A quantitative theory for genomic offset statistics

Clément Gain[1], Bénédicte Rhoné[2,3], Philippe Cubry[2], Israfel Salazar[4], Florence Forbes[4], Yves Vigouroux[2], Flora Jay[5], Olivier François[1,4]

Authors' affiliations:

[1] Université Grenoble-Alpes, Centre National de la Recherche Scientifique, Grenoble INP, TIMC UMR 5525, 38000 Grenoble, France.

[2] DIADE, Université de Montpellier, Institut de Recherche pour le Développement, French Agricultural Research Centre for International Development (CIRAD), Montpellier, France

[3] CIRAD, UMR AGAP Institut, F-34398 Montpellier, France, and UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro, Montpellier, France

[4] Université Grenoble-Alpes, Centre National de la Recherche Scientifique, Grenoble INP, Inria Grenoble - Rhône-Alpes, LJK UMR 5224, 655 Avenue de l'Europe, 38335 Montbonnot, France.

[5] Université Paris-Saclay, Centre National de la Recherche Scientifique, Inria, Laboratoire Interdisciplinaire des Sciences du Numérique, UMR 9015 Orsay, France.

Corresponding author: `olivier.francois@univ-grenoble-alpes.fr`

Abstract

Genomic offset statistics predict the maladaptation of populations to rapid habitat alteration based on association of genotypes with environmental variation. Despite substantial evidence for empirical validity, genomic offset statistics have well-identified limitations, and lack a theory that would facilitate interpretations of predicted values. Here, we clarified the theoretical relationships between genomic offset statistics and unobserved fitness traits controlled by environmentally selected loci, and proposed a geometric measure to predict fitness after rapid change in local environment. he predictions of our theory were verified in computer simulations and in empirical data on African pearl millet (*Cenchrus americanus*) obtained from a common garden experiment. Our results proposed a unified perspective on genomic offset statistics, and provided a theoretical foundation necessary when considering their potential application in conservation management in the face of environmental change.

Keywords: Predictive Ecological Genomics, Genomic Offset, Climate Change, Local Adaptation, Pearl Millet.

Introduction

Maladaptation across environmental changes. Predicting maladaptation resulting from traits that evolved in one environment being placed in an altered environment is a long-standing question in ecology and evolution, originally termed as evolutionary traps or mismatches (Schlaepfer *et al.*, 2002; Cook and Saccheri, 2013). With the increasing availability of genomic data, a recent objective is to determine whether those shifts could be predicted from the genetic loci that control adaptive traits and the fitness effects of these loci in spatially varying environments, bypassing any direct phenotypic measurements (Capblancq *et al.*, 2020; Waldvogel *et al.*, 2020). This question is crucial to understand whether sudden changes in the species ecological niche, *i.e.*, the sum of the habitat conditions that allow individuals to survive and reproduce, can be sustained by natural populations (Grinnell, 1917; Hutchinson, 1957; Sork *et al.*, 2010; Jay *et al.*, 2012; Aitken and Whitlock, 2013; Schoville *et al.*, 2012; Foden *et al.*, 2019). To this aim, several approaches have incorporated genomic information on local adaptation into predictive measures of population maladaptation across ecological changes, called genomic offset (or genomic vulnerability) statistics (Fitzpatrick and Keller, 2015; Capblancq *et al.*, 2020; Waldvogel *et al.*, 2020).

Genomic offset statistics and their limitations. Genomic offset statistics first estimate a statistical relationship between environmental gradients and allelic frequencies using genotype-environment association (GEA) models (Forester *et al.*, 2018). The inferred relationship is then used to evaluate differences in predicted allelic frequencies at pairs of points in the ecological niche (Fitzpatrick and Keller, 2015; Rellstab *et al.*, 2016; Gougherty *et al.*, 2021). The central hypothesis is that

those statistics are predictive of changes in fitness traits that occur under altered environmental conditions (Capblancq *et al.*, 2020). Recent efforts combining trait measurements in common garden experiments or natural population censuses with landscape genomic data have shown that the loss of fitness due to abrupt environmental shift correlates well with genomic offset predictions (Bay *et al.*, 2018; Ruegg *et al.*, 2018; Rhoné *et al.*, 2020; Ingvarsson and Bernhardsson, 2020; Fitzpatrick *et al.*, 2021; Chen *et al.*, 2022; Sang *et al.*, 2022). Experiments in which organisms are placed into an environment that differs from the one in which the traits evolved are, however, not always feasible (or efficient). Genomic offsets – that can be calculated in field studies – offer then a reasonable alternative to common garden experiments in a wide spectrum of applications to model and non-model organisms.

Despite substantial evidence for empirical validity, the proposed measures of genomic offset have well-identified limitations due to migration and gene flow (but see Gougherty *et al.* (2021)), population structure or genomic load. They also have difficulties to account for polygenic effects or correlated predictors (Rellstab *et al.*, 2021; Aguirre-Liguori *et al.*, 2021; Hoffmann *et al.*, 2021). More importantly, different types of genomic offset statistics have been proposed in recent years (Fitzpatrick and Keller, 2015; Rellstab *et al.*, 2016; Capblancq and Forester, 2021), and the inferred values for each of those statistics have not been explicitly linked to fundamental measures in quantitative and population genetics. The proposed measures lack theoretical foundations that would clarify how those different statistics are related to fitness and to each other. Thus, there is an urgent need to propose theoretical developments that will facilitate biological interpretations of genomic offset statistics. Here, we developed a theoretical framework that links genomic offset statistics to adaptive trait

values controlled by ecological conditions, unifies existing approaches and addresses their limitations.

Results

Geometry of the ecological niche. We developed a geometric approach to the concept of genomic offset (GO) by defining a dot product of ecological predictors built on effect sizes of those predictors on allelic frequencies. Effect sizes, $(\mathbf{b}_\ell) = (b_{\ell j})$, were obtained from a GEA model of centered allelic frequencies on scaled predictors observed at a set of sampling locations. In that notation, ℓ stands for a locus, and j stands for a predictor. Effect sizes were corrected for the confounding effects of population structure and missing predictors (Methods: "GEA studies"). Given d ecological predictors, recorded in vector $\mathbf{x}$, and their altered versions based on some change in time or space, recorded in $\mathbf{x}^\star$, we defined a geometric GO – implemented as *genetic gap* in the computer package LEA – as a quadratic distance between the two vectors $\mathbf{x}$ and $\mathbf{x}^\star$

$$G^2(\mathbf{x}, \mathbf{x}^\star) = (\mathbf{x} - \mathbf{x}^\star)\mathbf{C}_\mathbf{b}(\mathbf{x} - \mathbf{x}^\star)^T \,, \tag{1}$$

where $\mathbf{C}_\mathbf{b} = \mathbb{E}[\mathbf{b}^T\mathbf{b}]$ is the empirical covariance matrix of environmental effect sizes. Here the notation $\mathbb{E}[.]$ stands for the empirical mean across genomic loci in the analysis, ideally the number of loci controlling adaptive traits. Because the reference allele defining the genotype at a particular locus can be changed without any impact on the GEA analysis, we assume that the average value of effect sizes across all genomic loci is null, $\mathbb{E}[\mathbf{b}] \approx 0$. Considering allelic frequencies predicted from the GEA model, $f(\mathbf{x}) = \mathbf{x}\mathbf{b}^T + \sum_{k=1}^K \mathbf{u}_k\mathbf{v}_k^T$ and $f(\mathbf{x}^\star) = \mathbf{x}^\star\mathbf{b}^T + \sum_{k=1}^K \mathbf{u}_k\mathbf{v}_k^T$, where the $\mathbf{u}_k$ represents

