MOLECULAR ECOLOGY
RESOURCES

**Supplemental Information for:**

# Sampling strategy optimization to increase statistical power in landscape genomics: a simulation-based approach

Oliver Selmoni, Elia Vajana, Annie Guillaume, Estelle Rochat, Stéphane Joost

MOLECULAR ECOLOGY
RESOURCES

## Supplementary Tab. 1. List of environmental variables employed.

| Name | Geographic resolution | Source |
|---|---|---|
| **Annual Mean Temperature** | 2.5 minutes | Bioclim[1](BIO1) |
| **Mean Diurnal Range** | 2.5 minutes | Bioclim[1] (BIO2) |
| **Temperature Seasonality** | 2.5 minutes | Bioclim[1](BIO4) |
| **Mean Temperature of Wettest Quarter** | 2.5 minutes | Bioclim[1](BIO8) |
| **Annual Precipitation** | 2.5 minutes | Bioclim[1](BIO12) |
| **Precipitation Seasonality** | 2.5 minutes | BIoclim[1](BIO15) |
| **Precipitation of Warmest Quarter** | 2.5 minutes | Bioclim[1](BIO18) |
| **Altitude** | 100 m | Marine Geoscience Data System[2] |

[1] WorldClim - Global Climate Data | Free climate data for ecological modeling and GIS. Available at: http://www.worldclim.org/. (Accessed: 26th September 2018)

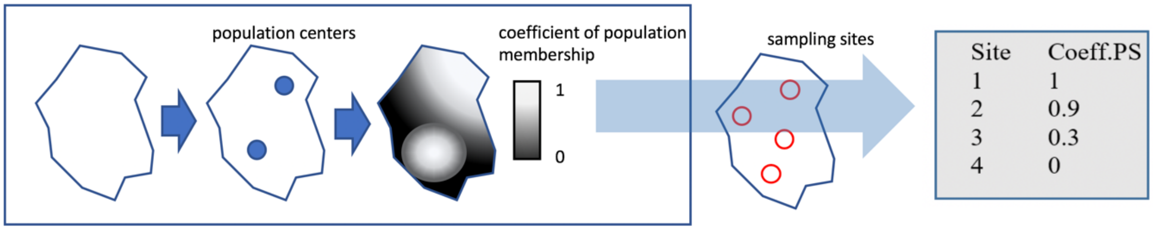
[2] MGDS. Global Multi-Resolution Topography Data Synthesis. Available at: http://www.marine-geo.org/portals/gmrt/. (Accessed: 22nd August 2017)

MOLECULAR ECOLOGY RESOURCES

**Supplementary Box 1. Computation of the genotype matrix.** The vignettes describe how genotypes were computed during simulations. At each iteration, a new genotype matrix counting 1'000 loci was generated. Ten of them were set as adaptive and followed the respective pipeline, while the others were set as neutral and computed accordingly.

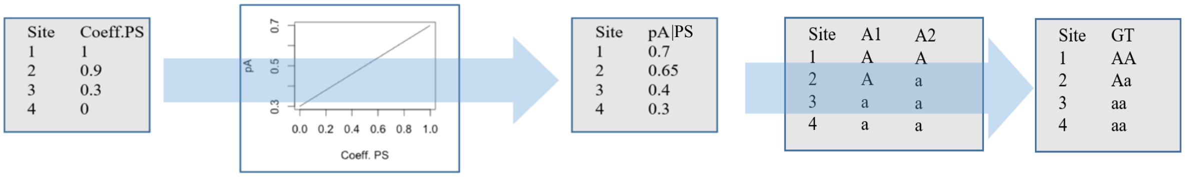
## A) Neutral Locus

i. An artificial population membership coefficient is computed as the distance from randomly located population centers. The membership coefficient is extracted then at each sampling site.



| Site | Coeff.PS |
|---|---|
| 1 | 1 |
| 2 | 0.9 |
| 3 | 0.3 |
| 4 | 0 |

ii. A function translates the coefficient of population structure in the probability of carrying the allele characteristic of the population. Finally, alleles are sampled at each site using the probability associated. This step is reiterated if more than one individual is sampled at the same site and for all the loci related to a same population membership coefficient.



