# Predicting species invasiveness with genomic data: Is genomic offset related to establishment probability?

Louise Camus, Mathieu Gautier, Simon Boitard

▶ **To cite this version:**

Louise Camus, Mathieu Gautier, Simon Boitard. Predicting species invasiveness with genomic data: Is genomic offset related to establishment probability?. Evolutionary Applications, 2024, 17 (6), pp.e13709. 10.1111/eva.13709 . hal-04621972

**HAL Id: hal-04621972**
**https://hal.inrae.fr/hal-04621972v1**

Submitted on 24 Jun 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Check for updates

Received: 19 February 2024 | Revised: 30 April 2024 | Accepted: 4 May 2024

DOI: 10.1111/eva.13709

Evolutionary Applications Open Access WILEY

ORIGINAL ARTICLE

# Predicting species invasiveness with genomic data: Is genomic offset related to establishment probability?

Louise Camus | Mathieu Gautier | Simon Boitard

CBGP, INRAE, CIRAD, IRD, L'institut Agro, Université de Montpellier, Montpellier, France

**Correspondence**
Louise Camus, Mathieu Gautier and Simon Boitard, CBGP, INRAE, CIRAD, IRD, L'institut Agro, Université de Montpellier, Montpellier, France.
Email: louise.camus@inrae.fr; mathieu.gautier@inrae.fr and simon.boitard@inrae.fr

**Funding information**
Région Occitanie Pyrénées-Méditerranée; Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement

## Abstract

Predicting the risk of establishment and spread of populations outside their native range represents a major challenge in evolutionary biology. Various methods have recently been developed to estimate population (mal)adaptation to a new environment with genomic data via so-called Genomic Offset (GO) statistics. These approaches are particularly promising for studying invasive species but have still rarely been used in this context. Here, we evaluated the relationship between GO and the establishment probability of a population in a new environment using both in silico and empirical data. First, we designed invasion simulations to evaluate the ability to predict establishment probability of two GO computation methods (Geometric GO and Gradient Forest) under several conditions. Additionally, we aimed to evaluate the interpretability of absolute Geometric GO values, which theoretically represent the adaptive genetic distance between populations from distinct environments. Second, utilizing public empirical data from the crop pest species *Bactrocera tryoni*, a fruit fly native from Northern Australia, we computed GO between "source" populations and a diverse range of locations within invaded areas. This practical application of GO within the context of a biological invasion underscores its potential in providing insights and guiding recommendations for future invasion risk assessment. Overall, our results suggest that GO statistics represent good predictors of the establishment probability and may thus inform invasion risk, although the influence of several factors on prediction performance (e.g., propagule pressure or admixture) will need further investigation.

**KEYWORDS**
*Bactrocera tryoni*, biological invasions, GEA, genomic offset, local adaptation

## 1 | INTRODUCTION

Predicting how natural species adapt to environmental change is crucial in evolutionary biology. This knowledge not only furthers our understanding of evolutionary processes involved in adaptation but also holds practical implications, particularly in the current context of global changes. At the within-species level, the variation of abiotic and biotic conditions over a species distribution range favors the evolution of locally adapted populations,

Mathieu Gautier and Simon Boitard are joint senior authors on this work.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2024 The Author(s). *Evolutionary Applications* published by John Wiley & Sons Ltd.

which has been documented in a wide diversity of taxonomic groups (Wadgymar et al., 2022). This dynamic process, rooted in genetic variations, intricately shapes the fitness-related traits of organisms in diverse habitats. Thus, characterizing the genetic underpinnings of local adaptation over varied habitats as snapshots in space can inform how populations might respond or adapt over time to local environmental alterations. With the advent and democratization of high-throughput sequencing technologies, such a "space-for-time substitution" approach has thus been recently introduced in the population genomics field to forecast the potential impact of changing conditions on species' vulnerability (Capblancq, Fitzpatrick, et al., 2020; Rellstab et al., 2021).

In practice, these population-based approaches rely on Genome–Environment Association (GEA) methods to detect associations between the frequencies of adaptive genetic variants and environmental or phenotypic covariables. The GEA step also enables modeling the structure (or composition) of genetic diversity across populations and discerning which adaptive genetic variants and extrinsic variables may play a role in adaptation (Bogaerts-Márquez et al., 2021; Capblancq, Morin, et al., 2020; Ingvarsson & Bernhardsson, 2020; Ruegg et al., 2018). Subsequently, these statistical associations can be used to predict the optimal theoretical genetic composition (i.e., frequencies for different adaptive genetic variants) providing the highest fitness in a new given environment. The difference between the optimal genetic make-up in a new environment (according to the GEA modeling) and that of a population of interest has been referred to as Genomic Offset (GO) (Fitzpatrick & Keller, 2015; Gain et al., 2023; Rellstab et al., 2021). It is aimed at quantifying maladaptation risk: higher GO indicates greater risk of mismatch between population genetic composition and the new environment. Alternatively, GO measures can be viewed as weighted environmental distances based on covariables characterizing the environment of the populations, weighted according to their relative impact on the adaptive genetic composition of the populations (Fitzpatrick & Keller, 2015; Gain et al., 2023).

In recent years, GO approaches have been widely adopted in the field of conservation biology, being employed across a diverse array of taxa to estimate populations vulnerability to climate change to inform future management action (Borrell et al., 2020; Capblancq, Morin, et al., 2020; Morgan et al., 2020; Rhoné et al., 2020; Ruegg et al., 2018; Zhang, Chen, et al., 2023). These methods indeed allow accounting for population local adaptation within species, which is not achievable through conventional Species Distribution Modeling (SDM) methods that assume niche uniformity. Another application field where GO measures could prove useful is for the study of invasive populations. Indeed, global trade and climate change amplify the need to develop strategies allowing to predict future biological invasions (Gallien et al., 2010) and to mitigate their negative impacts (Hulme, 2017, 2021), especially for species of agricultural interest, for example, crop pests which represent a significant threat to global food security (Bradshaw et al., 2016; Bruce, 2010). In this context, GO measures could help predict the optimal regions for population establishment, taking into account intraspecific local adaptation.

Over the past few years, several methods have been proposed to compute GO. Among them, two widely used are the Risk Of Non-Adaptedness or RONA (Rellstab et al., 2016) and Redundancy Analysis or RDA (Capblancq & Forester, 2021), which model a linear relationship between allele frequencies and extrinsic covariates. Another recent linear method, known as Geometric GO (gGO), has shown particularly promising performance (Gain et al., 2023). Geometric GO relies on estimates of regression coefficients between population allele frequencies and environmental variables using Latent Factor Mixed Modeling or LFMM (Frichot et al., 2013), that is, the effect sizes of each environmental covariable on the variation of allele frequencies, while accounting for neutral population structure through the simultaneous estimation of latent factors (i.e., confounding factors). Under certain conditions, gGO is strictly equivalent to GO computed with RDA (Gain et al., 2023). Conversely, two alternative methods have been recently proposed to estimate GO in the pioneering study by Fitzpatrick and Keller (2015). These methods rely on the Gradient Forest (GF) algorithm (Ellis et al., 2012) and the Generalized Dissimilarity Modeling (GDM) approach (Ferrier et al., 2007) that accommodate nonlinear relationships between allele frequencies and covariates. They rely on turnover curves that describe the rate of genetic change along a gradient of environmental values but do not account directly for the confounding effects of neutral population structure. The GF method, based on the machine learning random forest approach (Breiman, 2001), has been used in many recent studies (Adam et al., 2022; Lachmuth et al., 2023; Ruegg et al., 2018; Zhang, Guo, et al., 2023).

Beyond the modeling differences summarized above, all GO approaches assume that (i) the populations adapt to their environment through pre-existing variants (rather than de novo mutations); (ii) the genome–environment relationship remains constant over time for future predictions ("space-for-time" hypothesis); and (iii) the populations studied are already adapted to their current environment (Capblancq, Fitzpatrick, et al., 2020; Rellstab et al., 2021). Several studies recently evaluated GO-based methods by comparing predicted maladaptation (quantified with GO) against fitness-related traits. This was done through in silico simulations (Gain et al., 2023; Láruson et al., 2022; Lotterhos, 2023) or by studying populations in common gardens (Archambeau, 2022; Fitzpatrick et al., 2021; Rhoné et al., 2020). Encouragingly, these studies often found that higher predicted maladaptation aligned with reduced realized fitness. However, none of these studies considered the specific situation of biological invasions, although GO measures are beginning to be applied within this framework (Chen et al., 2021, 2023).

In response to this gap, we here propose an evaluation of GO measures in the context of biological invasions. A primary objective of our study is to assess the predictive performance of GO, and we therefore evaluate the correlation between several GO measures and invasive population establishment probability. Through this analysis, our aim is twofold: first, to gauge the performances of GO in predicting establishment probabilities, and second, to identify factors that may hinder or enhance its predictive accuracy. This comprehensive investigation allows us to offer

17524571, 2024, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/eva.13709 by Cochrane France, Wiley Online Library on [24/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

practical recommendations for effectively employing GO with empirical data in invasion biology. We further propose some methodological innovations including (i) a novel approach to compute gGO (gGO BayPass) based on the BayPass GEA model (Gautier, 2015); (ii) an improvement in the computational efficiency of the Gradient Forest package to accommodate larger datasets; and (iii) an evaluation of interpretation of the absolute value of gGO, most GO metrics being relative and not directly tied to measurable aspects thereby complicating their biological interpretation across different datasets. These methodological advancements enable the broadening of the application of GO to diverse datasets, while the insights gained from evaluating the absolute gGO values can provide valuable guidance for understanding their interpretation for species management. Finally, for illustrative purpose and to exemplify the practical application of GO in the context of biological invasion, we also analyzed the data recently published by Popa-Báez et al. (2020) regarding *Bactrocera tryoni*, a tropical invasive fruit pest fly, native from Australia. Our analysis specifically aims to identify areas at risk of invasion, demonstrating the utility of GO in informing strategic invasion management decisions.

## 2 | MATERIALS AND METHODS

### 2.1 | Simulation study

To evaluate the relevance of GO in the context of biological invasion, we simulated the evolutionary dynamics of populations under biological invasion scenarios using SLiM v4.0.1 (Messer, 2013). In the first phase, we simulated 25 populations, each consisting of 1000 individuals, within a native area under a Wright–Fisher model including spatial structure and non-uniform selection constraints in order to produce genetic patterns of local adaptation. In this native area, each population was associated with two environmental optima (with values ranging from −1 to 1), to which individuals adapt through QTNs (Quantitative Trait Nucleotide). In the second phase, individuals from a given source population in the native area were randomly selected to invade a new environment, also characterized by given values for the two environmental optima.

The simulation framework in the native area was inspired by the work of Láruson et al. (2022), who evaluated the correlation between several GO measures and population fitness. However, because one of the primary objectives of our study was to examine GO in the context of biological invasion, we adapted the simulations, especially in the invaded area, to more closely match life history traits of invasive species, such as an important number of offspring, short generation time, and overlapping generations (Sakai et al., 2001; Zhao et al., 2023) (Note S1). While these characteristics are shared by many invasive species, the specific parameter values considered here were informed by literature focusing on invasive insect biology (Note S1), which represents a major concern in agriculture, as arthropods largely contribute to emerging alien species (Bradshaw et al., 2016; Seebens et al., 2021).