K confounding factors and $\mathbf{v}_k$ their loadings, we have

$$G^2(\mathbf{x}, \mathbf{x}^\star) = \mathbb{E}[((\mathbf{x} - \mathbf{x}^\star)\mathbf{b}^T)^2] = \mathbb{E}[(f(\mathbf{x}) - f(\mathbf{x}^\star))^2]\,. \tag{2}$$

Thus the geometric GO has a dual interpretation as a quadratic distance in environmental and in genetic space. The population genetic interpretation of the geometric GO is as the average value of Nei's $D_{\mathrm{ST}}/2$ ($= F_{\mathrm{ST}} \times H_{\mathrm{T}}/2$) for the set of loci assumed to be involved in local adaptation (Nei, 1973; François and Gain, 2021). As a genomic offset, the D_{ST} statistic can be calculated between pairs of population in space, but also in time, and it evaluates the genetic diversity between the populations in which $\mathbf{x}$ and $\mathbf{x}^\star$ are measured or forecasted.

Quantitative theory for genomic offset. We developed a quantitative theory for the geometric GO and for other GO statistics under the hypothesis of local stabilizing selection (Kimura, 1965; Lande, 1975). Under this hypothesis, observed allelic frequencies have reached local equilibria in which polygenic or quantitative characters are under natural selection for intermediate optimum phenotypes. The theory relies on a statistical model for an unobserved fitness trait for which a large number of small allelic effects mediate the effects of ecological predictors on fitness.

We defined $\omega(\mathbf{x}, \mathbf{x}^\star)$ to be the relative fitness value of a trait at equilibrium in environment $\mathbf{x}$ being placed in the altered environment $\mathbf{x}^\star$. Under local Gaussian stabilizing selection, we found that the value of the logarithm of altered fitness varies in proportion with the geometric GO (Figure 1, Box 1)

$$-\log \omega(\mathbf{x}, \mathbf{x}^\star) \propto G^2(\mathbf{x}, \mathbf{x}^\star)/2V_s\,, \tag{3}$$

where the V_s coefficient depends on the inherited variance and on the strength of stabilizing selection. In addition, the above equation remains valid when environmental predictors are indirectly related to the factors that influence the traits under selection, for example when those predictors are built on linear combinations of causal predictors for selection (Supporting Information: "Linear combination of predictors"). The geometric GO is thus robust to correlation in causal effects, and Eq. (3) extends to known and unknown linear combinations of those effects.

Unifying genomic offset statistics. Beyond defining a new geometric measure of genomic offset, the quantitative theory provides a unified framework for GO statistics based on redundancy analysis (RDA, Capblancq and Forester (2021)), the risk of nonadaptedness (Rona, Rellstab *et al.* (2016)), and gradient forests (GF, Fitzpatrick and Keller (2015)) (Supporting Information:"Relationships to other GO statistics"). The main result is that all GO statistics predict the logarithm of fitness, but not for the same shape of the (within-locality) selection gradient. When RDA is performed on both environmental and latent predictors, the RDA GO is theoretically equivalent to the geometric GO, and thus predicts relative fitness under the hypothesis of Gaussian selection within localities. The risk of nonadaptedness, which is defined as the average of allelic frequency differences instead of squared differences, makes the implicit assumption that the selection gradient is built upon an exponential (Laplace) curve. When the distribution of effect sizes is Gaussian, Rona is then related to the square root of the geometric GO (times $\sqrt{2/\pi}$). Like most machine learning techniques, GF is a nonparametric approach. In GF, no selection gradient is modelled a priori, but may be thought of as being estimated from the observed data. This might

be one reason for which GF require more information than linear approaches based on low-dimensional parameters. The GF GO nevertheless follows a construction similar to the geometric GO and the RDA GO.

Box 1. Genomic offset theory. Consider an (unobserved) fitness trait, z, for which a large number of genes mediate the effects of ecological predictors on organismal viability. Using Eq. (7) in (Barton *et al.*, 2017), the trait value is assumed to be controlled by L mutations each having infinitesimally small allelic effect of equal size, $a_\ell \approx \pm a/\sqrt{L}$, defining the trait value as a polygenic score, $z = \sum_{\ell=1}^{L} a_\ell y_\ell + e$. Here, y_ℓ is the allelic frequency at locus ℓ, expressed as deviation from the population mean, a_ℓ has random sign, a^2 controls the additive genetic variance, and the random term e models the non-genetic variance. The definition is equivalent to the more traditional decomposition of variance into inherited and non-inherited components (Figure S1). Assuming a local Gaussian stabilizing selection model, the relative fitness of the trait in environment $\mathbf{x}$ is equal to $\omega(z|\mathbf{x}) = \exp(-(z - z_{\text{opt}}(\mathbf{x}))^2/2V_S)$, where $1/V_S$ represents the strength of stabilizing selection. Conditional on local environment, the optimum, $z_{\text{opt}}(\mathbf{x})$, corresponds to the mean (or predicted) value of the trait, $\bar{z} = \sum_{\ell=1}^{L} a_\ell f_\ell(\mathbf{x})$. The logarithm of fitness for a trait at equilibrium in environment $\mathbf{x}$ being placed in the altered environment $\mathbf{x}^\star$ is thus equal to

$$-\log \omega(\mathbf{x}, \mathbf{x}^\star) = (\bar{z} - \bar{z}^\star)^2/2V_S \tag{4}$$

where $\bar{z}^\star = \sum_{\ell=1}^{L} a_\ell f_\ell(\mathbf{x}^\star)$. The difference in fitness traits, $(\bar{z} - \bar{z}^\star)$, is equal to $a(\mathbf{x} - \mathbf{x}^\star) \sum_{\ell=1}^{L} \mathbf{b}_\ell^T/\sqrt{L}$. According to the central limit theorem, the conditional distribution of $(\bar{z} - \bar{z}^\star)$ is Gaussian $N(0, a^2\mathcal{G}^2(\mathbf{x}, \mathbf{x}^\star))$, where $\mathcal{G}^2(\mathbf{x}, \mathbf{x}^\star)$ is defined from the theoretical – instead of empirical – effect size covariance matrix. The distribution of $(\bar{z} - \bar{z}^\star)^2$ is a non-standard chi-squared distribution with one degree of freedom

$$(\bar{z} - \bar{z}^\star)^2 \sim a^2\, \mathcal{G}^2(\mathbf{x}, \mathbf{x}^\star)\, \chi_1^2\,. \tag{5}$$

