Computing Genomic Offset with the Gradient Forest R package

Olivier François

Université Grenoble-Alpes

SSMPG 2025, Aussois, Savoie

Caution

- The GradientForest does not seem to be maintained. The package may not be available for the most recent version of R (> 4.5) in CRAN.

- Using GF is still possible, but requires installing a previous version of R (R 4.4.3 works).

https://cran.r-project.org/bin/macosx/big-sur-arm64/base/R-4.4.2-arm64.pkg

- We hope that this is a temporary issue, and that GradientForest will be back very soon

Gradient Forest (GF)

- GF is a machine learning method based on Random Forests, designed to study nonlinear relationships in environmental and landscape genomic or ecological data.

- In Gradient Forest, the importance curve (IC) is a way to show where along an environmental gradient a variable has the most influence on the response (e.g., allelic frequencies, species composition)

Fitzpatrick, M. C., Keller, S. R. (2015). Ecological genomics meets community-level modelling of biodiversity: Mapping the genomic landscape of current and future environmental adaptation. Ecology Letters, 18(1), 1-16.

Ecological predictors, $\mathbf{x} = (x_1, \ldots, x_d)$

- **Bioclimatic predictors**: temperature, precipitation or solar radiation
- **Biotic predictors**: local abundance of species sharing ecological interactions with the studied organisms
- **Altered environment**, $x^\star$, could result from **translocation** in geographic space or represent **future conditions** at unchanged geographic location
- Predictors are **unitless**, and expressed as deviations from the sample mean (e.g., centered)

GO principles (2 stages)

1. Adjust a genotype–environment association (GEA) model to predict allele frequencies from environmental variables across many genetic loci, and retain only those loci that exceed the significance threshold.

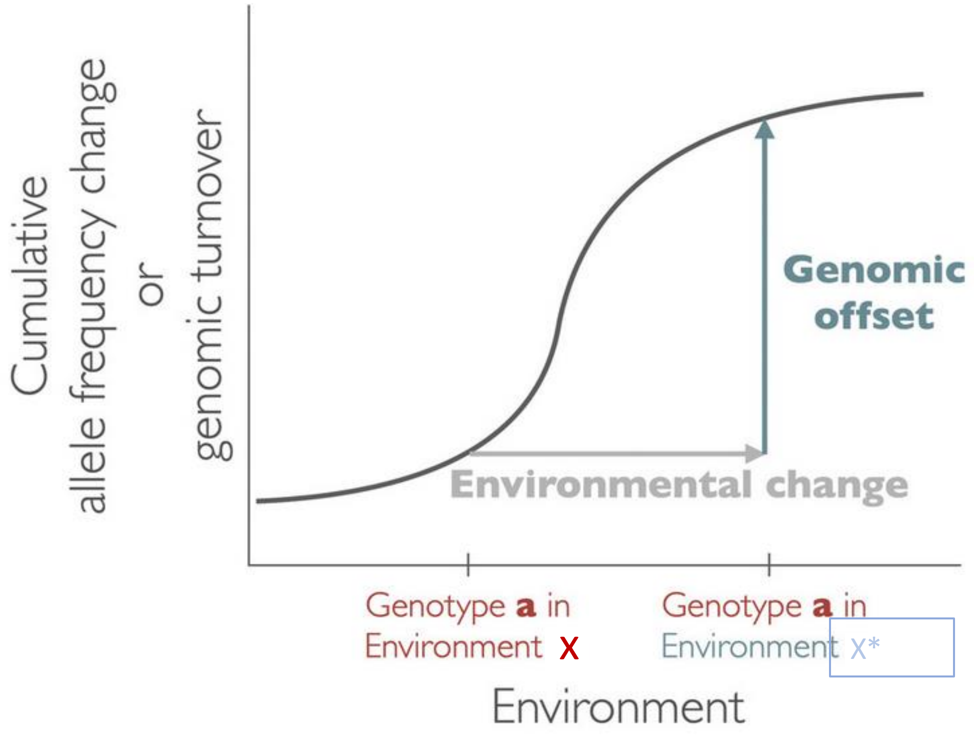
2. Evaluate dissimilarity between allelic frequencies predicted under current conditions and altered conditions.

Capblancq et al Genomic prediction of (mal)adaptation across current and future climatic landscapes 2020

GF Genomic offset

Genomic offset in GF is defined as

$$\mathrm{GF}^2(\mathbf{x}, \mathbf{x}^\star) = \sum_{j=1}^{d} .(\mathrm{IC}_j(x_j) - \mathrm{IC}_j(x_j^\star))^2$$



Lotterhos, K. E. (2024). Principles in experimental design for evaluating genomic forecasts. *Methods in Ecology and Evolution*, *15*(9), 1466-1482.