#### 2.1.1 | Genome and trait architecture

The simulated genome totaled 250 cM and consisted of five chromosomes of 50 cM each modeled as a segment of $5 \times 10^5$ sites with a per-site per-generation recombination rate of $10^{-5}$. Mutations affecting the phenotypes (QTNs) were simulated on the first four chromosomes at a rate of $2.5 \times 10^{-8}$ with effect sizes (in units of standard phenotypic deviations) independently sampled from a Gaussian distribution with a mean of 0 and a standard deviation of 0.1. Each QTN was characterized by two effect sizes contributing to two distinct phenotypes. For each phenotype, the effect sizes of all the QTNs present in an individual genome were summed to form its phenotypic value. The two phenotypes determined the individual fitness, which reached its maximum value of 1 when the two phenotypes matched the two environmental optima and decreased when the distance between the phenotypes and the environmental optima increased (Note S1). As fitness represents an individual probability of being selected as a parent for the next generation in the native area, this mechanism favored the spread of locally advantageous mutations. Note that given the distribution of QTNs effect sizes, a given phenotypic value between −1 and 1 could be obtained through multiple combinations of approximately 15 mutations per individual, implying a high degree of genetic redundancy.

In addition to the QTNs simulated directly during SLiM forward approach, neutral mutations were added afterward using the Python packages msprime (v1.2.0) and pyslim (v1.0.3) according to the *recapitation and overlay* procedure described by Haller et al. (2019), at a rate of $1.0 \times 10^{-7}$ per generation per site. Mutations with an MAF < 1% were filtered out using the program VCFtools (v 0.1.16) (Danecek et al., 2011).

#### 2.1.2 | Simulations design: Native area

Evolution in the native area was simulated under a 2D stepping stone model with selection, where slight modifications were made compared to Láruson et al. (2022) work (Note S1). The native area consisted of 25 populations of 1000 individuals each, organized in a 5 × 5 grid that evolved under a Wright–Fisher demographic model with nonoverlapping generations and constant population size. Populations exchanged individuals with either high (0.05) or low (0.005) migration rates, resulting in varying degrees of local adaptation. As mentioned above, the grid position was associated with two different values of environmental optima.

To investigate the impact of the adaptive landscape and its relation with neutral genetic structure, three different distributions of the environmental optima were considered (Figure 1). Simulations of the native area included 3000 SLiM generations, which were divided into the three following phases: (i) an initialization of 1000 generations without any environmental variation in order to generate the necessary adaptive genetic variation; (ii) 1000 generations with a gradual transition of the environment toward the specified optima value; and (iii) 1000 final generations

17524571, 2024, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/eva.13709 by Cochrane France, Wiley Online Library on [24/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

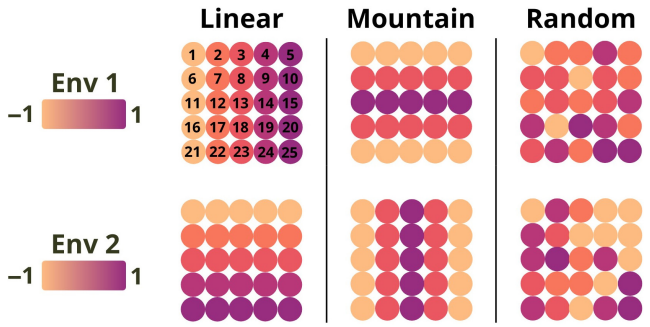


**FIGURE 1** Schematic representation of the three simulated native environment types. For each environment type (columns), the two grids represent the optimal values of environmental variables $e_1$ (top) and $e_2$ (bottom) for each of the 25 populations. Population indices are indicated on the top left panel and specify the source populations used for invasion (see Simulations design: invaded area).

under the adaptive landscape. This procedure and its duration of 3000 generations were established as sufficient to ensure robust alignment between populations and their environment (Note S1 and Figure S1).

Ten random replicates of the evolutionary history were simulated for each of the six native environment scenarios (3 environment types ×2 migration rates). "Mountain" (M) and "Random" (R) environments aimed to reduce environment–demography correlation seen in the "Linear" (L) environment.

The realized genome-wide $F_{ST}$ across all populations for each native area was computed with the *computeFST* function (*poolfstat* R package, version 2.1.1) with default settings.

### 2.1.3 | Simulations design: Invaded area

After 3000 simulated generations in the native area, a few founder individuals from a given source population were randomly chosen to invade a new environment. In order to test the effect of the founding bottleneck intensity, we considered either 10 or 100 founder individuals. For each native environment scenario, three distinct source populations were selected for invasion based on their native environmental optima $e_1$ and $e_2$: (i) $e_1 = e_2 = -1$ ("−1/−1"); (ii) $e_1 = e_2 = 0$ ("0/0"); or (iii) $e_1 = e_2 = 1$ ("1/1"). For example, in the case of the L scenario, these three source populations corresponded, respectively, to the populations 1, 13 and 25 (Figure 1). The source population could then invade nine possible environments, which correspond to the nine combinations of the three possible environmental values (−1, 0, or 1) for $e_1$ and $e_2$ in the invaded area.

Considering the distribution pattern of environmental values in the three simulated native area types (L, R, and M), it is important to note that multiple populations could correspond to any of the three potential source environments (e.g., population 1, 5, 21, and 25 all corresponding to the −1/−1 source population in the M environment, Figure 1). In this case, a single source population was arbitrarily selected to simulate invasion.

In the invaded area, we relied on the so-called non-Wright–Fisher simulation mode of SLiM (Messer, 2013) to allow for population extinction and estimate establishment probabilities. Reproduction and death events were disconnected in the invasive population, allowing in particular for overlapping generations. The fitness of an individual, derived from its QTN genotypes, quantified its probability of surviving to the next simulated time step (up to a maximum age of 3 time steps). Note that for simplicity, we then further assume that all living individuals in a given time step had the same probability of reproducing.

Each invasion was repeated 250 times, and the establishment probability (EP) in the new environment was determined by tallying successful establishment events among these repetitions. A population was deemed established if it exceeded 50,000 individuals or persisted for at least 100 time steps. These thresholds were selected based on preliminary tests, which showed that populations meeting these criteria never faced extinction (Note S1 and Figure S2).

## 2.2 | GO estimation

In this study, we considered two types of approach to estimate GO, differing in their underlying modeling of the relationship between genomic composition of the populations and their local environment. The first, named Gradient Forest and proposed by Fitzpatrick and Keller (2015) relies on a Random Forest machine learning algorithm, and the resulting GO estimate will be hereafter referred to as $GO_{gf}$. The second was proposed by Gain et al. (2023) and is based on a linear regression model to compute the geometric GO (gGO) defined as

$$\mathrm{gGO} = \frac{1}{n_{\mathrm{snp}}}(\boldsymbol{e} - \boldsymbol{e}^{\star})\boldsymbol{B}'\boldsymbol{B}(\boldsymbol{e} - \boldsymbol{e}^{\star})$$

where $n_{\mathrm{snp}}$ is the number of genotyped SNPs; $\boldsymbol{e}$ and $\boldsymbol{e}^{\star}$ are the vectors of $m$ environmental covariables values for the two compared environments; and $\boldsymbol{B}$ is the $n_{\mathrm{snp}} \times m$ matrix of the regression coefficients $\beta_{jk}$ of environmental variable $k$ on population allele frequencies at SNP $j$ as estimated under a GEA linear model. Three different methods were considered to compute gGO differing on how the $\beta_{jk}$ 's were estimated. First, we considered the original approach implemented in the *lfmm2* function of the R package *LEA* (v.3.12.2) (Gain & François, 2021), where the $\beta_{jk}$ were estimated under a Latent Factor Mixed Model (LFMM) (Caye et al., 2019). We hereafter refer to the resulting gGO estimator as $\mathrm{gGO}_{\mathrm{lfmm}}$. For all simulations, we modeled $K = 2$ latent factors since (as expected from the simulation design) these were found sufficient to capture the neutral population structure based on a Principal Component Analysis (PCA) of the genotypes matrix. Alternatively, we considered two estimations of gGO based on $\beta_{jk}$'s estimated under the BayPass Bayesian hierarchical GEA model (Gautier, 2015) which are both implemented in the newly developed R function compute_genetic_offset available from the latest

17524571, 2024, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/eva.13709 by Cochrane France, Wiley Online Library on [24/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

version (2.41) of the software package (Gautier, 2024). One important difference between BAYPASS compared to LFMM resides in the full modeling (and estimation) of the neutral covariance structure among the population allele frequencies through the so-called Ω matrix. First, the standard covariate model was run with default options to estimate Ω and the posterior mean of the regression coefficients $\beta_{jk}$'s based on an Importance Sampling (IS) algorithm. The resulting gGO estimator derived from the IS estimates of the $\beta_{jk}$'s is referred to as $gGO_{is}$. We also ran BAYPASS with option -covmcmc (setting Ω to the posterior mean estimated in the first run) to obtain estimates of the $\beta_{jk}$'s as the posterior means from values directly sampled via a Markov Chain Monte Carlo (MCMC) algorithm (instead of relying on the IS approximation). This gGO estimator will be hereafter referred to as $gGO_{mc}$. When considering a single environmental covariate (univariate case), estimates of the regression coefficients were found to be far more accurate with the MCMC than with the IS algorithm (Gautier, 2015). However, it is important to stress that when multiple covariates are analyzed (multivariate case), the MCMC algorithm considers all these covariates jointly (similar to LFMM), while the IS algorithm treats each covariate independently.

Note that other popular methods based on linear modeling to compute GO, namely RONA (Rellstab et al., 2016) or RDA-based (Capblancq & Forester, 2021), were not considered here since their properties and relationship with $gGO_{lfmm}$ have been explored in a recent study, and they were shown to be either strictly equivalent (for RDA-based under certain hypotheses) or less accurate (Gain et al., 2023). In this previous study, $GO_{gf}$ was also found to generally provide lower prediction accuracy than $gGO_{lfmm}$, at least under the simulation scenario the authors explored. Nevertheless, we kept it in our study for its ability to account for nonlinear relationships between allele frequencies and covariates. To estimate $GO_{gf}$, we used a customized version of the original *Gradient Forest* package (v.0.1.32) (Ellis et al., 2012) to allow efficient analyses of large number of SNPs (Note S2). Note that this optimized version is so far restricted to continuous covariables. To account for neutral population structure, we used as response variable in the GF modeling the residuals of the SNPs allele frequencies obtained after fitting a linear model with $K=2$ latent factors (Caye et al., 2019) as described below (see details in Note S2). Finally, following Gain et al. (2023), the resulting $GO_{gf}$ estimates were squared to ensure similar scaling than other distance measures.

We also computed the Euclidean distance between the environmental covariables as $\Delta_e = \frac{1}{n_e}(\boldsymbol{e}-\boldsymbol{e}^\star)'(\boldsymbol{e}-\boldsymbol{e}^\star)$ to establish a baseline prediction performance of GO measures that would only include the environmental information but not the genetic one. Indeed, as highlighted by the above gGO definition expression (but less clear with $GO_{gf}$), the GO may directly be interpreted as a weighted environmental distance, whose weights are related to the influence of environmental covariables on the structuring of genetic diversity (quantified by the $B$ matrix of regression coefficients in gGO). In other words, a covariable with no impact on genetic diversity (i.e., no SNPs with frequencies associated with its variation) would be highly penalized, and the GO is expected to be null if no SNP is found associated to any covariable. Like $GO_{gf}$, all Euclidean distances were squared.

## 2.3 | Evaluation of GO measures

### 2.3.1 | GO to predict establishment probability

To evaluate the predictive accuracy of the different GO measures for establishment probabilities, allele frequencies (for QTNs and neutral markers) in the native area were estimated from 50 randomly sampled individuals per population, keeping only SNPs with an MAF >1% to train the GEA models (via GF or linear modeling with LFMM or BAYPASS).