Since $G^2(\mathbf{x}, \mathbf{x}^\star) \approx \mathcal{G}^2(\mathbf{x}, \mathbf{x}^\star)$ for large L, the value of the logarithm of altered fitness varies in proportion with the geometric GO, where the proportionality coefficient is equal to $a^2\chi_1^2/2V_S$. The expected value is thus approximately equal to $\mathcal{G}^2(\mathbf{x}, \mathbf{x}^\star)/2V_s$, where $V_s = V_S/a^2$. Consideration of traits that are not at equilibrium in environment $\mathbf{x}$ adds an intercept term to the expected value, equal to $a^2\sigma_\epsilon^2/2V_S + \sigma_e^2/2V_S$, where σ_ϵ^2 is the residual variance in the GEA model and σ_e^2 is the non-inherited variance (Supporting Information: "Logarithm of altered fitness for non-optimal traits").

Validation of the theory. To illustrate the above theory, we analyzed simulated data in which adaptive traits were matched to ecological gradients by local Gaussian stabilizing selection (Figure 2A, Methods: "Simulation study", Supporting Information: "Extended simulation study") (Haller and Messer, 2019). Two environmental predictors playing the role of temperature and precipitation in the studied range were considered, as well as two additional non-causal predictors correlated to the first ones (Figure 2B). The median values of temperature and precipitation determined four broad types of environments from *dry/warm* to *wet/cold* conditions. As an outcome of the simulation, the genetic groups resulting from selection, drift and gene flow matched the environmental classes, generating high levels of correlation between environmental predictors and population structure in the GEA analysis (Figure S2). As predicted by equation (3), the values of the geometric GO computed according to equation (1) varied linearly with the logarithm of fitness after alteration of local conditions ($r^2 \approx 78\%, P < 0.001$, Figure 2C-D). The predictive power of the geometric GO was much higher than the predictive power of squared Euclidean environmental distance between predictors and their altered values ($r^2 \approx 45\%$, $J = 11.3$, $P < 0.001$). Although it was calculated on both causal and non-causal predictors, the GO adjusted almost perfectly to the quadratic function that determines the intensity of local Gaussian stabilizing selection ($r^2 = 97\%$, $P < 0.001$, Figure S3). The first two eigenvalues of the covariance matrix of environmental effect sizes were much larger than the last ones (Figure 2C). We found that the loadings on the first axes gave more weight to predictors associated with natural selection, while the loadings on the last axes weighted predictors that did not play a role in the simulated evolutionary process. Uninformative predictors were given only low weights in the calculation of

the GO statistic. Those results provided evidence that the largest eigenvalues that characterize the geometric GO contain useful information about local adaptation.

Extended simulation study. Expanding our case analysis, additional simulation scenarios were considered with traits under local stabilizing selection having distinct levels of polygenicity. Some cases were complicated by a strong correlation of environmental predictors with population structure. To overcome this complication, correction based on latent factors were included in all GO calculations (Methods:"GO computations"). As predicted by the theory, the values of the squared correlation between the GO statistic and the logarithm of fitness were very close to each other in all investigated cases (Figure 3, Figure S4). As expected, methods that did not use correction (undercorrection) or include population structure covariates (overcorrection) worked less well than methods with latent factor correction (Figures S5-S6). Once corrected, the GO statistics ranked similarly in all simulation scenarios. The ability of the geometric GO to predict the logarithm of fitness was equal to that of corrected RDA GO. It was slightly superior to that of Rona and to that of the GF GO. All GO statistics were highly correlated with the geometric GO (Figure S7). The geometric GO also exhibited high correlation with the quadratic distance between causal predictors explaining the traits under local stabilizing selection in the simulation model (Figure S8). This result supported the evidence of near-optimal fitness prediction by the GO statistics in all simulated evolutionary scenarios. When all genomic loci in the genotype matrix were included in the GO calculations, the predictions stayed close to those based on subsets of loci identified in the GEA analysis, GF GO reaching then performances similar to the other GO statistics (Figure

S9).

Evaluating the bias of linear allelic frequency predictions. An approximation made by the geometric and other GO statistics is that allelic frequencies are predicted by unconstrained linear functions of environmental predictors. To evaluate the impact of this approximation, we compared linear predictions to those of a logistic regression model, which are constrained between zero and one. For small environmental change, the effect sizes in the linear GEA model could be approximated by the effect sizes in the logistic regression multiplied by the heterozygosity at each locus (Supporting Information: "Bias of linear predictors"). The geometric GO was then accurately approximated by the squared distance between constrained genetic predictors, $\mathbb{E}[(f_c(\mathbf{x}) - f_c(\mathbf{x}^\star))^2]$ (Figure S10). Using a nonlinear machine learning model (Supporting Information: "Variational autoencoder GO"), we found again that the squared genetic distance between constrained genetic predictors strongly correlated with the geometric GO, supporting the approximation of fitness in altered environment using linear models (Figure S11).

Pearl millet common garden experiment. We hypothesized that GO statistics could predict the logarithm of fitness in pearl millet, a nutritious staple cereal cultivated in arid soils in sub-Saharan Africa (Rhoné *et al.*, 2020). Pearl millet is grown in a wide range of latitudes and climates with wide variety of ecotypes (landraces). The geometric GO and other measures of GO were estimated from 138,948 single-nucleotide polymorphisms for 170 Sahelian landraces in a two-year common garden experiment conducted in Sadoré (Niger) using loci identified in the GEA study (Figure 4A, Methods: "Pearl millet experiment"). For each landrace grown in the common

garden, the total weight of seeds was measured as a proxy of landrace fitness, which was explained by a Gaussian selection gradient (Figure S12). Including latent factor correction, GO statistics were computed using the climate condition at the location of origin of the landrace and the climate at the experimental site. All GO statistics displayed a consistent relationship with the logarithm of seed weight (Figure 4B, Figure 5). Loci identified in the GEA study increased the performance of GO statistics compared to using whole genomic data, and the improvements were substantial compared to methods that did not include correction for confounding factors (Figures S13-S14 and Table S1). The best predictions of fitness in the common garden were obtained with the geometric GO and with the corrected version of Rona ($r^2 = 61\%$, $P < 0.001$, Figure 5). The eigenvalues and eigenvectors of the covariance matrix of environmental effect sizes suggested that climatic conditions could be summarized in three axes. Temperature predictors were given higher importance in driving fitness variation than precipitation and solar radiation predictors (Figure S15).