| Site | Coeff.PS |
|---|---|
| 1 | 1 |
| 2 | 0.9 |
| 3 | 0.3 |
| 4 | 0 |

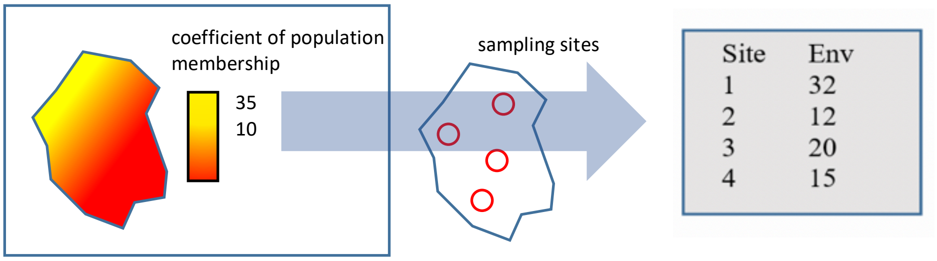
| Site | pA\|PS |
|---|---|
| 1 | 0.7 |
| 2 | 0.65 |
| 3 | 0.4 |
| 4 | 0.3 |

| Site | A1 | A2 |
|---|---|---|
| 1 | A | A |
| 2 | A | a |
| 3 | a | a |
| 4 | a | a |

| Site | GT |
|---|---|
| 1 | AA |
| 2 | Aa |
| 3 | aa |
| 4 | aa |

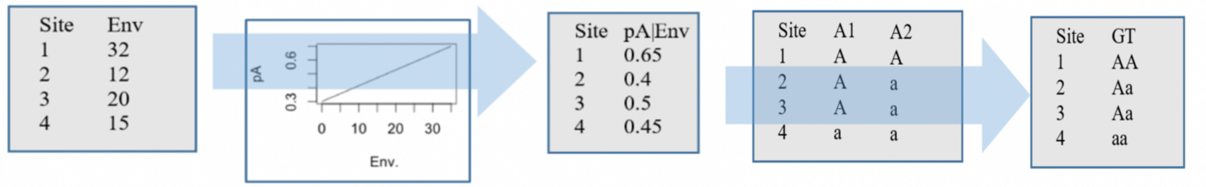
## B) Adaptive Locus

i. For each sampling site, the environmental values are extracted.



| Site | Env |
|---|---|
| 1 | 32 |
| 2 | 12 |
| 3 | 20 |
| 4 | 15 |

ii. A function computes the probability of carrying an allele conferring a selective advantage against the environmental condition. Alleles are sampled at each site using the probability associated. This step is reiterated if more than one individual are sampled at the same site.



| Site | Env |
|---|---|
| 1 | 32 |
| 2 | 12 |
| 3 | 20 |
| 4 | 15 |

| Site | pA\|Env |
|---|---|
| 1 | 0.65 |
| 2 | 0.4 |
| 3 | 0.5 |
| 4 | 0.45 |

| Site | A1 | A2 |
|---|---|---|
| 1 | A | A |
| 2 | A | a |
| 3 | A | a |
| 4 | a | a |

| Site | GT |
|---|---|
| 1 | AA |
| 2 | Aa |
| 3 | Aa |
| 4 | aa |

MOLECULAR ECOLOGY RESOURCES

**Supplementary Box 2. Formulae and parameters for genotype computations**

The probability function for the allele A depending on a population membership coefficient is calculated as follows:

$$p(A|PS) = \left(\frac{1-2c}{\max(PS) - \min\,(PS)}\right) PS + c - \left(\frac{1-2c}{\max(PS) - \min(PS)}\right) \min\,(PS)$$

where *PS* is a population membership coefficient and *c* a parameter representing the strength of the relationship. This parameter can range between 0 (strongest relation, *i.e.* maximal and minimal *PS* returns *p*=1 and *p*=0, respectively) and 0.5 (no relation, any level of *PS* returns *p*=0.5).

