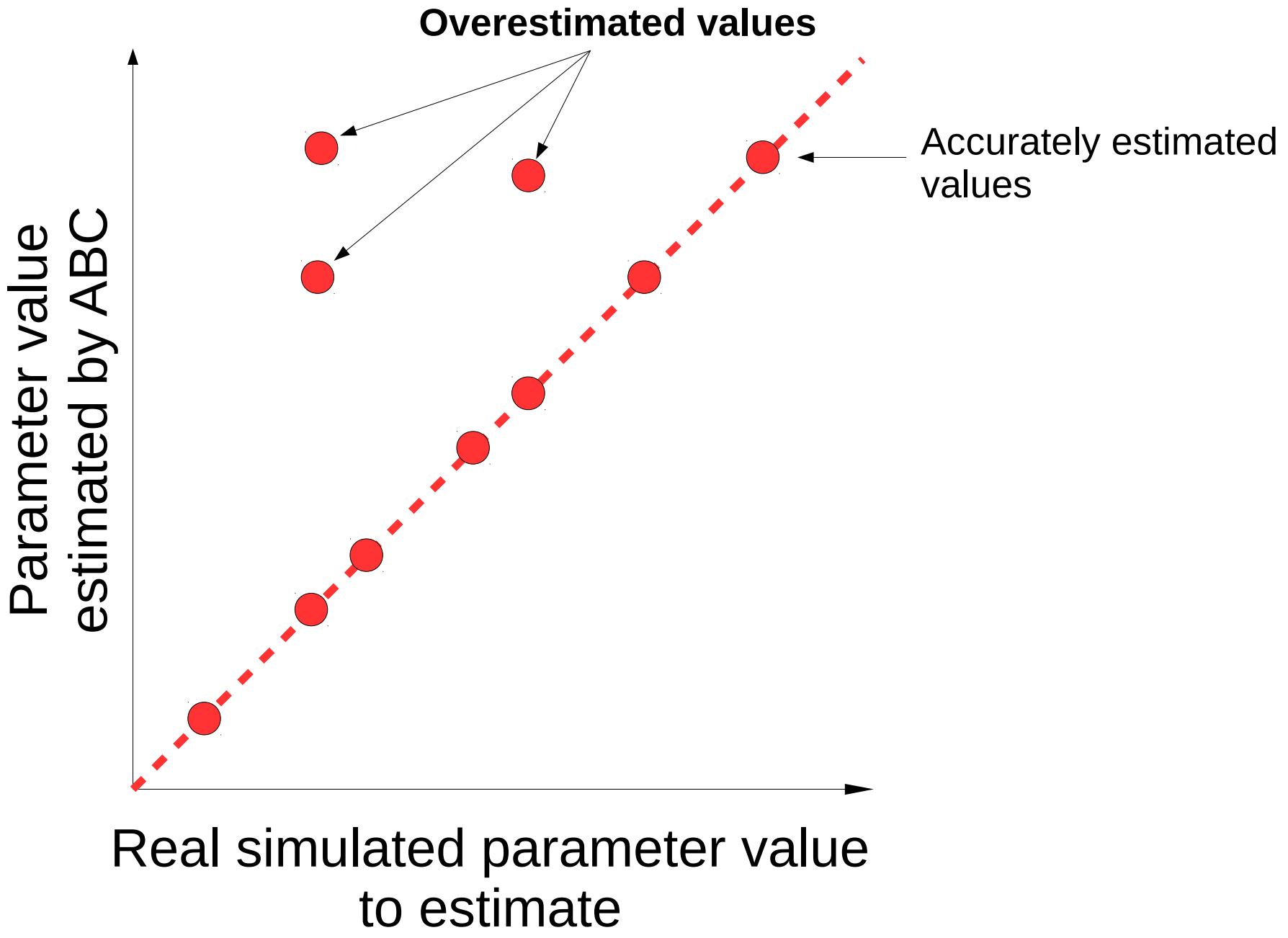
SF : randomly sampled from uniform ***U*(0, 400,000)** generations