GO calculations included either all (neutral and QTN) the SNPs or the top 10% overly differentiated SNPs based on the $XtX^\star$ statistics estimated with BAYPASS (Gautier, 2015; Olazcuaga et al., 2020). Further, we either considered (i) the two causal environmental covariables alone (assuming an unrealistic situation where these would be known); (ii) eight covariables including these two causal plus six confounding variables; or (iii) the first four PCs obtained after performing a PCA of these eight covariables with the R package *ade4* (v1.7-22) (Dray & Dufour, 2007). The six confounding covariables consisted of two "fake" covariables without any link with the causal variables; two covariables correlated with causal environment 1 ($r=0.4$ and $r=0.8$); and two variables correlated with causal environment 2 ($r=0.4$ and $r=0.8$). The fake variables were randomly generated for each possible environment (25 in the native grid, and 9 potentially invaded environments) from a Gaussian distribution so that their correlations with the two causal covariables were close to 0 (ranging between −0.1 and 0.1).

The different GO estimators were then compared based on their Spearman's correlation $R^2$ with the logarithm of the establishment probabilities ($\log(p_e)$) obtained with simulations as detailed above; each $R^2$ value was based on a total of 90 observations, arising from the combination of nine possible invaded environments (for a given source population) and 10 replicates of the native environment scenario (Figure S3). In cases where GO was computed with confounding variables, each GO computation was performed for three distinct random draws of confounding environments and the mean $R^2$ value across these three confounding environments was reported. The associations between GO and population fitness, as well as between GO and population growth rate, were investigated using the same approach.

### 2.3.2 | Interpretation of the absolute value of gGO in terms of $f_2$

As demonstrated by Gain et al. (2023), gGO values calculated for new environments vary proportionally with fitness logarithms. However, the proportionality coefficient is challenging to compute making it difficult to accurately predict fitness values in a new environment

17524571, 2024, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/eva.13709 by Cochrane France, Wiley Online Library on [24/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

based on estimated gGO. Alternatively, gGO can be interpreted as the expected squared allelic frequency difference across all adaptive loci between two populations, aligning with the definition of the $f_2$ statistic (Patterson et al., 2012). This provides a way to evaluate the accuracy of the absolute value of different GO estimators based on simulated data. We thus compared estimated GO with $f_2$ for pairs of populations taken from different environments within the native area. The choice of native area populations ensured that allele frequencies were at equilibrium, as assumed in the theoretical prediction of Gain et al. (2023).

The gGO calculations were performed between the three potential source populations for invasion and the nine possible combinations of environmental values used to build invaded environments. We here only compared the $gGO_{lfmm}$ and $gGO_{mc}$ estimators since both rely on the same modeling approach consisting of treating all the covariables jointly. The $f_2$ statistics were estimated using the R package *poolfstat* (version 2.1.1) (Gautier et al., 2022), based on the allelic frequencies computed with the genotypes for all the 1000 individuals of a population. In all cases, we computed Mean Percentage Absolute Error (MAPE) to compare the estimated gGO to their corresponding estimated $f_2$ (expected to represent the truth).

Two different settings were investigated. We first evaluated the accuracy of gGO under ideal conditions, where the estimators were computed using all causal QTNs; the allele frequencies were computed based on the genotypes for all the 1000 simulated individuals of a population; and only the two causal environmental variables were used to compute $gGO_{lfmm}$ and $gGO_{mc}$. Second, we evaluated gGO estimates under more realistic conditions, where causal loci are unknown. gGO measures were then computed using QTNs and neutral SNPs, and population allele frequencies were obtained from 50 randomly sampled individuals (discarding all SNPs with an MAF < 1%). These more 'realistic' GO estimates were compared to estimated $f_2$ computed solely with QTNs or with both QTNs and neutral markers (filtered on MAF), derived from allele frequencies of all individuals within populations. Theoretically, gGO should reasonably predict the $f_2$ computed solely with QTNs, although a perfect match is not expected due to the exclusion of some QTNs with low MAF. On the other hand, the $f_2$ computed with both QTNs and neutral markers (MAF-filtered) should be seen as a higher bound that would be reached by a gGO estimation procedure failing to distinguish adaptive and neutral markers.

## 2.4 | *Bactrocera tryoni* case study

### 2.4.1 | Studied populations

We used publicly available data on 28 populations of *Bactrocera tryoni* including 15 native and 13 non-native populations (Figure S13), which were previously analyzed by Parvizi et al. (2023) and Popa-Báez et al. (2020). The dataset consisted of 6707 SNPs, which were obtained through Diversity Arrays Technology (DArT) sequencing data for 301 individuals (from 4 to 31 individuals per population).

### 2.4.2 | Environmental data

Environmental data were downloaded from the Chelsa (v2.1, accessed the March 27, 2023) database (Karger et al., 2017) using the *dismo* v 1.3.5 (Hijmans, Phillips, & Elith, 2023) and *raster* v 3.5.15 (Hijmans, Etten, et al., 2023) R packages. A total of 21 environmental covariables were extracted for each of the studied populations consisting of the averaged values over the period 1981–2010 (at a 30 arc sec resolution) for the 19 commonly used bioclimatic variables, the mean monthly climate moisture index, and the mean monthly-near surface wind speed. Indeed, previous works have shown the importance of humidity on fitness-related traits or geographic distribution in *B. tryoni* (Dominiak et al., 2006; Hulthen & Clarke, 2006; Sutherst & Yonow, 1998; Weldon & Taylor, 2010). While moisture's and humidity impact has been established, wind-related variables emerge as potentially influential factors for *B. tryoni*. Wind patterns might affect the presence of dew (Dominiak et al., 2006) and could also impact predation dynamics (Dominiak, 2012). To address variable interdependence, we carried out a PCA on all the covariates and retained the first five PCs (explaining 95% of the total variance) for GO calculation. To also adopt a similar method as that used by Parvizi et al. (2023) for managing variable correlation, we performed the identical analysis but with the six bioclimatic variables they selected (bio_3, bio_5, bio_8, bio_9, and bio_12), showing Pearson correlation coefficients inferior to 0.6.

### 2.4.3 | GO computation

Following the simulation study, we estimated $GO_{gf}$, $gGO_{lfmm}$, $gGO_{is}$, and $gGO_{mc}$ (see above). BayPass analysis was performed here from the allele count file already formatted by Parvizi et al. (2023). To ensure comparability across the different estimators in the context of real data including missing genotypes, $gGO_{lfmm}$ was estimated using the allele frequencies obtained from the allele count file, missing data (for three different SNPs in three different populations) being replaced by the mean of allele frequencies across the 28 populations. For the LFMM analysis (and $gGO_{lfmm}$ estimation), we included $K = 3$ latent factors following Parvizi et al. (2023). For the GF analysis, neutral population structure was accounted for by using the residuals of the LFMM analysis as input variable, similar to what was done in the simulation study (Note S2).

The different GO estimators were then computed between a "source" population and 12,838,400 positions encompassing an extensive area in Oceania, covering areas that have been invaded or are potentially at risk of invasion. The choice of the source population was based on the estimation of heterozygosities with *poolfstat* package (v.2.2.0). Among the native range, population

17524571, 2024, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/eva.13709 by Cochrane France, Wiley Online Library on [24/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

1 exhibited the highest heterozygosity (Figure S4). This aligns with *B. tryoni*'s historical records (Parvizi et al., 2023; Popa-Báez et al., 2020), so this population was selected as the source population to calculate GO.

## 3 | RESULTS

### 3.1 | GO to predict establishment probability

The primary objective of our simulation study was to assess the predictive capacity of several GO measures in estimating the establishment probability of invasive populations originating from diverse locations within a native area. This evaluation included locally adapted populations across three distinct types of native areas and considered variations in demographic parameters such as migration and the number of invading individuals (see Section 2 and Figure 1).

Between 1507 and 1568 QTNs were simulated under conditions of low migration, while between 1756 and 1955 QTNs were obtained for high migration scenarios, as detailed in Table 1. The majority of these QTNs had a low polymorphism level, with only between 44 and 64 retained with MAF > 1% for the low migration scenario, and between 80 and 109 for the high migration scenario. Mean fitness at the end of the 3000 simulated generations was lower for populations experiencing higher migration rates. Furthermore, two levels of local adaptation were generated, with low migration exhibiting stronger population differentiation, with $F_{ST}$ values ranging from 3.4% to 7.2%, in contrast to higher migration scenarios, where $F_{ST}$ values ranged from 0.34% to 0.46%.

We first focus on the results obtained for the low migration scenario (strongest population structuring) with 10 founding individuals in the invaded area. Following Gain et al. (2023), we compare Pearson's correlation of different GO estimates with $\log(p_e)$ in order to evaluate their ability to correctly rank the establishment probabilities of a population in different environments. Only the results based on all SNPs (without SNP pre-selection) are presented in the main text; they are depicted in Figure 2.

#### 3.1.1 | Superior performances of gGO measures when computed on causal variables

The $R^2$ values between $\log(p_e)$ and gGO remained consistently high when using only causal variables to compute gGO, with $R^2$ exceeding 0.75 in most cases. Overall, there was no noticeable difference between the performances of the three gGO estimators ($gGO_{lfmm}$, $gGO_{mc}$, and $gGO_{is}$). In contrast, $GO_{gf}$ exhibited less consistent performance than gGO and the Euclidean distance, with a $R^2$ value of only 0.25 in the worst case. Notably, $R^2$ values between $\log(p_e)$ and gGO using causal variables were similar to those between $\log(p_e)$ and Euclidean distance, as previously observed by Láruson et al. (2022). Indeed, one main interest of using GO measure is to weight variables according to their genetic importance. In the ideal scenario where only causal variables are employed for its computation, GO is not anticipated to demonstrate better performance than the Euclidean distance, as there is no need to discern which variables are related to adaptation. Despite their unrealistic nature, these results demonstrate the existence of a relationship between GO and $\log(p_e)$ under ideal conditions. Additionally, it is noteworthy that for the 0/0 source population, $R^2$ values tended to decrease in comparison to the other two source populations. This reduction can be attributed to the fact that these populations, with environmental values equal to 0, occupy a mid-range position within the spectrum of possible environmental values. This positioning results in less extreme GO values as well as less variable establishment probability values, making the ranking more challenging.

#### 3.1.2 | Robustness of GO measures to confounding covariables

When introducing additional confounding variables in the GO computation, differences between the different GO methods and the Euclidean distance became more apparent. $R^2$ values for all methods remained relatively high in the L environment, exceeding 0.75 for the −1/−1 and 1/1 source populations, and notably outperformed the Euclidean distance. However, for the M environment, $R^2$ values decreased for all GO methods, with the majority falling

**TABLE 1** Information about native area simulations.

| Migration rate | Environment type | Mean nb. Of QTNs | Mean nb. Of QTNs (MAF > 1%) | Mean nb. Of neutral mut. (MAF > 1%) | Mean fitness | Mean $F_{ST}$ (in %) |
|---|---|---|---|---|---|---|
| 0.005 | Linear | 1507 | 44 | 11,860 | 0.952 | 3.4 |
| | Mountain | 1568 | 64 | 11,781 | 0.933 | 4.9 |
| | Random | 1535 | 62 | 11,873 | 0.945 | 7.2 |
| 0.05 | Linear | 1756 | 82 | 11,548 | 0.883 | 0.34 |
| | Mountain | 1955 | 109 | 11,504 | 0.807 | 0.40 |
| | Random | 1852 | 80 | 11,509 | 0.815 | 0.46 |

*Note*: Means are computed over 10 replicates for each native area, and at the end of the 3000 simulated generations (see Table S1 for standard deviations that are all two to three orders of magnitude lower).