Discussion

Quantitative theory. The geometric theory presented in our study provided a unified framework that not only explains why and when a GO statistic differs from the standard Euclidean environmental distance, but also allowed for a better understanding of previous measures of genomic offset. Based on models of local selection gradients, a theoretical analysis of GO statistics relying on Fisher's infinitesimal trait model was developed. In this framework, the geometric GO decays linearly with the logarithm of fitness in the altered environment. Although of much lower computational complexity, the geometric GO was proved to be equivalent to a GO based on

RDA, which justifies the use of RDA approaches under local Gaussian selection. The square root of the geometric GO was connected to Rona, and justifies the use of absolute differences in allele frequencies under exponential selection gradient.

Improving GO statistics. According to Rellstab *et al.* (2021), current GO statistics may provide wrong predictions due to the correlation between population structure at selectively neutral loci and environmental predictors. Built on unbiased effect sizes, the geometric GO, which is based on a unique model for GEA estimation and for GO prediction, addressed this problem by including latent factors as covariates in the prediction model. Latent factor corrections were then incorporated into all considered GO statistics, which increased their predictive performance compared to their traditional usage. Our versions of RDA GO and Rona – that slightly differ from original proposals – were implemented in the `R` package LEA. Although those changes led to improved statistics, the geometric GO reached higher predictive performance than the other GO approaches. Next, the geometric GO addressed the problem of correlated predictors by modeling the covariance of their effect sizes. The importance of predictors could be assessed by examining the eigenvalues and eigenvectors of the environmental effect size covariance matrix. The eigenvalues provide a natural ranking of the importance of each axis, similar to the cumulative importance curves in GF. When a statistical analysis includes redundant predictors, reproducing information already present in a reduced set of predictors, the geometric GO gave lower weight to those redundant predictors, and differed substantially from the Euclidean environmental distance. Generally, the principal benefit of genomic offset over purely environmental distances in predicting maladaptation comes from the weighting of en-

vironmental predictors by their effect sizes (Làruson *et al.*, 2022). All proposed GO approaches share the principle of weighting the environmental predictors by their strength of genetic association. For the vast majority of organisms where the most important predictors are unknown or for which common garden experiments are not efficient or unfeasible, genomic offset therefore provides a useful means for weighting the environmental predictors based on the information contained in allele frequencies.

Limitations. Our simulation models and our theoretical developments relied upon a model of genotype $\times$ environment interaction for fitness related to antagonistic pleiotropy, whereby native alleles are best adapted to local conditions (Kawecki, 2004; Anderson *et al.*, 2011). While antagonistic pleiotropy is an important mechanism for local adaptation, there are other types of interactions for fitness. If local adaptation is caused by conditional neutrality at many loci, where alleles show difference in fitness in one environment, but not in a contrasting environment, the predictive performances of GO statistics remain to be explored. In addition, GO statistics (except GF) are based on linear models for the relationship between genotype and environment. Linear models generate GO statistics that are invariant under translation in the niche, making predictions relevant at the center of the species distribution, but perhaps less relevant at margins of the range. While translational invariance could be corrected for by defining the offset as the average of squared differences between allelic frequencies in nonlinear models, we found that the results were very close to the linear models. An explanation may be that nonlinear machine learning models offer more flexible GO statistics than linear models, but that linear models achieve a better bias-variance trade-off than machine learning models, likely because less data

are needed for their application. Other conceptual limitations include gene flow and constraints on adaptive plasticity that might mitigate the effect of environmental change on fitness (Aguirre-Liguori *et al.*, 2021; Kawecki, 2004). As they do not use any observed information on fitness traits, GO statistics provide measures of expected fitness loss based on the indirect effects of environment mediated by loci under selection (Baron and Kenny, 1986). GO statistics are more accurate when non-genetic effects do not covary with environmental predictors. Lastly, we found that using candidate loci based on statistical significance in GEA improved prediction of fitness in altered conditions both in simulation and in real data analysis. We think that this happens because those studies may generally be underpowered, *i.e.*, a much larger sample size would increase the predictive power of GO statistics. Using a liberal threshold in GEA studies was considered as a trade-off between polygenicity and statistical significance, so that the GO measures could actually be based on polygenic scores while not erasing or blurring the genomic signals of local adaptation.

Pearl millet experiment. To compare predictions of local adaptation with empirical data, GO statistics were estimated in a common garden experiment on pearl millet landraces in sub-Saharan Africa. Using GF, the original study reported a squared correlation of $r^2 \approx 9.5 - 17\%$ for seed weight, indicating that higher genomic vulnerability was associated with lower fitness under the climatic conditions at the experimental site (Rhoné *et al.*, 2020). In our reanalysis, signals of local adaptation were consistent across all GO statistics, and improved fitness prediction substantially, up to a value of squared correlation equal to $r^2 \approx 61\%$. The results strengthened the conclusions of (Rhoné *et al.*, 2020), and supported the use of GO statistics in

predictions of fitness values across the sub-Saharan area.

Conclusions. Considering a duality between genetic space and environmental space, we developed a theoretical framework that linked GO statistics to a non-Euclidean geometry of the ecological niche. The geometric GO, as well as the modified Rona statistic, were implemented in the genetic gap function of the `R` package LEA (Gain and François, 2021). As a result of the quantitative theory, interpretations in terms of fitness in the altered environment were proposed, unifying several existing approaches, and addressing some of their limitations. Based on extensive numerical simulations and on data collected in a common garden experiment, our study indicated that GO statistics are important tools for conservation management in the face of climate change.

Materials and Methods

GEA studies. GEA studies and estimates of environmental effect sizes were performed based on LFMMs in the computer package LEA v3.9 (Caye *et al.*, 2019; Gain and François, 2021). In LFMMs, allelic frequencies are modelled at each genomic locus of a genotype matrix as a mixed response of observed environmental variables with fixed effects and K unobserved latent factors. The number of latent factors was estimated from the screeplot of a principal component analysis of the genotype matrix. Loci with minor allele frequency less than 10% were filtered out the analysis. Statistical significance was determined by using the `R` package `qvalue` at a level of false discovery rate equal to 10%.

GO computations. RDA was performed by using principal components of fitted values of the GEA regression model. Rona was computed as the average value of the absolute distance between predicted allelic frequencies across genomic loci (de Aquino *et al.*, 2022; Rellstab *et al.*, 2016). GF computations were performed using the `R` package `gradientForest` version 0.1. For consistency, we reported squared values of GO statistics in RDA and GF. Unless specified, GO statistics were computed on the loci detected in the GEA study, i.e., a same set of loci for all methods. To correct statistics for the confounding effect of population structure, all analyzes were performed conditional on the factors estimated in the LFMM analysis (Supporting Information:"GO computations").