Similarly, probability for the allele A depending on environmental selection is calculated as follows:

$$p(A|Env) = \left(\frac{1-2s_1}{\max(Env) - \min(Env)}\right) E + s_1 - \left(\frac{1-2s_1}{\max(Env) - \min(Env)}\right) \min(Env) + \ s_2$$

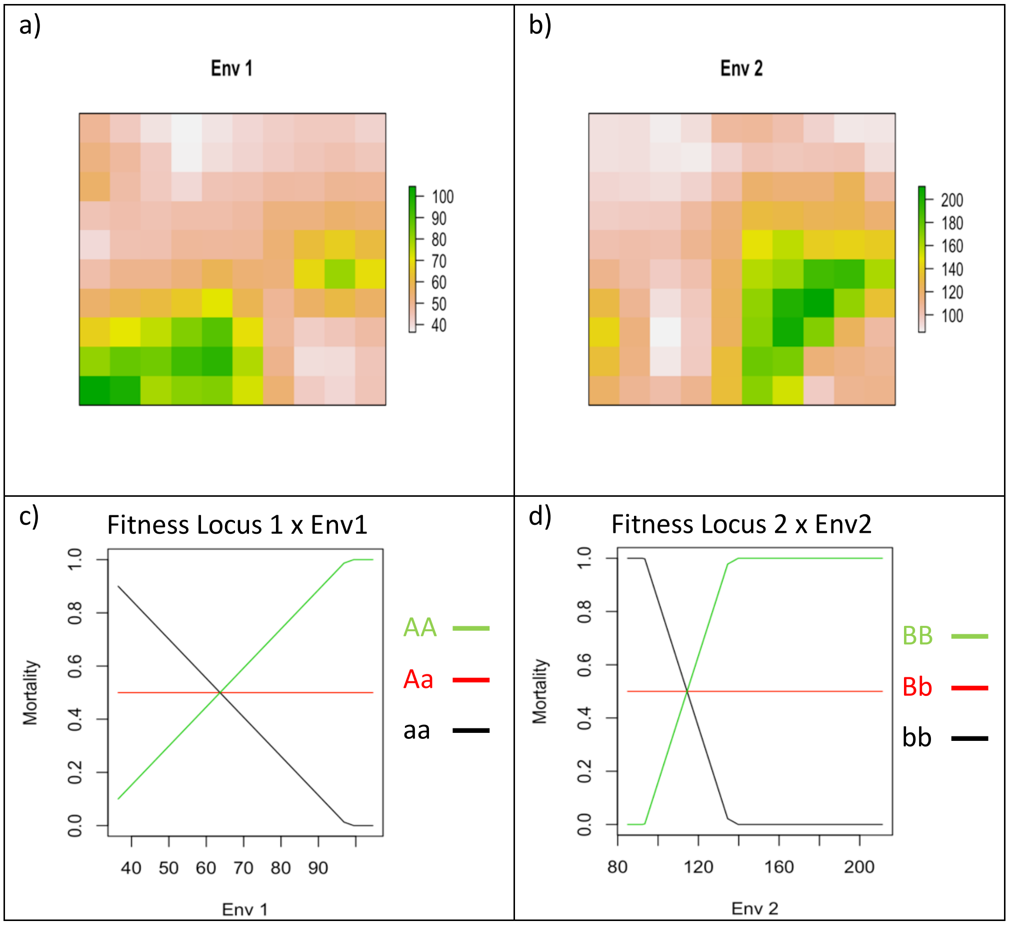
where *Env* are the values of the environmental variable and $s_1$ represents the strength of selection and behaves as the *c* in the previous equation. The additional parameter $s_2$ provides a baseline of allele frequency.