17524571, 2024, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/eva.13709 by Cochrane France, Wiley Online Library on [24/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

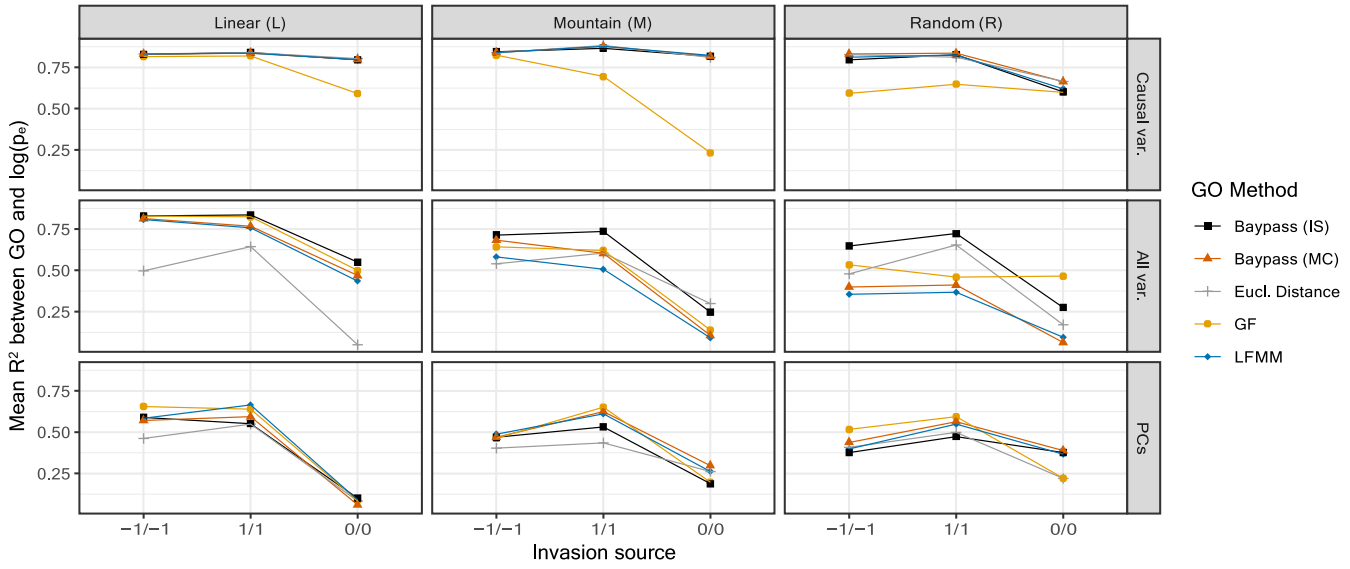


**FIGURE 2** Mean $R^2$ values between GO and $\log(p_e)$ (for the low migration rate and 10 invading individuals) for the different native environment types (L on the left; M in the middle; and R on the left) as a function of the covariables included in the computation of GO (two causal variables on top; eight covariables including two causal and six confounding covariables on center; and five PCs at bottom). Each panel represents the mean $R^2$ value over 90 observations for each of the three possible source population for invasion (−1/−1, 0/0, and 1/1); specified on the *x*-axis; and over 10 replicated simulations for the different GO estimators ($GO_{gf}$, $gGO_{lfmm}$, $gGO_{is}$, and $gGO_{mc}$, see the main text for details) alongside with Euclidean environmental distance. All SNPs were used for GO computation.

below 0.75. For the R environment, the results exhibited a more significant decline. $gGO_{lfmm}$ and $gGO_{mc}$ $R^2$ values dropped to around 0.5 and decreased to less than 0.125 in the case of the 0/0 source population, while Euclidean distance yielded higher $R^2$ values than these two gGO methods. Conversely, $GO_{gf}$, exhibited higher $R^2$ values compared to Euclidean distance for the 0/0 and −1/−1 source populations. Interestingly, the $gGO_{is}$ generally outperformed all other methods and maintained $R^2$ values close to 0.75 in all environments for the 1/1 and −1/−1 source populations (but see below).

### 3.1.3 | Using PCs reduce differences in method performances

When computing GO on PCs of (true and confounding) variables, the results showed similarities to those obtained using all variables, but the differences between methods were less pronounced and $R^2$ values were lower. In the case of the Linear (L) environment, the results were slightly less favorable than with all variables, ranging between 0.5 and 0.75 for all methods and the 1/1 and −1/−1 source populations, but dropping to less than 0.125 for the 0/0 source population. For the Mountain (M) and Random (R) environments, the reductions in $R^2$ values were less pronounced, with most $R^2$ values remaining relatively similar to those observed when using all variables. Moreover, employing PCs resulted in a narrower performance gap between $gGO_{is}$ and other gGO methods, occasionally leading to $gGO_{is}$ being outperformed by alternative methods. Conversely, when using PCs, Euclidean distance exhibited comparable or slightly worse performance compared to GO methods.

### 3.1.4 | Impact of covariables correlation on the gGO estimators

It might seem at first surprising that $gGO_{is}$ performed equally (when considering the two causal covariables only) or even better (when considering all eight covariables) than $gGO_{mc}$ and $gGO_{lfmm}$ since the underlying GEA models are similar and previous work showed that the IS estimation of regression coefficients was suboptimal compared to MCMC-based estimation (Gautier, 2015). However, the trend was less clear when considering PCs suggesting a possible negative impact of the correlation of covariables in the estimation of $gGO_{mc}$ and $gGO_{lfmm}$, both relying on a joint modeling of all the covariables while the IS estimation of regression coefficients (used to estimated $gGO_{is}$) amounts to treating all covariables separately. We thus hypothesized that the suboptimal performance of $gGO_{mc}$ and $gGO_{lfmm}$ when additional confounding variables were introduced, might be related to a poorer estimation of regression coefficients of correlated covariables. To explore this, we conducted a comparative analysis between BayPass IS, BayPass MC, and LFMM. However, this time, we estimated regression coefficients independently for each variable (i.e., running BayPass MC and LFMM separately for each covariable). Results for a low migration and 10 invading individuals are shown in Figure 3. Univariate calculation of regression coefficients resulted in clear improvements for $gGO_{lfmm}$ and $gGO_{mc}$, aligning them closely with $gGO_{is}$ in most instances, notably in the M and R environments. Although $gGO_{is}$ mostly maintained slightly higher $R^2$ values, particularly in the L environment, $gGO_{lfmm}$ occasionally outperformed it, as seen in the M and R environments for 0/0 source population.

17524571, 2024, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/eva.13709 by Cochrane France, Wiley Online Library on [24/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

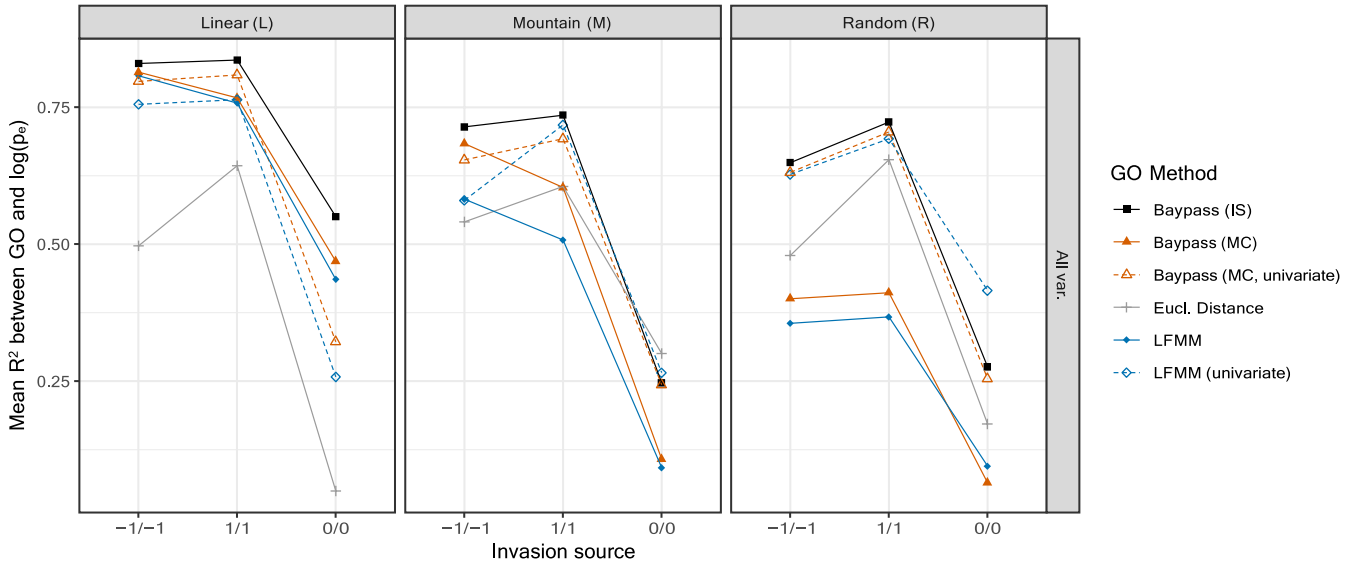


**FIGURE 3** Mean $R^2$ values between GO and $\log(p_e)$, for the low migration rate and 10 invading individuals, depending on the type of native environment (L, M, or R). Each panel represents the mean $R^2$ value (for each of the three possible source population for invasion, −1/−1, 0/0, and 1/1) between $\log(p_e)$ and GO obtained through five gGO computation methods, including $gGO_{lfmm}$ and $gGO_{mc}$ modified in order to treat variables independently (noted as "univariate"), alongside Euclidean distance. All SNPs and covariables were used for GO computation.

### 3.1.5 | Influence of SNPs pre-selection and simulation parameters

Overall, with the gGO method, the effect of SNP pre-selection on performance was not clear, as it did not consistently lead to better or worse $R^2$ values. However, when using $GO_{gf}$, a clear trend emerged where SNPs pre-selection consistently resulted in lower $R^2$ values compared to cases where SNPs are not pre-selected (Figures S5 and S6). Similar to the SNPs pre-selection, the migration rate had a small impact on the results. Whether the migration rate was low (Figure 2) or high (Figure S7), the results remained overall quantitatively and qualitatively comparable. Indeed, for high migration rate, substantial $R^2$ values were observed across all GO methods when exclusively employing causal variables, and a decline in $R^2$ values occurred upon the addition of confounding variables. As for the number of invading individuals, while a linear relationship between log() and GO is observed for 10 invading individuals, the same does not hold when the invading population size increases to 100 (see Figure S3 for an example). In the latter case, where EP values are consistently close to one, calculating $R^2$ values becomes less meaningful. That is why the $R^2$ results for the 100-individual scenario are not presented here. However, it is noteworthy that the results regarding the association between GO and fitness, as well as GO and population growth rate, exhibited comparable outcomes to the ones between GO and establishment probability (Figures S8 and S9), and that a robust relationship between GO and fitness or growth rate persisted even with 100 invading individuals, as depicted in Figure S10.

## 3.2 | Interpreting gGO absolute value in terms of $f_2$

To provide insights into the biological interpretation of gGO, we compared the accuracy of the different gGO estimators based on the theoretical expectations of Gain et al. (2023), who showed that the value of gGO between two locally adapted populations should be equal to the $f_2$ measured at causal QTNs for these two populations.