Simulation study. Spatially-explicit individual-based simulations were performed using SLiM 3.7 (Haller and Messer, 2019) (Supporting Information:"Extended simulation study"). Each individual genome contained neutral mutations and quantitative trait loci (QTLs) under local stabilizing selection from a two-dimensional environment. The probability of survival of an individual genome in the next generation was computed as the product of density regulation and fitness. We designed four classes of scenarios, including weakly or highly polygenic traits, and weak or high correlation of environment with population structure. In scenarios with high polygenicity, traits controlled by 120 mutations with additive effects were matched to each environmental variable by local stabilizing selection. In weakly polygenic scenarios, the traits were controlled by 10 mutations. Scenarios with high confounding effects were initiated in a demographic range expansion process, creating correlation between environment and allelic frequencies at the genome level. For each scenario, thirty replicates were

run with distinct seed values of the random generator. At the end of a simulation, individual geographic coordinates, environmental variables and individual fitness values before and after instantaneous environmental change were recorded. Paired t-tests were used to test statistical differences in the mean of predictive performances for the geometric GO and the other GO statistics.

Empirical study. Methods regarding the common garden experiment on Pearl millet landraces conducted in Sadoré (13° 14' 0" N, 2° 17' 0" E, Niger, Africa) were described by Rhoné *et al.* (2020). For each of 170 landraces grown in the common garden, the total weight of seeds was measured by harvesting the main spike in ten plants per landrace sown during two consecutive years and was used as a proxy of landrace fitness. For each landrace grown in the common garden, environmental predictors, $\mathbf{x}$, were obtained at the location of origin of the landrace, and $\mathbf{x}^\star$ corresponded to the local conditions in Sadoré. We made the hypothesis that the mean total weight of seeds for a landrace was proportional to $\omega(\mathbf{x}, \mathbf{x}^\star)$ in the common garden. Using 100 plants per landrace in a pool-sequencing design, allelic frequencies were inferred at 138,948 single-nucleotide polymorphisms. Climate data were used to compute 157 metrics in three categories, precipitation, temperature (mean, maximum and minimum near surface air temperature), and surface downwelling shortwave radiation, that were reduced by principal component analysis (27 axes). GO statistics were computed using the climate condition ($\mathbf{x}$) at the location of origin of the landrace and the climate conditions ($\mathbf{x}^\star$) at the experimental site. For each GO statistic, we estimated a linear relationship with the logarithm of the mean total weight of seeds and used Pearson's squared correlation to evaluate the goodness of fit. The J-test

was used to test differences between predictive performances, corresponding to R-squared for distinct regression models, of the geometric GO and other GO statistics (Davidson and MacKinnon, 1981).

Data Availability. The pearl millet data have already been published, and have permissions appropriate for fully public release.

Code Availability. The codes necessary to reproduce the simulations and data analyses of this study are available at `https://github.com/bcm-uga/geneticgap` under GNU General Public License v3.0 The geometric GO is implemented in the genetic gap function of the `R` package LEA (version number $> 3.9.5$) available from the public repository `bioconductor` and `https://github.com/bcm-uga/LEA` (latest version).

References

Anderson JT, Willis JH, Mitchell-Olds T. 2011. Evolutionary genetics of plant adaptation. *Trends Genet.* 27(7):258-266.

Aitken SN, Whitlock MC. 2013. Assisted gene flow to facilitate local adaptation to climate change. *Annu Rev Ecol Evol Syst.* 44:367-388.

Aguirre-Liguori JA, Ramirez-Barahona S, Gaut BS. 2021. The evolutionary genomics of species' responses to climate change. *Nat Ecol Evol.* 5(10):1350-1360.

de Aquino SO, Kiwuka C, Tournebize R, Gain C, Marraccini P, Mariac C, Bethune

K, Couderc M, Cubry P, Andrade AC, et al. 2022. Adaptive potential of Coffea canephora from Uganda in response to climate change. *Mol Ecol.* 31:1800-1819.

Baron RM, Kenny DA. 1986. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J Pers Soc Psychol.* 5:1173-1182.

Barton NH, Etheridge AM, Véber A. 2017. The infinitesimal model: Definition, derivation, and implications. *Theor Pop Biol.* 118:50-73.

Bay RA, Harrigan RJ, Le Underwood V, Gibbs HL, Smith TB, Ruegg K. 2018. Genomic signals of selection predict climate-driven population declines in a migratory bird. *Science* 359(6371):83-86.

Capblancq T, Fitzpatrick MC, Bay RA, Exposito-Alonso M, Keller SR. 2020. Genomic prediction of (mal)adaptation across current and future climatic landscapes. *Annu Rev Ecol Evol Syst.* 51:245-269.

Capblancq T, Forester BR. 2021. Redundancy analysis: A Swiss army knife for landscape genomics. *Methods Ecol Evol.* 12(12):2298-2309.

Caye K, Jumentier B, Lepeule J, François O. 2019. LFMM 2: Fast and accurate inference of gene-environment associations in genome-wide studies. *Mol Biol Evol.* 36(4):852-860.

Chen Y, Jiang Z, Fan P, Ericson PG, Song G, Luo X, Lei F, Qu Y. 2022. The combination of genomic offset and niche modelling provides insights into climate change-driven vulnerability. *Nat. Commun.* 13:1-15.

Cook LM, Saccheri IJ. 2013. The peppered moth and industrial melanism: Evolution of a natural selection case study. *Heredity* 110:207-212.

Davidson R, MacKinnon J. 1981. Several tests for model specification in the presence of alternative hypotheses. *Econometrica* 49:781-793.

Fitzpatrick MC, Keller SR. 2015. Ecological genomics meets community-level modelling of biodiversity: Mapping the genomic landscape of current and future environmental adaptation. *Ecol Lett.* 18(1):1-16.

Fitzpatrick MC, Chhatre VE, Soolanayakanahally RY, Keller SR. 2021. Experimental support for genomic prediction of climate maladaptation using the machine learning approach Gradient Forests. *Mol Ecol Res.* 21(8):2749-2765.

Foden WB, Young BE, Akçakaya HR, Garcia RA, Hoffmann AA, Stein BA, Thomas CD, Wheatley CJ, Bickford D, Carr JA, et al. 2019. Climate change vulnerability assessment of species. *Wiley Interdiscip Rev Clim Change.* 10(1):e551.

Forester BR, Lasky JR, Wagner HH, Urban DL. 2018. Comparing methods for detecting multilocus adaptation with multivariate genotype-environment associations. *Mol Ecol.* 27(9):2215-2233.

François O, Gain C. 2021. A spectral theory for Wright's inbreeding coefficients and related quantities. *PLoS Genet.* 17(7):e1009665.

Gain C, François O. 2021. LEA 3: Factor models in population genetics and ecological genomics with R. *Mol Ecol Res.* 21(8):2738-2748.

Gillespie JH. 2004. Population genetics: a concise guide. John Hopkins University Press, Baltimore, Maryland, USA.

Grinnell J. 1917. The niche-relationships of the California thrasher. *The Auk* 34:427-433.