S_{AA} : randomly sampled from uniform ***U*(0.002, 0.5)**

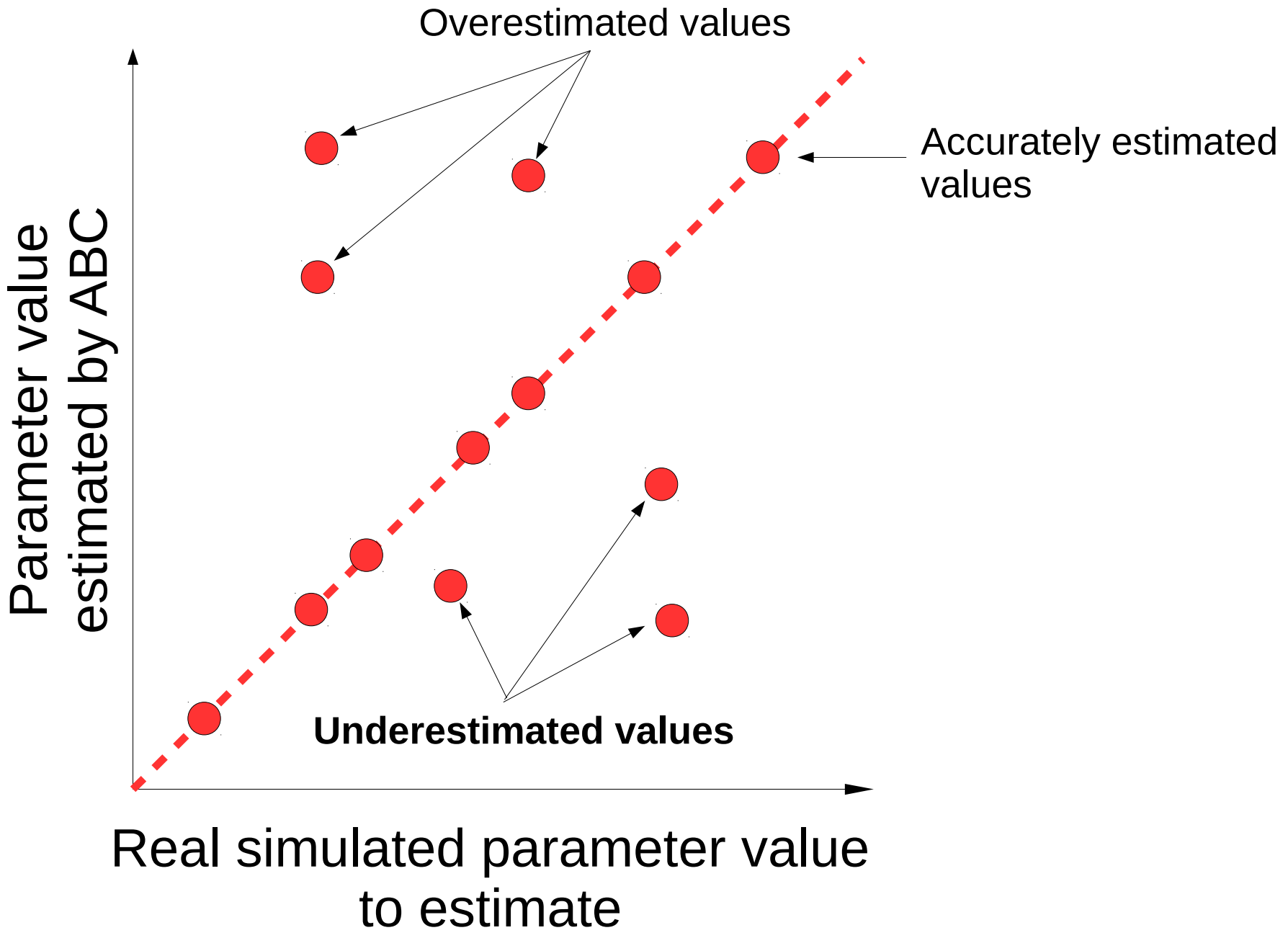
Schematic representation + interpretation