In our simulations, we set two scenarios employing the following parameters:

- *panmictic population scenario* (random neutral structure): *c=0.5*, $s_1$*=Unif(0.3, 0.4)*, $s_2$*=Unif(-0.2,0.1)*
- *structured population scenario* (strong population structure): *c=Unif(0.2,0.4)*, $s_1$*=0*, $s_2$*=Unif(-0.1,0.2)*

MOLECULAR ECOLOGY
RESOURCES

**Supplementary Figure 1. Environmental gradients and fitness constraint employed in the CDPOP validation run.** Panel a) and b) show the distribution of the two environmental variables across the 10-by-10 cells grid used for the CDPOP simulation. Plots in panels c) and d) show the fitness constraint set for the two environmental variables and how the respective adaptive genotypes modulate mortality.



MOLECULAR ECOLOGY
RESOURCES

**Supplementary Table 2. CDPOP vs. our simulative approach comparison metrics.** The tables show the rank of the simulative variants computed with our method (and defined by parameters *m, c* $s_1$ *and* $_{s2}$) that best matched the CDPOP replicates. In a) and b) are shown the metrics used to compare the neutral genetic structure with the CDPOP case of a panmictic population and a structured population, respectively. The three metrics employed are 1) the average random mean squared error (RMSE) when comparing the curves describing the differential of explained variation by the genetic principal components; 2) the Kullback-Leibler Divergence (KLD) used to compare the pairwise-Fst distributions; 3) the difference in the average mantel correlation (ΔmR), which describes the link between genetic and geographic distances. The ranking coefficient is the sum of the three scaled metrics. In c) and d) the comparison concerns the adaptive genotypes computed in panmictic structured scenario of CDPOP, respectively. Here the RMSE compares, for our simulation and CDPOP runs, the allelic frequency of the adaptive genotype as a function of the environmental variable causing adaptation

## a) Panmictic Scenario: Neutral structure metrics

| *rank* | *m* | *c* | RMSE (PCA) | KLD (Fst) | ΔmR | Ranking Coefficient |
|---|---|---|---|---|---|---|
| 1 | 1 | 0.5 | 0.000780575 | 7.33E-06 | 0.003577 | -4.35661 |
| 2 | 25 | 0.4-0.5 | 0.000771722 | 7.70E-06 | 0.022455 | -4.25828 |
| 3 | 10 | 0.4-0.5 | 0.000771901 | 7.93E-06 | 0.023357 | -4.24377 |
| 4 | 20 | 0.4-0.5 | 0.000780659 | 8.58E-06 | 0.022308 | -4.21677 |
| 5 | 5 | 0.4-0.5 | 0.000770043 | 7.46E-06 | 0.034877 | -4.21321 |
| 6 | 15 | 0.4-0.5 | 0.000766353 | 9.31E-06 | 0.025071 | -4.17643 |
| 7 | 5 | 0.4-0.4 | 0.000796873 | 1.15E-05 | 0.067273 | -3.88113 |
| 8 | 10 | 0.4-0.4 | 0.000763216 | 1.12E-05 | 0.074199 | -3.87217 |
| 9 | 25 | 0.4-0.4 | 0.000771422 | 1.27E-05 | 0.072328 | -3.81237 |
| 10 | 20 | 0.4-0.4 | 0.000761967 | 1.38E-05 | 0.073625 | -3.7593 |

## c) Panmictic Scenario: adaptive genotypes metrics

| Moderate Selection | | | |
|---|---|---|---|
| rank | $s_1$ | $s_2$ | RMSE (AF) |
| 1 | 0 | -0.1 | 0.7417767 |
| 2 | 0.1 | -0.1 | 0.75108 |
| 3 | 0.1 | -0.2 | 0.7681983 |
| 4 | 0 | -0.2 | 0.78917 |
| 5 | 0.2 | -0.1 | 0.7946361 |
| Strong Selection | | | |
| rank | $s_1$ | $s_2$ | RMSE (AF) |
| 1 | 0 | 0.2 | 0.676855 |
| 2 | 0.1 | 0.2 | 0.683247 |
| 3 | 0.1 | 0 | 0.710474 |
| 4 | 0 | 0.1 | 0.715619 |
| 5 | 0.2 | 0.1 | 0.728321 |