Gougherty AV, Keller SR, Fitzpatrick MC. 2021. Maladaptation, migration and extirpation fuel climate change risk in a forest tree species. *Nat Clim Change.* 11:166-171.

Haller B, Messer PW. 2019. SLiM 3: Forward genetic simulations beyond the Wright-Fisher model. *Mol Biol Evol.* 36(3):632-637.

Hoffmann AA, Weeks AR, Sgrò CM. 2021. Opportunities and challenges in assessing climate change vulnerability through genomics. *Cell* 184(6):1420-1425.

Hutchinson GE. 1957. Concluding Remarks. *Cold Spring Harb Symp Quant Biol* 22:415-427.

Ingvarsson PK, Bernhardsson C. 2020. Genome-wide signatures of environmental adaptation in European aspen (*Populus tremula*) under current and future climate conditions. *Evol Appl.* 13(1):132-142.

Jay F, Manel S, Alvarez N, Durand EY, Thuiller W, Holderegger R, Taberlet P, François O. 2012. Forecasting changes in population genetic structure of alpine plants in response to global warming.*Mol Ecol.* 21(10):2354-2368.

Kawecki TJ, Ebert D. 2004. Conceptual issues in local adaptation. *Ecol. Lett.* 7(12):1225-1241.

Kimura M. 1965. A stochastic model concerning the maintenance of genetic variability in quantitative characters. *Proc Natl Acad Sci USA.* 54:731-736.

Lande R. 1975. The maintenance of genetic variability by mutation in a polygenic character with linked loci. *Genet Res.* 26(3):221-35.

Làruson ÀJ, Fitzpatrick MC, Keller SR, Haller BC, Lotterhos KE. 2022. Seeing the forest for the trees: Assessing genetic offset predictions with Gradient Forest. *Evol Appl.* 15(3):403-416.

Nei M. 1973. Analysis of gene diversity in subdivided populations. *Proc Natl Acad Sci USA.* 70:3321-23.

Rellstab C, Zoller S, Walthert L, Lesur I, Pluess AR, Graf R, Bodénès C, Sperisen C, Kremer A, Gugerli F. 2016. Signatures of local adaptation in candidate genes of oaks (Quercus spp.) with respect to present and future climatic conditions. *Mol Ecol.* 25(23), 5907-5924.

Rellstab C, Dauphin B, Exposito-Alonso M. 2021. Prospects and limitations of genomic offset in conservation management. *Evol Appl.* 14(5):1202-1212.

Rhoné B, Defrance D, Berthouly-Salazar C, Mariac C, Cubry P, Couderc M, Dequincey A, Assoumanne A, Kane NA, Sultan B, et al. 2020. Pearl millet genomic vulnerability to climate change in West Africa highlights the need for regional collaboration. *Nat. Commun.* 11(1):1-9.

Ruegg K, Bay RA, Anderson EC, Saracco JF, Harrigan RJ, Whitfield M, Paxton EH , Smith TB. 2018. Ecological genomics predicts climate vulnerability in an endangered southwestern songbird. *Ecol Lett.* 21(7):1085-1096.

Sang, Y., Long, Z., Dan, X., Feng, J., Shi, T., Jia, C., Zhang X , Lai Q, Yang G, Zhang H, et al. 2022. Genomic insights into local adaptation and future climate-induced vulnerability of a keystone forest tree in East Asia. *Nat. Commun.* 13(1):1-14.

Schoville SD, Bonin A, François O, Lobreaux S, Melodelima C, Manel S. 2012. Adaptive genetic variation on the landscape: methods and cases. *Annu Rev Ecol Evol Syst.* 43:23-43.

Schlaepfer MA, Runge MC, Sherman PW. 2002. Ecological and evolutionary traps. *Trends Ecol Evol.* 17:474-480.

Sork VL, Davis FW, Westfall R, Flint A, Ikegami M, Wang H, Grivet D. 2010. Gene movement and genetic association with regional climate gradients in California valley oak (Quercus lobata Née) in the face of climate change. *Mol Ecol.* 19(17):3806-3823.

Waldvogel A-M, Feldmeyer B, Rolshausen G, Exposito-Alonso M, Rellstab C, Kofler R, Mock T , Schmid K, Schmitt I, Thomas Bataillon T, et al. 2020. Evolutionary genomics can improve prediction of species' responses to climate change. *Evol Lett.* 4(1), 4-18.

Acknowledgments

This work received support from the French National Research Agency, projects PEG (grant number ANR-22-CE45-0033), Afradapt (grant number ANR-22-CE32-0008), and ETAPE (grant number ANR-18-CE36-0005). The authors are grateful to Thibaut Capblancq for many interactions and fruitful discussions.

Author contributions

B.R., P.C., Y.V., I.S., F.F. contributed analyzes and helped drafting the manuscript. C.G., F.J. and O.F. conceived the study, developed the method, carried out analyzes, and wrote the manuscript.

Competing Interests

The authors declare no competing interests.

Additional information

Supporting methods and materials, supplementary figures and tables, are available online.

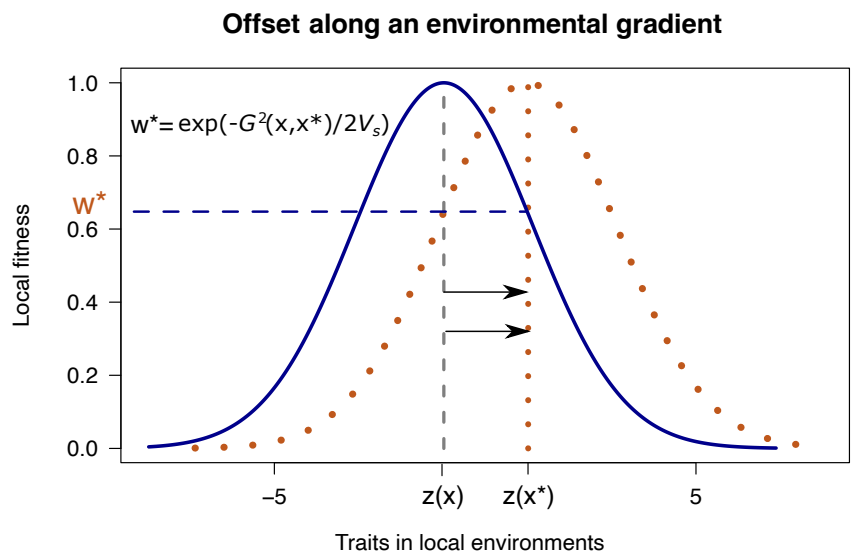

Figure 1. Geometric offset (genetic gap) under local Gaussian stabilizing selection. The two points, $z(\mathbf{x}) = \bar{z}$ and $z(\mathbf{x}^\star) = \bar{z}^\star$, represent locally optimal values of an adaptive trait in respective environments $\mathbf{x}$ and $\mathbf{x}^\star$. The curves display the fitness values for the trait in each environment. An organism with trait $z(\mathbf{x})$, optimal in environment $\mathbf{x}$, being placed in altered environment $\mathbf{x}^\star$, has a fitness value equal to $\omega^\star = \exp(-G^2(\mathbf{x}, \mathbf{x}^\star)/2V_s)$, where $G^2(\mathbf{x}, \mathbf{x}^\star)$ is the genomic offset (horizontal dashed line), and V_s is defined in text.