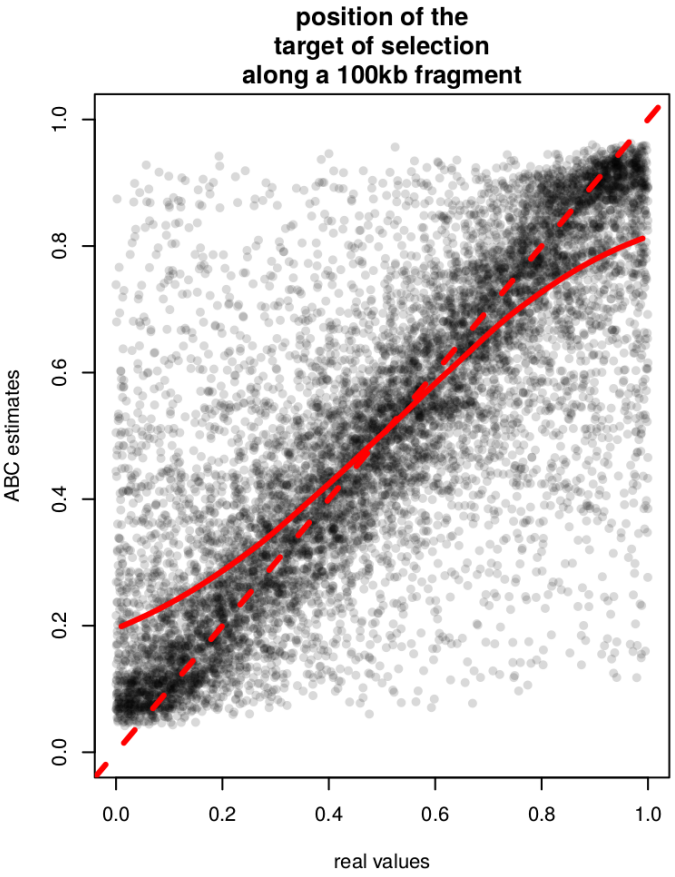
Schematic representation + interpretation