## b) Structured Scenario: Neutral structure metrics

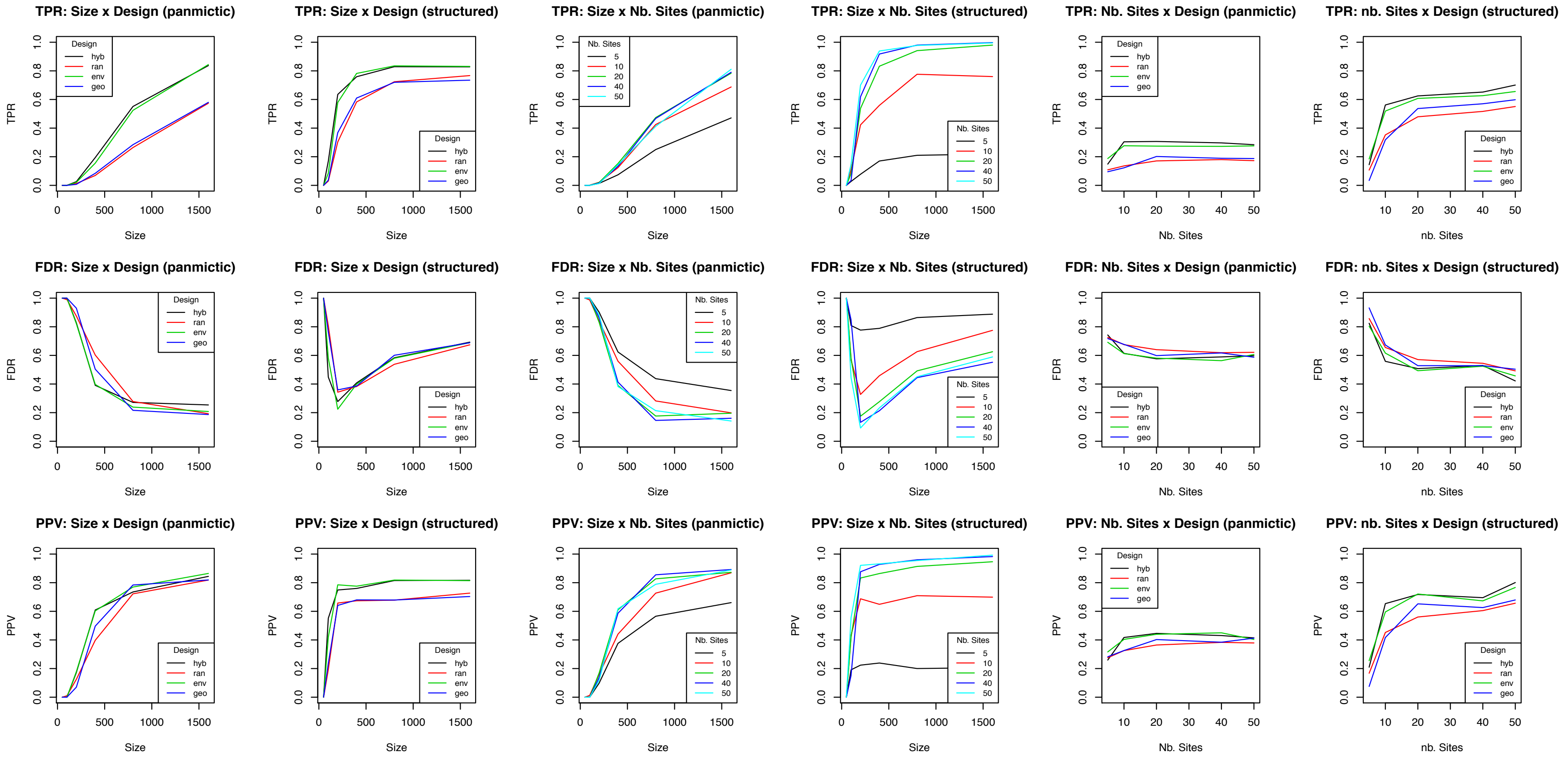
| *rank* | *m* | *c* | RMSE (PCA) | KLD (Fst) | ΔmR | Ranking Coefficient |
|---|---|---|---|---|---|---|
| 1 | 10 | 0.2-0.4 | 0.00290909 | 8.17E-06 | 0.320549 | -3.63827 |
| 2 | 20 | 0.1-0.5 | 0.00266099 | 8.85E-06 | 0.339198 | -3.63027 |
| 3 | 5 | 0.3 | 0.003023145 | 8.38E-06 | 0.312132 | -3.45645 |
| 4 | 15 | 0.1-0.5 | 0.002793301 | 7.57E-06 | 0.37057 | -3.43066 |
| 5 | 25 | 0.2-0.4 | 0.003250162 | 8.42E-06 | 0.314625 | -3.31517 |
| 6 | 15 | 0.2-0.3 | 0.002468453 | 6.72E-06 | 0.422087 | -3.31507 |
| 7 | 5 | 0.2-0.4 | 0.003092629 | 9.91E-06 | 0.329403 | -3.27752 |
| 8 | 10 | 0.3 | 0.002819477 | 9.84E-06 | 0.295631 | -3.26125 |
| 9 | 25 | 0.1-0.5 | 0.002947686 | 8.05E-06 | 0.373038 | -3.23848 |
| 10 | 15 | 0.2-0.5 | 0.002799946 | 1.02E-05 | 0.280361 | -3.09366 |