Using all QTNs and the two causal environmental variables yielded estimated gGO closely related to $f_2$ in the L environment, with small MAPE for both the $gGO_{lfmm}$ and $gGO_{mc}$ estimators across varying migration rates (Figure S11). For instance, for the highest migration rate, the $gGO_{lfmm}$ estimator deviates from the true $f_2$ by only 17%. Predictions were less accurate for the M environment, where $gGO_{lfmm}$ MAPE value reached 64% in the case of low migration, and for the R environment with MAPE ranging from 69% to 81% (Figure S11). $gGO_{mc}$ and $gGO_{lfmm}$ exhibited very similar results, with $gGO_{mc}$ appearing slightly more accurate in most cases.

In practice, calculating GO on all causal SNPs only is unrealistic notably because the driving covariables are usually not all included in the analysis and even in this case, no GEA method could be expected to classify perfectly (i.e., with a decision criterion leading to a power of 1 and a no false discoveries) all the underlying associated SNPs. The results of the comparison between a more "realistic" gGO computed using both causal and neutral SNPs (with an MAF filter) and the $f_2$ computed either with QTNs only or with both QTNs and neutral markers are presented in Figure 4. For improved readability and to reduce computational intensity associated with computing allele frequencies for all individuals and markers, we present results for only two environmental seeds and the L environment. Results for

17524571, 2024, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/eva.13709 by Cochrane France, Wiley Online Library on [24/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

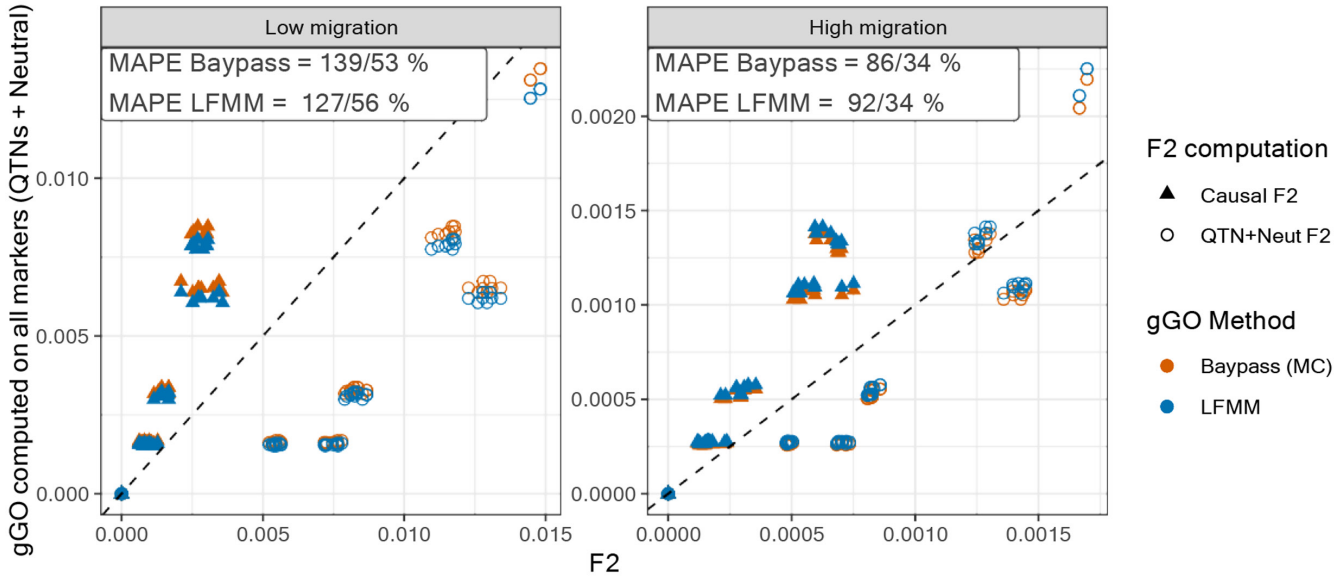


**FIGURE 4** Comparison between gGO and $f_2$ among QTNs (i.e., "ground truth" GO value) or among QTNs and neutral SNPs. The estimated values with the two gGO estimators ($gGO_{lfmm}$ and $gGO_{mc}$) were obtained using QTNs and neutral markers (with MAF > 0.01) for the scenarios with low (left panel) or high (right panel) migration within the native area under the L (linear) environment. The inset in each panel gives the two corresponding MAPEs separated by a slash.

the M and R environments are shown in Figure S12. These "realistic" gGO calculations consistently overestimated the $f_2$ calculated for QTNs alone, suggesting an imperfect estimation of the regression coefficients leading to an incorrect consideration of some neutral QTNs as adaptive. MAPE values were overall quite high, indicating limited accuracy in predicting true adaptive $f_2$ values. As already observed for relative establishment probabilities, absolute gGO values were the most accurate in the L environment and decreased in the M and R environments. On a more positive note, we observed that estimated gGOs were clearly lower than the $f_2$ computed with both QTNs and neutral markers, implying that these categories of SNPs could be at least partly distinguished.

### 3.3 | *B. tryoni* case study

GO computations conducted between the source population (population 1) and an extensive area in Oceania, employing PCs as predictor variables, demonstrated consistent outcomes across $gGO_{mc}$, $gGO_{lfmm}$, and $GO_{gf}$ (Figure 5).

Notably, all three methods identified outlying geographical zones in the North West area of New Zealand, the West part of Indonesia, and in the central region of Papua New Guinea, while also identifying New Zealand and Tasmania as being some of the areas having the highest GO values. Similarly, all methods exhibited the lowest GO in the North of Australia, Southern Papua New Guinea, South of Western Australia, and Southern Indonesia islands. Islands situated to the east of Australia (comprising Loyalty and Fiji Islands) exhibited medium GO. Some of the highest GO values in mainland Australia concentrated in the central part of Western Australia and the Eastern region shared by Northern Australia, Queensland, and South Australia. These "visual" similarities are confirmed by a high correlation of all GO estimators values, particularly between $gGO_{lfmm}$ and $gGO_{mc}$ (Figure S14). Euclidean distance showed similar results to all GO methods, notably showing high correlation level with $gGO_{lfmm}$ and $gGO_{mc}$ (Figure S14).

When using the six environmental covariables selected by Parvizi et al. (2023), results for multivariate methods closely resembled those obtained with PCs (Figure S15). Disparities between "univariate" and "multivariate" methodologies were also observed in agreement with the simulation study. More precisely, results obtained using univariate methods ($gGO_{lfmm}$ "univariate" and $gGO_{is}$) differed from those obtained using PCs, likely due to correlation between variables. Among multivariate methods, $gGO_{mc}$ and $gGO_{lfmm}$ were again very similar and overall identified the same low/high GO regions as those identified with PCs. $GO_{gf}$ differed from these two approaches, for example, not showing high GO regions in the central part of Western Australia and the Eastern region shared by Northern Australia, Queensland, and South Australia. Again, Euclidean distances gave comparable results to $gGO_{lfmm}$ and $gGO_{mc}$ but displayed lower correlations compared to those observed when using PCs (Figure S16).

## 4 | DISCUSSION

The purpose of this study was to conduct a comprehensive evaluation of GO measures, specifically focusing on their performances within the framework of biological invasions. The primary objective was to determine whether GO measures could effectively predict establishment probabilities through a simulation study. Furthermore, the application of GO measures to the biological invasion of *B. tryoni*

17524571, 2024, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/eva.13709 by Cochrane France, Wiley Online Library on [24/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

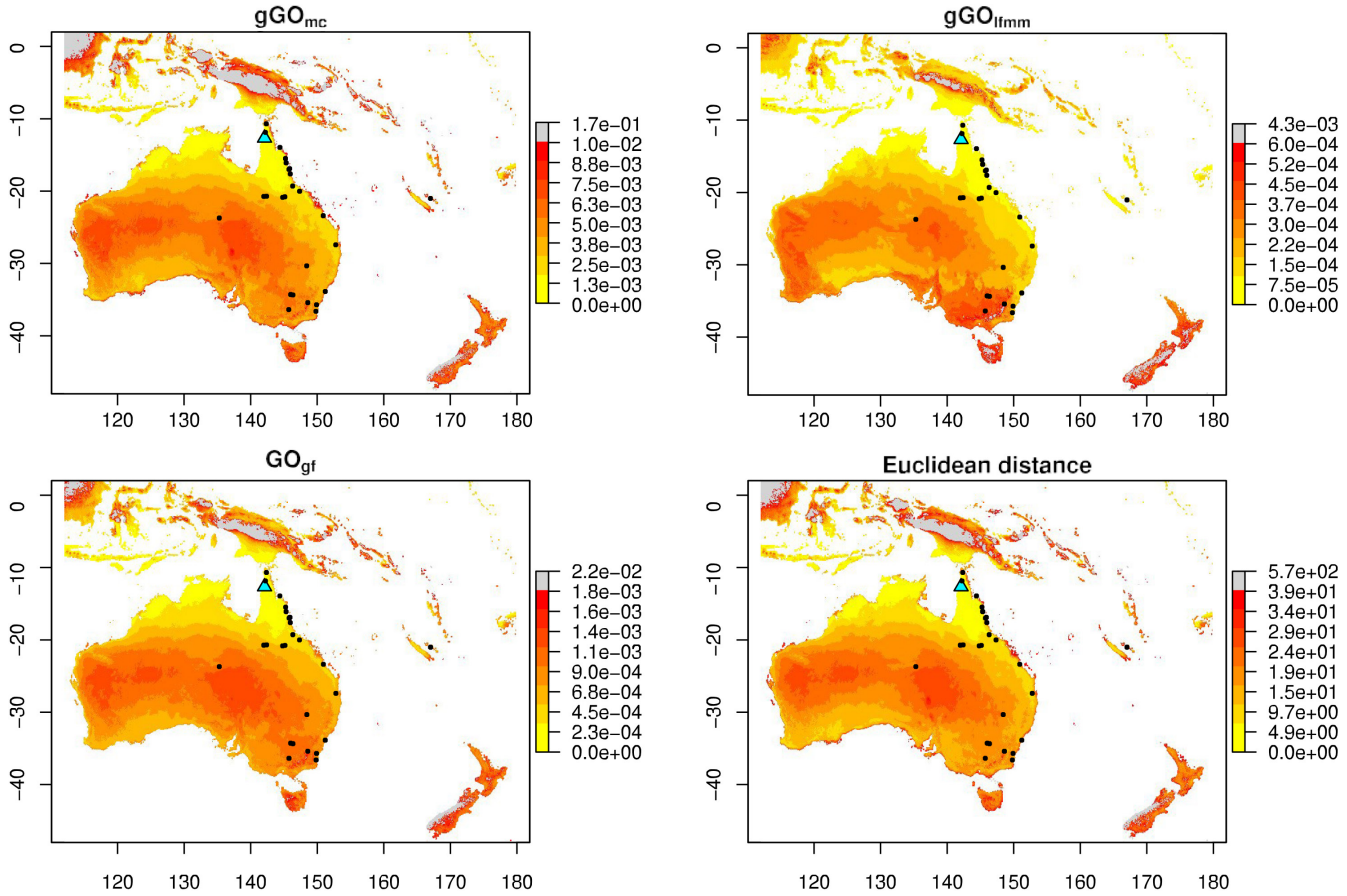


**FIGURE 5** Application of GO to *B. tryoni* populations. GO was estimated between population 1 ("source" population, identified with a blue triangle) and a large area in Oceania, with $gGO_{mc}$, $gGO_{lfmm}$, and $GO_{gf}$. Squared Euclidean distance to the source population is also displayed. Shades of yellow indicate lower GO values, while red shades higher GO values. Grey pixels represent outliers values, and black dots the studied populations.

served to illustrate their practical utility in this context. Finally, our study also sought to provide insights into the feasibility of interpreting the absolute values of gGO measures, while introducing methodological innovations through the computation of gGO with BAYPASS software and the optimization of Gradient Forest R package.