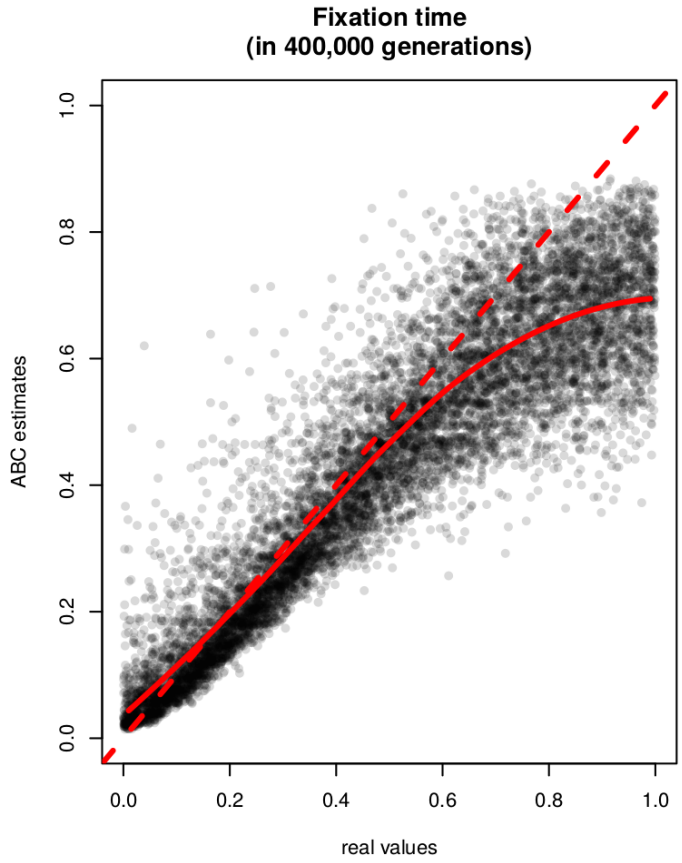


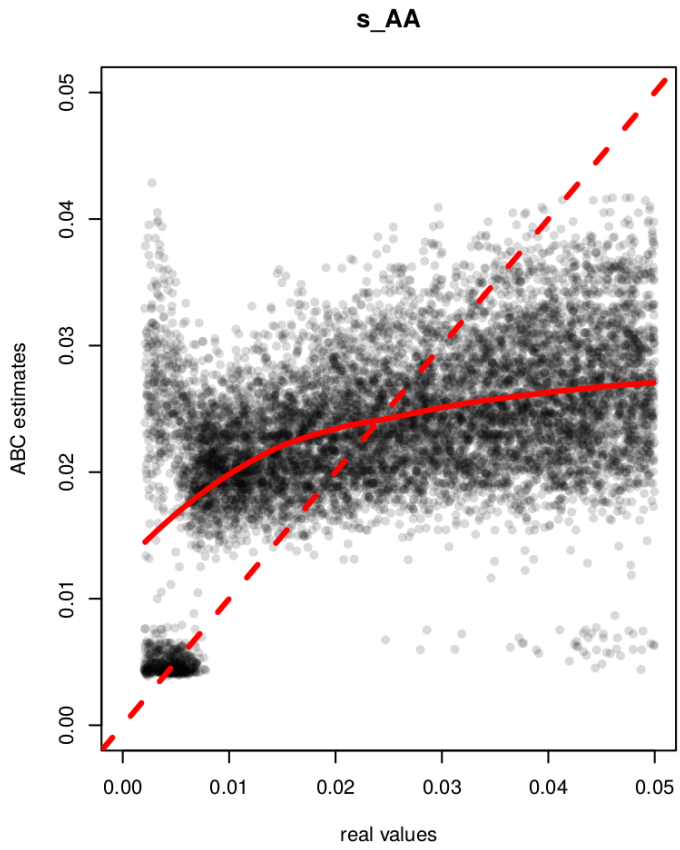
Schematic representation + interpretation



Results : ABC analyses of 10,000 of simulated datasets
1. Plot of 10,000 points (real values, estimated values)







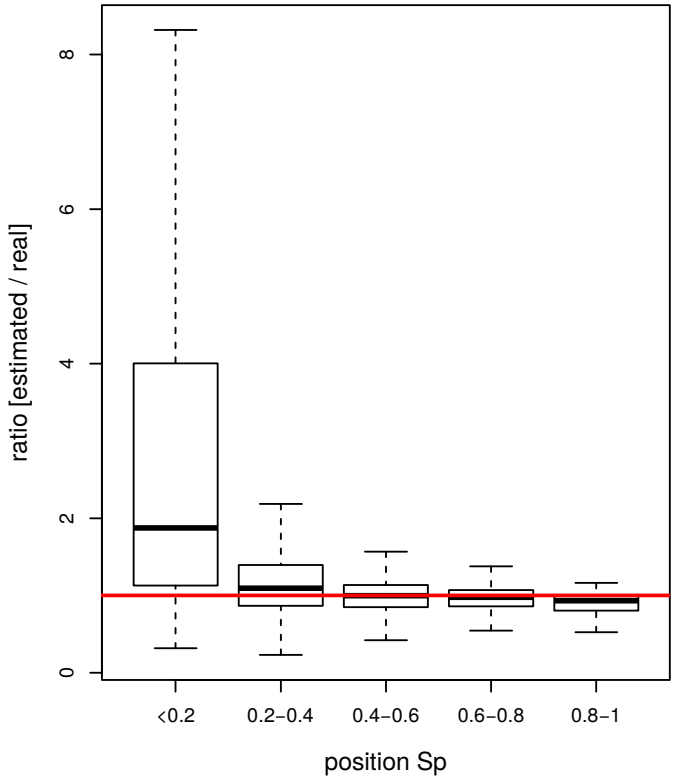
Visual interpretation :

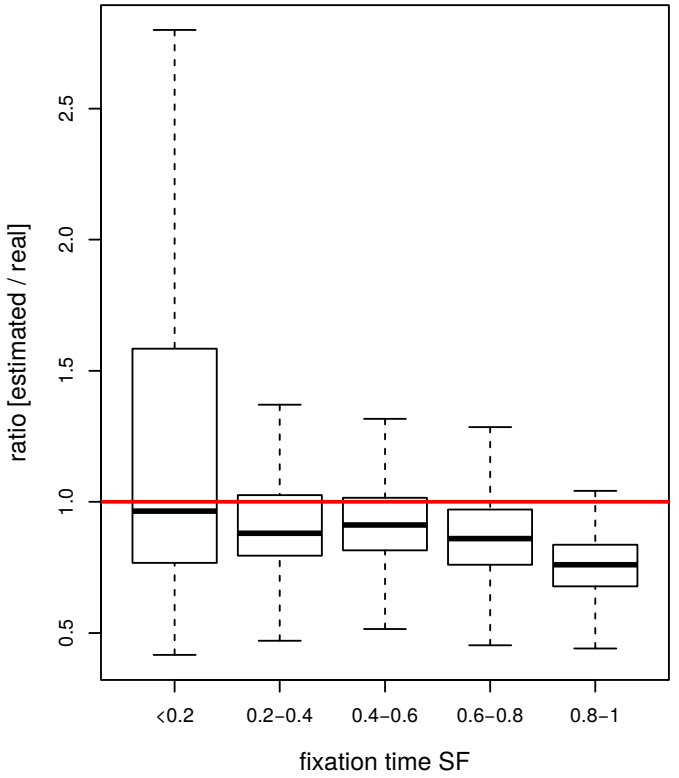
Estimating the position : not bad, except in the proximity of extremities (general problem with ABC). Not a big issue if genomes are scanned using a sliding window.