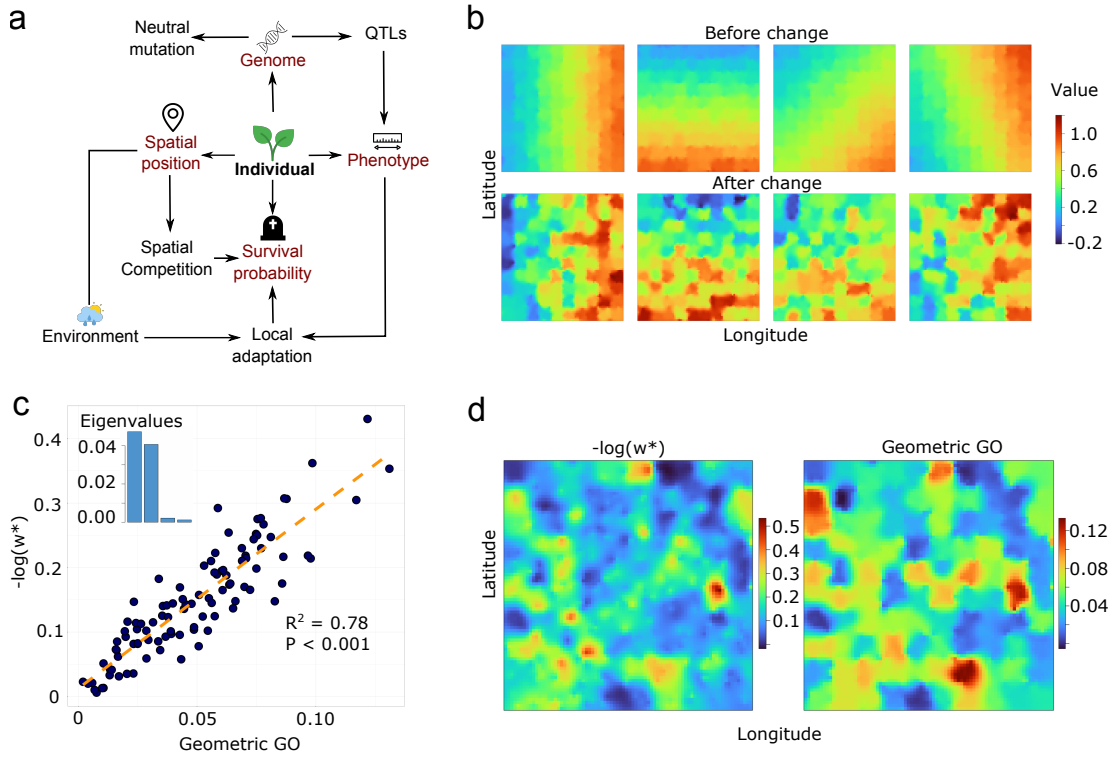

Figure 2. Simulation of fitness traits and geometric offset. **a**) Spatial individual-based forward simulations: Adaptive traits were matched to ecological gradients by local Gaussian stabilizing selection. **b**) Geographic maps of four environmental predictors before and after change. **c**) Logarithm of altered fitness values as a function of geometric offset. The eigenvalues of the covariance matrix of environmental effect sizes are displayed in the top left corner. **d**) Geographic maps of the logarithm of altered fitness values (left) and geometric offset (right).

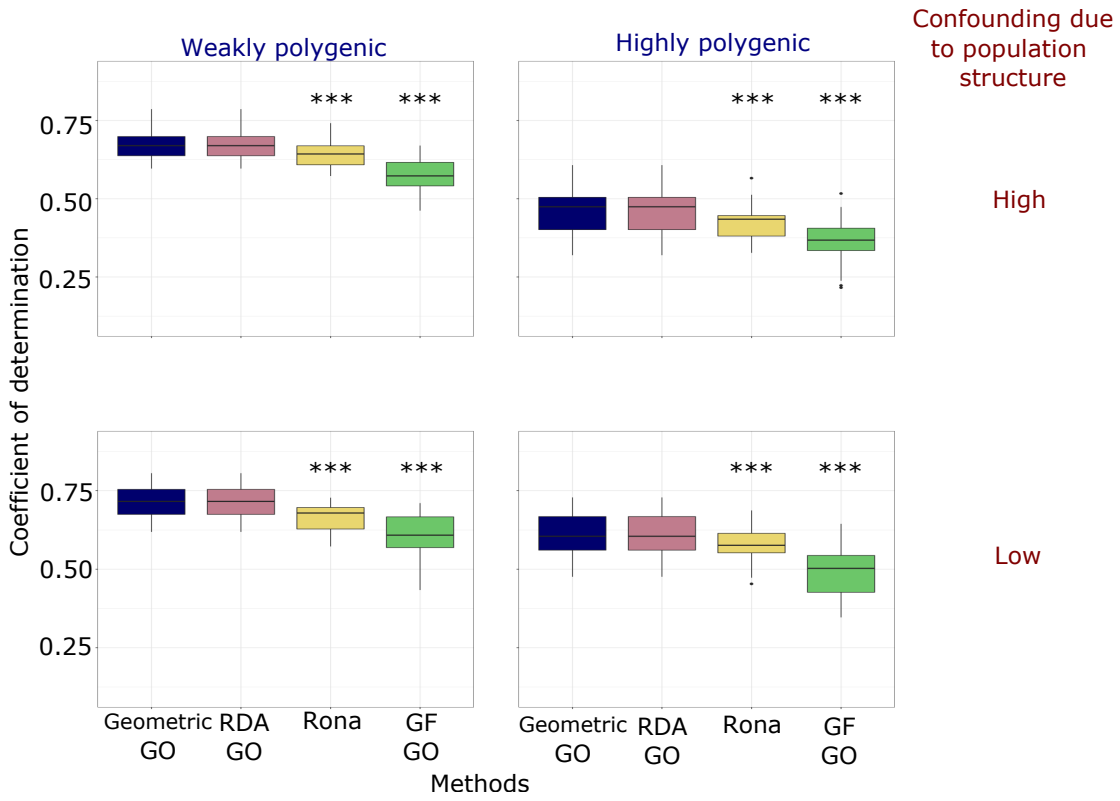

Figure 3. Predictive performances of GO statistics. Proportion of variance of fitness in the altered environment explained by GO statistics (coefficient of determination). Four scenarios with distinct levels of polygenicity in adaptive traits and correlation of environmental predictors with population structure were implemented. Significance values were based on paired t-tests of the difference in mean performance for each GO statistic relative to the geometric GO (***: $P < 0.001$). Boxplots display the median, the first quartile, the third quartile, and the whiskers of distributions. The upper whisker extends from the hinge to the largest value no further than 1.5 inter-quartile range (IQR) from the hinge. The lower whisker extends from the hinge to the smallest value at most 1.5 IQR of the hinge. Extreme values are represented by dots.