## 4.1 | Predicting relative establishment probabilities using GO

In general, at least one GO measure outperformed Euclidean distance in predicting EP, except when predicting EP using causal variables only, where all approaches performed similarly. These results were expected, as the unrealistic scenario where causal variables are known suppresses one of the main advantages of GO: weighting variables according to their genetic importance. The better performance of GO methods over Euclidean Distance when confounding variables are added underlines the effectiveness of using genetic information to improve the prediction of establishment probabilities, providing insights into the genetic make-up of the populations studied.

While our simulations indicate overall good performances of GO in predicting EP, these performances differed between native environments. The L environment, defined by gradual transitions between populations' environmental optima, tends to generate clearer relationships between allele frequencies and the environment. Conversely, the M environment contains geographically distant populations with similar environments (e.g., populations 1, 5, 21, and 25, Figure 1), and the R environment juxtaposes dissimilar environments. This complexity may lead to populations with different genetic compositions under similar environmental conditions due to low migration, or with similar genetic composition under different environments due to high migration, likely resulting in slightly poorer GO performances. Moreover, our simulations were characterized by high polygenicity and genetic redundancy, where numerous potentially pleiotropic QTNs were segregating and multiple combinations of these QTNs could lead to the same trait values. The interplay between this high polygenicity and the complex patterns of local adaptation in the M and R environments likely explains the poorer performances of GO methods in these scenarios. These results are in line with those of Lotterhos (2023), illustrating that high polygeny, genotypic redundancy, and pleiotropy can result in non-monotic

17524571, 2024, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/eva.13709 by Cochrane France, Wiley Online Library on [24/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

patterns between allele frequencies and environmental variables. Such patterns challenge GEA methods, which rely on the assumption of clinal patterns between allele frequencies and environmental variables.

Overall, gGO methods outperformed GF in predicting EP, especially when each variable was considered individually in gGO calculations. This is in line with the findings of Gain et al. (2023), who focused on the relationship between GO and fitness. In the ideal case with only causal variables, $GO_{gf}$ often showed lower performance than Euclidean distance, indicating a relatively limited ability to decipher the allele frequency–environment relationship. However, the performance gap between gGO and GF decreased when confounding variables or PCs were included in GO calculations. In scenarios where nonlinear relationships exist between allele frequencies and environment, $GO_{gf}$ may in theory outperform gGO due to its ability to accommodate such relationships. While this could not be clearly observed in the more complex M and R environments of our simulations, we note that GF is a machine learning-based method whose performance is likely more dependent on the amount of data; thus, we cannot rule out that our dataset of 25 populations may have been insufficient to make accurate predictions with this approach. As underlined by Gain et al. (2023), a linear model may also achieve a better bias-variance trade-off than a nonlinear machine learning model.

Among gGO methods, $gGO_{mc}$ and $gGO_{lfmm}$ yielded similar results, which was expected as these methods are based on the same principle. In practice, although LFMM is more computationally efficient, the Bayesian hierarchical framework underlying BayPass allows to accommodate and properly account for the specificities of non-standard (e.g., Pool-Seq data) or heterogeneous datasets (Camus et al., in prep), thereby opening new ways to apply gGO in some biological contexts. Likewise, promising directions would be to directly incorporate estimation of the matrix $\boldsymbol{B}$ of the SNP environmental effects (regression coefficients) in the model to allow accounting for correlation among covariables (if not using PCs) and to provide estimates of uncertainty (e.g., credibility interval) associated with the estimated GO.

In the presence of confounding variables, a noteworthy finding was that the prediction power of gGO methods was reduced only if regression coefficients were estimated jointly (i.e., the multivariate approach), but not if these coefficients were estimated independently for each variable (i.e., the univariate approach). The relatively stable performance of univariate methods can be related to the theoretical result of Gain et al. (2023), stating that a gGO computed from linear combinations of causal (and potentially non-causal) variables should be equivalent to the gGO computed with causal predictors only. Indeed, the variance–covariance matrix of regression coefficients that is used in gGO allows mitigating the redundancy of information resulting from the inclusion of both causal and correlated variables, while removing the noise arising from fake uncorrelated variables. However, a strong assumption underlying these expectations is that regression coefficients of observed variables are correctly estimated. This might explain the lower performance of multivariate approaches in our simulations, because the joint estimation of regression coefficients from a set of correlated variables is typically more challenging.

While these results suggest to always favor univariate methods to ensure accurate coefficient estimates, note that their prediction accuracy is certainly boosted by the inclusion of true causal variables in gGO computations, which is very unlikely in real-life studies. In comparison, simulation results based on PCs of environmental variables might better reflect the practical performance of GO methods. Prediction accuracy decreased for all methods in this case, and the differences between methods were also reduced. This result was expected since PCs potentially cause a loss of information about causal variables but avoid estimation issues by removing correlations between variables. Nevertheless, PC-based GOs maintained an $R^2$ value close to 0.5 (for −1/−1 and 1/1 source populations), which can be considered an acceptable predictive performance.

We therefore suggest caution when using multivariate gGO methods and recommend either the use of PCs to suppress variable correlations or the use of univariate gGO for untransformed variables, even though the behavior of this latter approach with real data featuring many highly correlated variables and possibly no causal ones remains uncertain. A limitation associated with the use of PCs lies in the interpretation of variable importance, a facet of interest in GO approaches, because GEA models then report the importance of PCs, which have no clear biological interpretation. To overcome this issue, note that obtaining the importance of original variables from those of the PCs is actually straightforward (Note S3). Finally, the impact of pre-selecting markers (e.g., based on XtX*) on gGO performance was found to be negligible, while it could compromise GF's performance, affirming previous findings of Gain et al. (2023). Therefore, we propose that pre-selecting markers is not an obligatory step for achieving a robust GO interpretation.

Despite the overall good performances of GO methods to predict EP, the scenarios considering the invasion of individuals originating from the 0/0 source populations highlighted some more conceptual limitations of GO-based EP prediction. Indeed, EP prediction was more difficult for 0/0 source populations, because their mid-range environmental values imply a relatively low adaptive challenge whatever the invaded environment. For instance, in the L environment with a low migration rate and 10 invading individuals, EP was approximately equal to one-third in the most extremes 1/1 or −1/−1 environments (Figure S17). Considering causal variables only, a strong correlation between GO and EP was found despite the quite similar values associated to the different invaded environments (Figure S17). However, the additional noise resulting from the inclusion of confounding variables was sufficient to affect the ranking of GO values among invaded environments, which lead to a strong decrease of $R^2$ values. This illustrates the problem of the lack of interpretability of absolute GO values. In such instances, a GO analysis should ideally conclude that *all* environments present a high invasion risk, not only that 0/1 environments are more at risk than 1/1 environments (for instance). In other words, interpreting absolute GO values would be crucial to determine whether the variations

17524571, 2024, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/eva.13709 by Cochrane France, Wiley Online Library on [24/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

of GO computed in distinct environments imply distinct or similar challenges for adaptation. This understanding can significantly impact species management in real-life scenarios, whether for invasive or endangered species.

Unfortunately, our attempt to evaluate the interpretability of absolute gGO values outlined the difficulty of this task, with gGO values often diverging from their expected $f_2$ values. This may be due to inaccurate estimation of regression coefficients, stemming from a variety of factors such as the small number of observed populations, the existence of non-monotonic clines between allele frequencies, and covariates or imperfect correction of population structure. Deviations from the conditions where gGO is expected to equal $f_2$ (namely the infinitesimal model) may also contribute to this imperfect match. Nonetheless, our findings are promising as they reveal a strong proximity between gGO and $f_2$ within the linear environment, under ideal conditions where only QTNs were considered. Moreover, a strong correlation between gGO and $f_2$ persists across all environments, even in more realistic conditions where gGO was computed from all SNPs. In these conditions, gGO also exhibited some expected behavior as it overestimated the QTNs' $f_2$ (likely due to erroneously attributing weight to neutral SNPs) while remaining below the overall $f_2$ (i.e., computed based on both QTNs and neutral SNPs) thus correctly excluding (or down-weighting) most neutral SNPs.

## 4.2 | Toward a more comprehensive modeling of population fitness

Our simulation framework demonstrated a strong correlation between GO and EP across various environment types and migration rates, even when considering confounding variables. This suggests, among others, that GO is resilient in scenarios with moderate population differentiation and imperfect adaptation, such as those with high migration. However, several open questions need to be addressed before applying GO to predict EP beyond idealized simulation frameworks.

All GO methods assume that populations adapt to new environments through pre-existing variants, so their ability to predict fitness and thus EP is expected to decrease if adaptation actually proceeds, at least partly, from de novo mutations. However, the simulations conducted in this study do not allow quantifying this effect, because their design implies that adaptation is mainly driven by standing variation. Indeed, the polygenic traits' architecture and the environment heterogeneity create high levels of standing genetic variation in the native area (Höllinger et al., 2019; Yeaman, 2022), facilitating rapid adaptation in invaded environments (Jain & Stephan, 2017). In addition, the relatively low mutation rates and the small founding population sizes make adaptation through de novo mutations very unlikely in the invaded area, at least in the short evolution time considered here. While a growing body of literature supports the idea that rapid adaptation to environmental change often results from standing genetic variation (Barrett & Schluter, 2008; Bitter et al., 2019; Chaturvedi et al., 2021) and deems it important for invasive species to adapt to invaded areas (Bock et al., 2015; Prentis et al., 2008), at least one reported case indicates adaptation through potential new mutations during colonization (Exposito-Alonso et al., 2018). Additionally, some adaptations to traits relevant to invasion biology, such as insecticide resistance in crop pests, are thought to result from the interplay between standing variation and de novo mutations (Hawkins et al., 2019). This emphasizes the importance of considering these processes when using GO methods.

Additionally, when relying on GO to anticipate biological invasions, it is implicitly assumed that population pre-adaptation plays a significant role in the successful establishment in a new environment. However, scenarios with 100 invading individuals illustrate that it might not always be the case: while GO maintained a strong correlation with fitness (Figure S10), accurate prediction of EP became challenging due to the substantial number of invading individuals buffering the adaptive challenges presented by the new environment. While meta-analyses have shown that invasive species often maintain their ecological niche in the invaded area (Aravind et al., 2022; Bomford et al., 2009; Liu et al., 2020), supporting the hypothesis that pre-adaptation plays a significant role in invasion success, successful population establishment can be influenced by various factors, including propagule pressure (Simberloff, 2009; Wittmann et al., 2014), hybridization/admixture (Barker et al., 2019; Rius & Darling, 2014), and epigenetic processes (Marin et al., 2020; Mounger et al., 2021). While our simulations partially explored the effects of propagule pressure by varying the number of invading individuals, the influence of successive introductions was not examined. Incorporating some of the above factors into simulations could refine the conditions under which GO can effectively predict EP in more realistic applications.