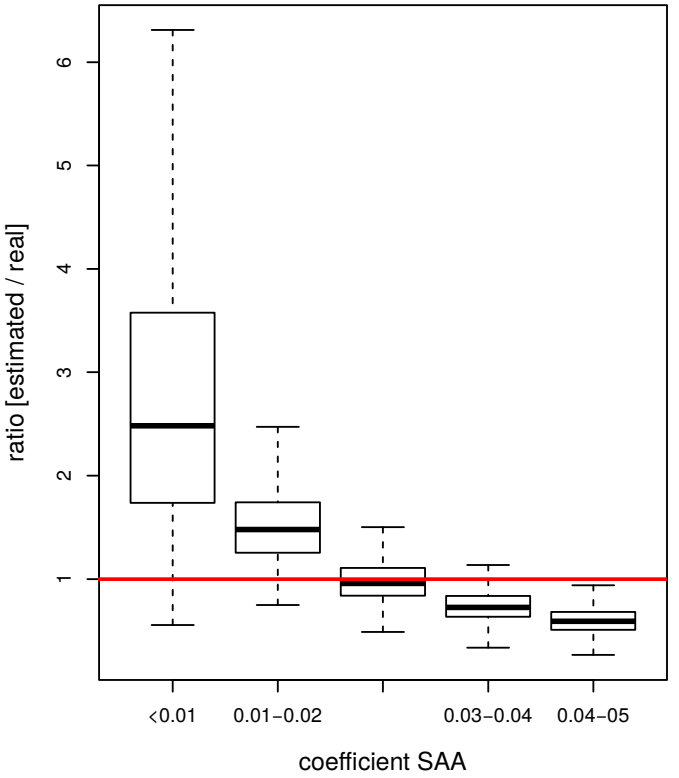
Fixation time : looks efficient, except for ancient fixation.

Selective coefficient : cannot be estimated for now. It's normal according to Pennings and Hermisson when using patterns of diversity.

Results : ABC analyses of 10,000 of simulated datasets
2. Distributions of the ratio estimates/real values within bins of parameters







Visual interpretation :

Estimating the position : not bad, except in the proximity of extremities (general problem with ABC). Not a big issue if genomes are scanned using a sliding window.

Fixation time : looks efficient, except for ancient fixation.

Selective coefficient : cannot be estimated for now. It's normal according to Pennings and Hermisson when using patterns of diversity.

Results of explicit model comparisons:

Since selective coefficients cannot be estimated using the current version, an explicit test of sweep can be an option :

Number of simulations: 9598 (4841 ‘neutral’ + 4757 ‘with sweep’)
(not exactly 10,000 after removing simulations with at least one ‘NA’ statistic)

Confusion matrix:

Simulated models		Classified as		
		Neutral	selection	class.error
	Neutral	4796	45	0.0092956001
	Selection	3	4754	0.0006306496

Mean error rate : 0.004963125

Very low error rate (~0.05%) when discriminating among «neutral» and «sweeped» datasets

Next step :

1) trying to add statistics taking informations from LD (i.e : H12 stat from http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004845)

2) exploring genomes :
For instance, it can be done by first :

i) comparing « neutral » versus « selective » scenarios in overlapping 100kb regions of genomes (divided themselves in windows like in slide 3)

ii) estimating the position Sp and the time SF for regions supporting the selective Scenarios

3) To do before : fitting a demographic scenario to avoid bias because of recent expansion/contraction of populations

4) all codes will be uploaded here : https://github.com/popgenomics/ABCsweep