## d) Structured Scenario: adaptive genotypes metrics

| Moderate Selection | | | |
|---|---|---|---|
| rank | $s_1$ | $s_2$ | RMSE (AF) |
| 1 | 0.4 | -0.2 | 0.6889893 |
| 2 | 0.3 | -0.2 | 0.6895106 |
| 3 | 0.2 | -0.2 | 0.7181186 |
| 4 | 0.3 | -0.1 | 0.7319583 |
| 5 | 0.2 | -0.1 | 0.7454251 |
| Strong Selection | | | |
| rank | $s_1$ | $s_2$ | RMSE (AF) |
| 1 | 0.3 | 0.1 | 0.624262 |
| 2 | 0.4 | 0.1 | 0.6417665 |
| 3 | 0.2 | 0.1 | 0.6484901 |
| 4 | 0.3 | 0 | 0.6709922 |
| 5 | 0.4 | 0 | 0.6831192 |

MOLECULAR ECOLOGY
RESOURCES

**Supplementary Figure 2. Interaction effects.** The table displays different combination of the elements defining the sampling strategy and their effect on the average of the three diagnostic parameters (row 1: TPR, row 2: FDR, row 3: PPV). For every diagnostic parameter, the two demographic scenarios are represented (column 1,3 and 5: panmictic, 2,4 and 6: structured). In columns 1-2, the combined effects of sample size (x axis) and sampling design (colored lines) are observed. In columns 3-4, the combined effects of sample size (x axis) and number of sampling locations (colored lines) are observed. In columns 5-6, the combined effects of number of sampling locations (x axis) and sampling design (colored lines) are observed.


TPR: Size x Design (panmictic)
TPR: Size x Design (structured)
TPR: Size x Nb. Sites (panmictic)
TPR: Size x Nb. Sites (structured)
TPR: Nb. Sites x Design (panmictic)
TPR: nb. Sites x Design (structured)
FDR: Size x Design (panmictic)
FDR: Size x Design (structured)
FDR: Size x Nb. Sites (panmictic)
FDR: Size x Nb. Sites (structured)
FDR: Nb. Sites x Design (panmictic)
FDR: nb. Sites x Design (structured)
PPV: Size x Design (panmictic)
PPV: Size x Design (structured)
PPV: Size x Nb. Sites (panmictic)
PPV: Size x Nb. Sites (structured)
PPV: Nb. Sites x Design (panmictic)
PPV: nb. Sites x Design (structured)
TPR
FDR
PPV
Size
Nb. Sites
nb. Sites
Design
hyb
ran
env
geo
Nb. Sites