Our simulations also did not explicitly include recessive deleterious mutations, since QTNs could be either deleterious or beneficial depending on the environment and their genetic background always affects the phenotype. This hinders our ability to study the influence of genetic load on invasion success. However, biological invasions are generally characterized by initial bottlenecks favoring drift and potential inbreeding, which can reduce population fitness due to the fixation and/or expression of strongly deleterious mutations. Nevertheless, some studies have shown evidence of genetic load purging in invasive species (Facon et al., 2011; Marchini et al., 2016; Mullarkey et al., 2013; Tayeh et al., 2013), and gene flow and/or admixture can also mask genetic load (Whiteley et al., 2015). Furthermore, cases exist where populations, despite showing evidence of high genetic load without a clear purging signal, have successfully established and persisted in new environments (Gautier et al., 2023; Zayed et al., 2007). Demographic parameters and stochastic factors can influence genetic load and therefore population persistence in contrasting ways (reviewed in Robinson et al. (2023) and Bouzat (2010)). Further work is thus needed to explore the nuanced dynamics of invasion success in the presence of genetic load, as GO alone does not account for its effects.

17524571, 2024, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/eva.13709 by Cochrane France, Wiley Online Library on [24/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

### 4.3 | Insights from the *B. tryoni* case study

Despite the limits mentioned above, practical use of GO in the case of *B. tryoni* appeared insightful. The results obtained with GO, especially with $gGO_{lfmm}$ and $gGO_{mc}$, are consistent with the existing knowledge of the species' establishment. Consistently, the lowest GO values were obtained within the native range of *B. tryoni*, this area being not expected to pose any adaptive challenge for the chosen reference population. It is also anticipated that regions with a higher GO may experience invasion at a later stage compared to those with lower GO, given the expectation of a higher adaptive challenge for population establishment. Interestingly, the medium GO estimated in the Loyalty Islands aligns with the early stages of expansion, as the first documented records in these islands date back to around 1969 (Popa-Báez et al., 2020). The higher, but still relatively low, GO values observed around population 26, populations 16 and 17, and the east coast of Australia, are consistent with a later establishment—circa 1987 for population 26 (Cameron, 2006) and approximately 1994 for populations 16 and 17, as well as the East Coast of Australia (Osborne et al., 1997; Popa-Báez et al., 2020).

The lower GO in areas not yet colonized (southern Papua New Guinea, southern Western Australia, and southern Indonesian islands) suggest that these regions may be vulnerable to establishment of individuals from the native area (population 1). Given its geographical proximity to the native range of *B. tryoni*, Papua New Guinea is at a higher risk of invasion. Regions with high GO in mainland Australia (for $gGO_{mc}$ and $gGO_{lfmm}$) coincide with areas where *B. tryoni* is not established. Nevertheless, in Western Australia, a few incursions have been documented, but eradication measures are promptly initiated upon the detection of more than five individuals, likely preventing the establishment of any population (Dominiak & Mapson, 2017). More broadly, it should be kept in mind that biocontrol strategies, which are frequently implemented (Maelzer et al., 2004), might have substantially impeded the establishment of *B. tryoni*, whether populations were pre-adapted or not.

High GO regions beyond mainland Australia also align with our knowledge of *B. tryoni*, as none of these areas exhibit established populations. However, Popa-Báez et al. (2021) have demonstrated that occasional incursions into New Zealand and Tasmania likely originated from New Caledonia or the east coast of Australia. Based on GO values, we could infer that the invasion risk stemming from Northern *B. tryoni* in these areas is low, but it is important to note that our analysis did not assess the invasion risk from other populations.

It is also important to highlight that, in this specific case study, incorporating genetic information to predict *B. tryoni* invasion risk only marginally affected the conclusions that would have been drawn by simply considering Euclidean distance between environmental covariables. This may reflect a limited genetic basis for adaptation in this species and/or a lack of genomic (number of SNPs) and environmental (available covariates) information captured by the specific dataset used to evaluate GO. Besides, employing PCs may lead to a loss of information regarding the association between genomic data and covariates, especially in real-life scenarios where causal variables are lacking. This could explain the higher correlation levels observed between GO methods and Euclidean distance that we observed when using PCs, compared to the less pronounced correlations when considering the set of ascertained untransformed covariates.

More generally, this case study effectively demonstrates a relatively straightforward but practical application of GO in a real biological invasion context. The identification of areas at higher risk of invasion can greatly benefit invasive species management, by informing intensified surveillance efforts in higher-risk areas, leading to early detection and thus enhancing the chance of eradication (Reaser et al., 2020). Moreover, the broader applicability of GO can extend to predicting and preventing biological invasions under climate change scenarios.

### 4.4 | General conclusions

Our study confirms the relevance of GO to predicting invasion success and provides several methodological tools and advice to enhance the performance of this approach. Regarding empirical application, we illustrate how GO measures can be utilized to provide recommendations for invasion risk. Further theoretical research is needed to determine the impact of several key factors of invasion success, such as propagule pressure and genetic load, on the accuracy of predicting establishment probabilities with GO. Additional research on other species and how to integrate other genomic tools or SDM to provide a more comprehensive evaluation of invasion risk will also be necessary in the future.

#### ACKNOWLEDGMENTS

We wish to thank Olivier François, Arnaud Estoup, Bénédicte Rhoné, Renaud Vitalis, and Joëlle Ronfort for helpful insights and discussion. Louise Camus was funded by the Occitanie Region (France), and the INRAE Scientific Department SPE. We wish to thank our two anonymous reviewers for their very helpful and constructive comments.

#### CONFLICT OF INTEREST STATEMENT

The authors have no conflict of interest to disclose.

#### DATA AVAILABILITY STATEMENT

The *B. tryoni* data have already been published and have permissions appropriate for fully public release. The code necessary to reproduce the simulations, compute simulations GO, and use the optimized version of Gradient Forest package is available at https://forgemia.inra.fr/simon.boitard/popgenomicprediction.

#### ORCID

*Louise Camus* https://orcid.org/0000-0001-5485-8078

#### REFERENCES

Adam, A. A. S., Thomas, L., Underwood, J., Gilmour, J., & Richards, Z. T. (2022). Population connectivity and genetic offset in the spawning

17524571, 2024, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/eva.13709 by Cochrane France, Wiley Online Library on [24/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

coral *Acropora digitifera* in Western Australia. *Molecular Ecology*, *31*(13), 3533–3547.

Aravind, N., Shaanker, M., Bhat, P., Charles, B., Uma Shaanker, R., Shah, M., & Ravikanth, G. (2022). Niche shift in invasive species: Is it a case of "home away from home" or finding a "new home"? *Biodiversity and Conservation*, *31*, 1–14.

Archambeau, J. (2022). *Understanding the origin and predicting adaptive genetic variation at large scale in the genomic era: A case study in maritime pine*. These de doctorat, Bordeaux.

Barker, B. S., Cocio, J. E., Anderson, S. R., Braasch, J. E., Cang, F. A., Gillette, H. D., & Dlugosch, K. M. (2019). Potential limits to the benefits of admixture during biological invasion. *Molecular Ecology*, *28*(1), 100–113.

Barrett, R. D. H., & Schluter, D. (2008). Adaptation from standing genetic variation. *Trends in Ecology & Evolution*, *23*(1), 38–44.

Bitter, M. C., Kapsenberg, L., Gattuso, J.-P., & Pfister, C. A. (2019). Standing genetic variation fuels rapid adaptation to ocean acidification. *Nature Communications*, *10*(1), 5821.

Bock, D. G., Caseys, C., Cousens, R. D., Hahn, M. A., Heredia, S. M., Hübner, S., Turner, K. G., Whitney, K. D., & Rieseberg, L. H. (2015). What we still don't know about invasion genetics. *Molecular Ecology*, *24*(9), 2277–2297.

Bogaerts-Márquez, M., Guirao-Rico, S., Gautier, M., & González, J. (2021). Temperature, rainfall and wind variables underlie environmental adaptation in natural populations of Drosophila melanogaster. *Molecular Ecology*, *30*(4), 938–954.

Bomford, M., Kraus, F., Barry, S., & Lawrence, E. (2009). Predicting establishment success for alien reptiles and amphibians: A role for climate matching. *Biological Invasions*, *11*, 713–724.

Borrell, J. S., Zohren, J., Nichols, R. A., & Buggs, R. J. A. (2020). Genomic assessment of local adaptation in dwarf birch to inform assisted gene flow. *Evolutionary Applications*, *13*(1), 161–175.

Bouzat, J. L. (2010). Conservation genetics of population bottlenecks: The role of chance, selection, and history. *Conservation Genetics*, *11*(2), 463–478.

Bradshaw, C. J. A., Leroy, B., Bellard, C., Roiz, D., Albert, C., Fournier, A., Barbet-Massin, M., Salles, J.-M., Simard, F., & Courchamp, F. (2016). Massive yet grossly underestimated global costs of invasive insects. *Nature Communications*, *7*(1), 12986.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Bruce, T. J. A. (2010). Tackling the threat to food security caused by crop pests in the new millennium. *Food Security*, *2*(2), 133–141.

Cameron, E. C. (2006). *Fruit fly pests of northwestern Australia*. Ph.D. thesis, Faculty of Science, School of Biological Sciences, The University of Sydney.

Capblancq, T., Fitzpatrick, M. C., Bay, R. A., Exposito-Alonso, M., & Keller, S. R. (2020). Genomic prediction of (mal)adaptation across current and future climatic landscapes. *Annual Review of Ecology, Evolution, and Systematics*, *51*(1), 245–269.

Capblancq, T., & Forester, B. R. (2021). Redundancy analysis: A swiss Army knife for landscape genomics. *Methods in Ecology and Evolution*, *12*(12), 2298–2309.

Capblancq, T., Morin, X., Gueguen, M., Renaud, J., Lobreaux, S., & Bazin, E. (2020). Climate-associated genetic variation in *Fagus sylvatica* and potential responses to climate change in the French Alps. *Journal of Evolutionary Biology*, *33*(6), 783–796.

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

Chaturvedi, A., Zhou, J., Raeymaekers, J. A. M., Czypionka, T., Orsini, L., Jackson, C. E., Spanier, K. I., Shaw, J. R., Colbourne, J. K., & De Meester, L. (2021). Extensive standing genetic variation from a small number of founders enables rapid adaptation in Daphnia. *Nature Communications*, *12*(1), 4306.

Chen, Y., Gao, Y., Huang, X., Li, S., Zhang, Z., & Zhan, A. (2023). Incorporating adaptive genomic variation into predictive models for invasion risk assessment. *Environmental Science and Ecotechnology*, *18*, 100299.

Chen, Y., Liu, Z., Régnière, J., Vasseur, L., Lin, J., Huang, S., Ke, F., Chen, S., Li, J., Huang, J., Gurr, G. M., You, M., & You, S. (2021). Large-scale genome-wide study reveals climate adaptive variability in a cosmopolitan pest. *Nature Communications*, *12*(1), 7206.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158.

Dominiak, B., Mavi, H., & Nicol, H. (2006). Effect of town microclimate on the Queensland fruit fly *Bactrocera tryoni*. *Australian Journal of Experimental Agriculture*, *46*, 1239.

Dominiak, B. C. (2012). Review of dispersal, survival, and establishment of *Bactrocera tryoni* (Diptera: Tephritidae) for quarantine purposes. *Annals of the Entomological Society of America*, *105*(3), 434–446.

Dominiak, B. C., & Mapson, R. (2017). Revised distribution of *Bactrocera tryoni* in eastern Australia and effect on possible incursions of Mediterranean fruit fly: Development of Australia's Eastern Trading Block. *Journal of Economic Entomology*, *110*(6), 2459–2465.