Starting a session

1. Start an R session with the gradientForest package

```
library(gradientForest)

# We use LEA in order to have access to the example dataset
library(LEA)
```

2. Load the example dataset

```
# loading the simulated data in R
data("offset_example")

# Y containes genotypes for 200 individuals
Y <- offset_example$geno

# X containes 4 environmental variables for 200 individuals
X <- offset_example$env
```

GEA study

1. Fit an LFMM with K = 3 latent factors

```r
# use lfmm2 to estimate the latent factors from the data
mod_lfmm2 <- lfmm2(input = Y,
                   env = X,
                   K = 3)
```

2. Compute *p*-values (one for each locus)

```r
# GEA analysis computing locus-specific p-values, P
pv_lfmm2 = lfmm2.test(object = mod_lfmm2,
                      input = Y,
                      env = X,
                      full = TRUE)$pvalues
```

Candidate loci

Decide which loci to include in GO

```
# FDR control: computing qvalues
qv_lfmm2  <- qvalue::qvalue(pv_lfmm2, fdr.level = 0.2)

# the most interesting targets
candidates <- which(qv_lfmm2$significant)
```

GF GO

1. Load future conditions (altered environmental variables)

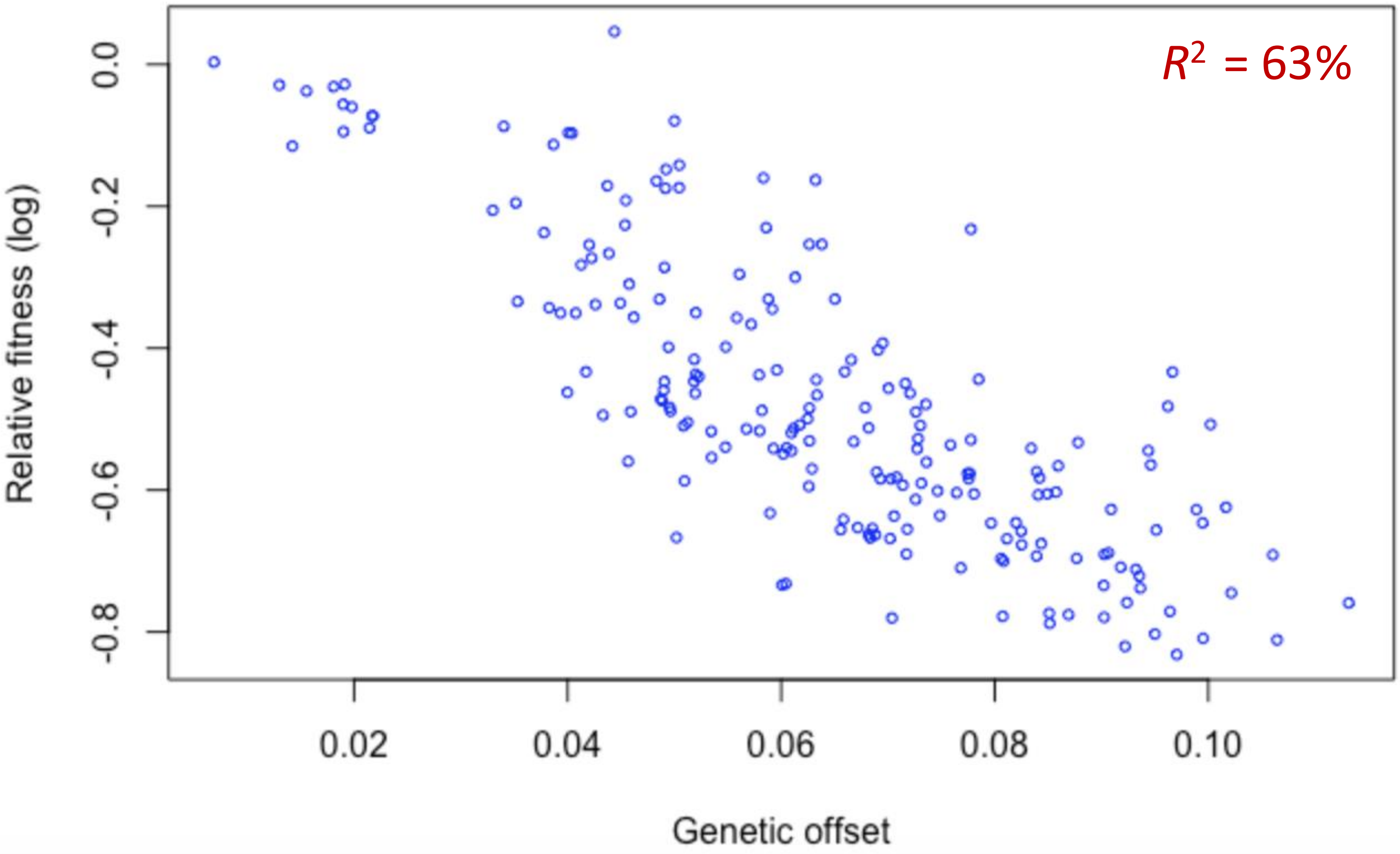
```r
## modified environment
X_pred <- offset_example$env.pred
```

2. Compute GF genomic offset (all locus or candidates)

```r
go_gf <- get_genetic_offset_gf(Y, X, X_pred, causal_set = 1:ncol(Y))
```

```r
causal_set = candidates)
```

Results



R^2 = 63%
Relative fitness (log)
0.0
-0.2
-0.4
-0.6
-0.8
0.02
0.04
0.06
0.08
0.10
Genetic offset

Correction for population structure

We can include latent factors as covariates into GF as follows

```
get_genetic_offset_gf(Y, X, X_pred, causal_set = 1:ncol(Y), confounding_var = mod_lfmm2@U)
```

In this example, using principal components or latent factors estimated in an LFMM as covariates did not lead to a substantial improvement in the predictions of genetic offset.

Resources

- For basic GEA and GO, there is a GF tutorial in the SSMPG 2025 GitHub repository

https://github.com/bcm-uga/SSMPG2025/blob/main/Tutorials/

- For more advanced questions on GEAs, GFs (and maps), ask us during the practical sessions