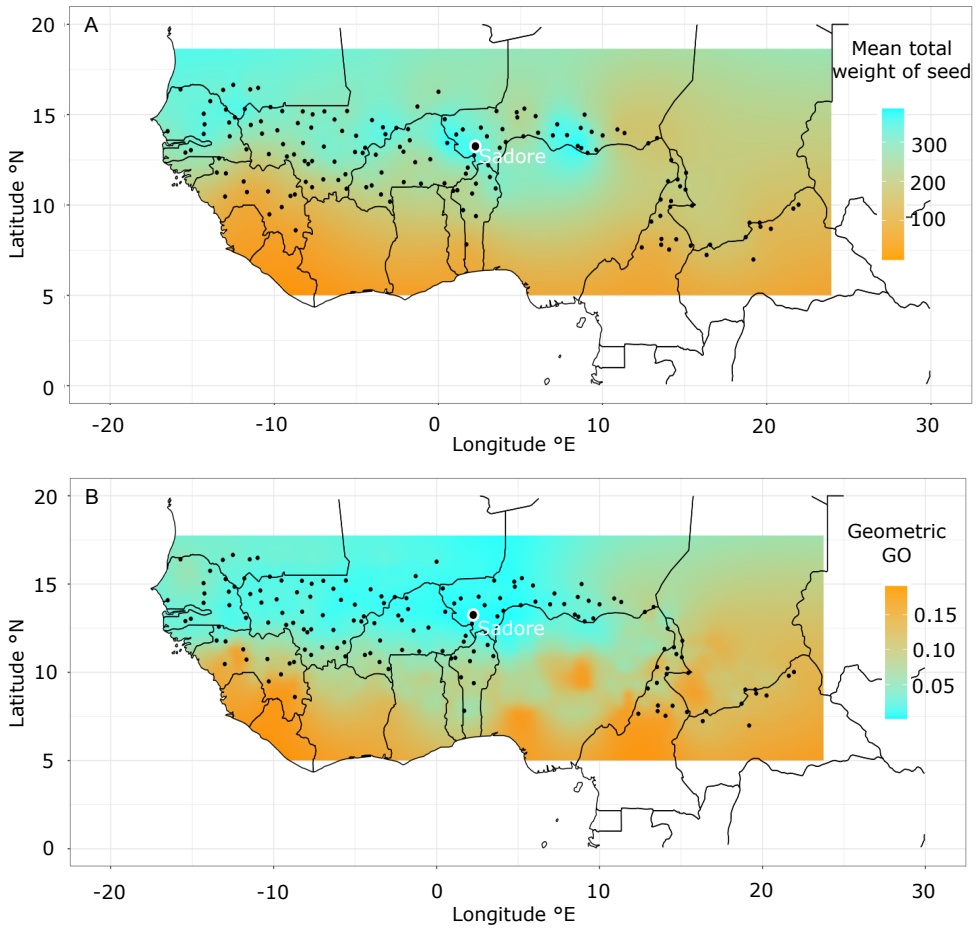

Figure 4. Interpolated fitness gradient and genomic offset for pearl millet landraces. A) Fitness values (log) measured as the mean total seed weight for each pearl millet landrace in the common garden experiment located in Sadoré (Niger). B) Values of the geometric genomic offset. Locations of landrace origin are represented as dots. Values at unsampled locations were interpolated from the nearest sampled location using the inverse distance weighting method.

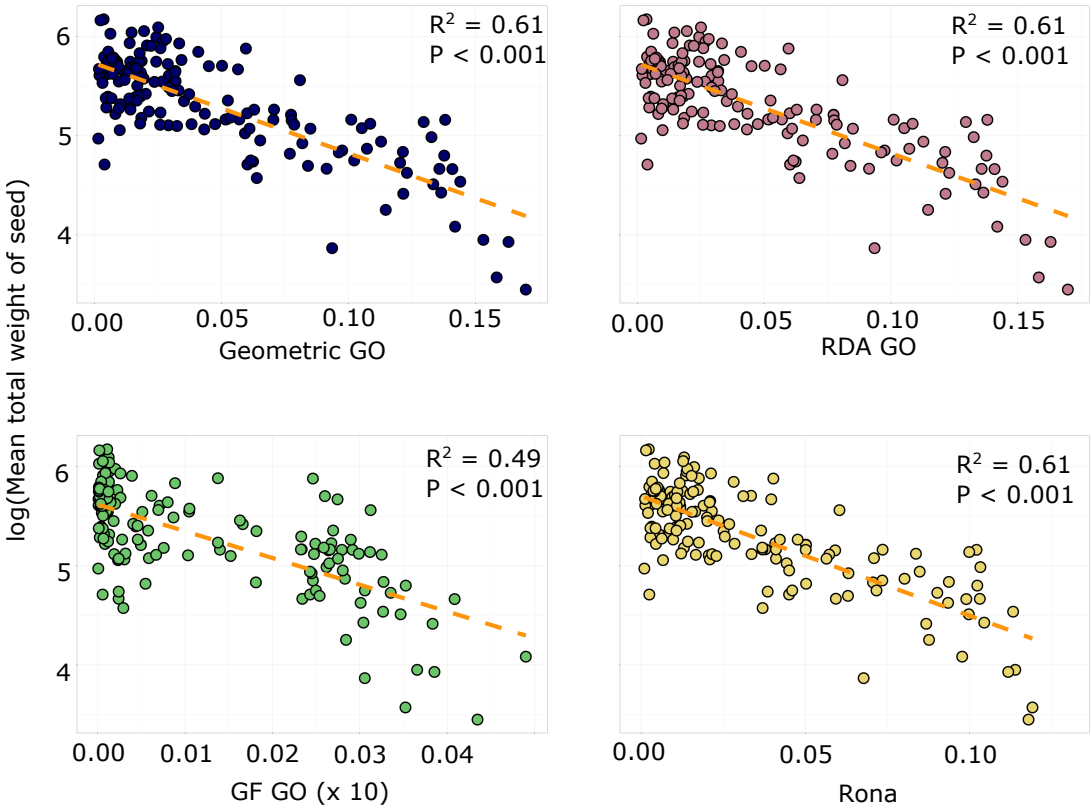

Figure 5. Logarithm of fitness in the common garden as a function of the GO statistic. Latent factor corrections were included in the calculation of all GO statistics (ten factors). Fitness was evaluated as the mean total weight of seed for 170 pearl millet landraces. GO values for GF were multiplied by a factor of ten.