Dray, S., & Dufour, A.-B. (2007). The ade4 Package: Implementing the duality diagram for ecologists. *Journal of Statistical Software*, *22*(4), 1–20.

Ellis, N., Smith, S. J., & Pitcher, C. R. (2012). Gradient forests: Calculating importance gradients on physical predictors. *Ecology*, *93*(1), 156–168.

Exposito-Alonso, M., Becker, C., Schuenemann, V. J., Reiter, E., Setzer, C., Slovak, R., Brachi, B., Hagmann, J., Grimm, D. G., Chen, J., Busch, W., Bergelson, J., Ness, R. W., Krause, J., Burbano, H. A., & Weigel, D. (2018). The rate and potential relevance of new mutations in a colonizing plant lineage. *PLoS Genetics*, *14*(2), e1007155.

Facon, B., Hufbauer, R. A., Tayeh, A., Loiseau, A., Lombaert, E., Vitalis, R., Guillemaud, T., Lundgren, J. G., & Estoup, A. (2011). Inbreeding depression is purged in the invasive insect *Harmonia axyridis*. *Current Biology*, *21*(5), 424–427.

Ferrier, S., Manion, G., Elith, J., & Richardson, K. (2007). Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. *Diversity and Distributions*, *13*(3), 252–264.

Fitzpatrick, M. C., Chhatre, V. E., Soolanayakanahally, R. Y., & Keller, S. R. (2021). Experimental support for genomic prediction of climate maladaptation using the machine learning approach Gradient Forests. *Molecular Ecology Resources*, *21*(8), 2749–2765.

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.

Frichot, E., Schoville, S. D., Bouchard, G., & François, O. (2013). Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular Biology and Evolution*, *30*(7), 1687–1699.

Gain, C., & François, O. (2021). LEA 3: Factor models in population genetics and ecological genomics with R. *Molecular Ecology Resources*, *21*(8), 2738–2748.

Gain, C., Rhoné, B., Cubry, P., Salazar, I., Forbes, F., Vigouroux, Y., Jay, F., & François, O. (2023). A quantitative theory for genomic offset statistics. *Molecular Biology and Evolution*, *40*(6), msad140.

Gallien, L., Münkemüller, T., Albert, C. H., Boulangeat, I., & Thuiller, W. (2010). Predicting potential distributions of invasive species: Where to go from here? *Diversity and Distributions*, *16*(3), 331–342.

Gautier, M. (2015). Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics*, *201*(4), 1555–1579.

17524571, 2024, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/eva.13709 by Cochrane France, Wiley Online Library on [24/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Gautier, M. (2024). *BayPass user manual*. https://forgemia.inra.fr/mathieu.gautier/baypass_public

Gautier, M., Micol, T., Camus, L., Moazami-Goudarzi, K., Naves, M., Guéret, E., Engelen, S., Colas, F., Flori, L., & Druet, T. (2023). Genomic reconstruction of the successful establishment of a feralized bovine population on the Subantarctic Island of Amsterdam. *bioRxiv*. https://doi.org/10.1101/2023.11.24.568563

Gautier, M., Vitalis, R., Flori, L., & Estoup, A. (2022). f-Statistics estimation and admixture graph construction with Pool-Seq or allele count data using the R package poolfstat. *Molecular Ecology Resources*, *22*(4), 1394–1416.

Haller, B. C., Galloway, J., Kelleher, J., Messer, P. W., & Ralph, P. L. (2019). Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. *Molecular Ecology Resources*, *19*(2), 552–566.

Hawkins, N. J., Bass, C., Dixon, A., & Neve, P. (2019). The evolutionary origins of pesticide resistance. *Biological Reviews*, *94*(1), 135–155.

Hijmans, R. J., Etten, J. V., Sumner, M., Cheng, J., Baston, D., Bevan, A., Bivand, R., Busetto, L., Canty, M., Fasoli, B., Forrest, D., Ghosh, A., Golicher, D., Gray, J., Greenberg, J. A., Hiemstra, P., Hingee, K., Ilich, A., Geosciences, I. F. M. A., … Wueest, R. (2023). *raster: Geographic data analysis and modeling*.

Hijmans, R. J., Phillips, S., & Elith, J. L. (2023). *dismo: Species distribution modeling*.

Höllinger, I., Pennings, P. S., & Hermisson, J. (2019). Polygenic adaptation: From sweeps to subtle frequency shifts. *PLoS Genetics*, *15*(3), e1008035.

Hulme, P. E. (2017). Climate change and biological invasions: Evidence, expectations, and response options. *Biological Reviews of the Cambridge Philosophical Society*, *92*(3), 1297–1313.

Hulme, P. E. (2021). Unwelcome exchange: International trade as a direct and indirect driver of biological invasions worldwide. *One Earth*, *4*(5), 666–679.

Hulthen, A. D., & Clarke, A. R. (2006). The influence of soil type and moisture on pupal survival of *Bactrocera tryoni* (Froggatt) (Diptera: Tephritidae). *Australian Journal of Entomology*, *45*(1), 16–19.

Ingvarsson, P. K., & Bernhardsson, C. (2020). Genome-wide signatures of environmental adaptation in European aspen (*Populus tremula*) under current and future climate conditions. *Evolutionary Applications*, *13*(1), 132–142.

Jain, K., & Stephan, W. (2017). Rapid adaptation of a polygenic trait after a sudden environmental shift. *Genetics*, *206*(1), 389–406.

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Zimmermann, N. E., Linder, H. P., & Kessler, M. (2017). Climatologies at high resolution for the earth's land surface areas. *Scientific Data*, *4*(1), 170122.

Lachmuth, S., Capblancq, T., Keller, S. R., & Fitzpatrick, M. C. (2023). Assessing uncertainty in genomic offset forecasts from landscape genomic models (and implications for restoration and assisted migration). *Frontiers in Ecology and Evolution*, *11*, 1155783.

Láruson, A. J., Fitzpatrick, M. C., Keller, S. R., Haller, B. C., & Lotterhos, K. E. (2022). Seeing the forest for the trees: Assessing genetic offset predictions from gradient forest. *Evolutionary Applications*, *15*(3), 403–416.

Liu, C., Wolter, C., Xian, W., & Jeschke, J. M. (2020). Most invasive species largely conserve their climatic niche. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(38), 23643–23651.

Lotterhos, K. E. (2023). The paradox of adaptive trait clines with nonclinal patterns in the underlying genes. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(12), e2220313120.

Maelzer, D. A., Bailey, P. T., & Perepelicia, N. (2004). Factors supporting the non-persistence of fruit fly populations in South Australia. *Australian Journal of Experimental Agriculture*, *44*(1), 109.

Marchini, G. L., Sherlock, N. C., Ramakrishnan, A. P., Rosenthal, D. M., & Cruzan, M. B. (2016). Rapid purging of genetic load in a metapopulation and consequences for range expansion in an invasive plant. *Biological Invasions*, *18*(1), 183–196.

Marin, P., Genitoni, J., Barloy, D., Maury, S., Gibert, P., Ghalambor, C. K., & Vieira, C. (2020). Biological invasion: The influence of the hidden side of the (epi)genome. *Functional Ecology*, *34*(2), 385–400.

Messer, P. W. (2013). SLiM: Simulating evolution with selection and linkage. *Genetics*, *194*(4), 1037–1039.

Morgan, K., Mboumba, J.-F., Ntie, S., Mickala, P., Miller, C. A., Zhen, Y., Harrigan, R. J., Le Underwood, V., Ruegg, K., Fokam, E. B., Tasse Taboue, G. C., Sesink Clee, P. R., Fuller, T., Smith, T. B., & Anthony, N. M. (2020). Precipitation and vegetation shape patterns of genomic and craniometric variation in the central African rodent *Praomys misonnei*. *Proceedings of the Royal Society B: Biological Sciences*, *287*(1930), 20200449.

Mounger, J., Ainouche, M. L., Bossdorf, O., Cavé-Radet, A., Li, B., Parepa, M., Salmon, A., Yang, J., & Richards, C. L. (2021). Epigenetics and the success of invasive plants. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, *376*(1826), 20200117.

Mullarkey, A. A., Byers, D. L., & Anderson, R. C. (2013). Inbreeding depression and partitioning of genetic load in the invasive biennial *Alliaria petiolata* (Brassicaceae). *American Journal of Botany*, *100*(3), 509–518.

Olazcuaga, L., Loiseau, A., Parrinello, H., Paris, M., Fraimout, A., Guedot, C., Diepenbrock, L. M., Kenis, M., Zhang, J., Chen, X., Borowiec, N., Facon, B., Vogt, H., Price, D. K., Vogel, H., Prud'homme, B., Estoup, A., & Gautier, M. (2020). A whole-genome scan for association with invasion success in the fruit fly *Drosophila suzukii* using contrasts of allele frequencies corrected for population structure. *Molecular Biology and Evolution*, *37*(8), 2369–2385.

Osborne, R., Meats, A., Frommer, M., Sved, J. A., Drew, R. A. I., & Robson, M. K. (1997). Australian distribution of 17 species of fruit flies (Diptera: Tephritidae) caught in Cue lure traps in February 1994. *Australian Journal of Entomology*, *36*(1), 45–50.

Parvizi, E., Vaughan, A. L., Dhami, M. K., & McGaughran, A. (2023). Genomic signals of local adaptation across climatically heterogenous habitats in an invasive tropical fruit fly (*Bactrocera tryoni*). *Heredity*, *132*, 18–29.

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., & Reich, D. (2012). Ancient admixture in human history. *Genetics*, *192*(3), 1065–1093.

Popa-Báez, A.-D., Catullo, R., Lee, S. F., Yeap, H. L., Mourant, R. G., Frommer, M., Sved, J. A., Cameron, E. C., Edwards, O. R., Taylor, P. W., & Oakeshott, J. G. (2020). Genome-wide patterns of differentiation over space and time in the Queensland fruit fly. *Scientific Reports*, *10*(1), 10788.

Popa-Báez, A.-D., Lee, S. F., Yeap, H. L., Westmore, G., Crisp, P., Li, D., Catullo, R., Cameron, E. C., Edwards, O. R., Taylor, P. W., & Oakeshott, J. G. (2021). Tracing the origins of recent Queensland fruit fly incursions into South Australia, Tasmania and New Zealand. *Biological Invasions*, *23*(4), 1117–1130.

Prentis, P. J., Wilson, J. R. U., Dormontt, E. E., Richardson, D. M., & Lowe, A. J. (2008). Adaptive evolution in invasive species. *Trends in Plant Science*, *13*(6), 288–294.

Reaser, J. K., Burgiel, S. W., Kirkey, J., Brantley, K. A., Veatch, S. D., & Burgos-RodrÃguez, J. (2020). The early detection of and rapid response (EDRR) to invasive species: A conceptual framework and federal capacities assessment. *Biological Invasions*, *22*(1), 1–19.

Rellstab, C., Dauphin, B., & Exposito-Alonso, M. (2021). Prospects and limitations of genomic offset in conservation management. *Evolutionary Applications*, *14*(5), 1202–1212.

Rellstab, C., Zoller, S., Walthert, L., Lesur, I., Pluess, A. R., Graf, R., Bodénès, C., Sperisen, C., Kremer, A., & Gugerli, F. (2016). Signatures of local adaptation in candidate genes of oaks (*Quercus* spp.) with respect

17524571, 2024, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/eva.13709 by Cochrane France, Wiley Online Library on [24/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

to present and future climatic conditions. *Molecular Ecology*, *25*(23), 5907–5924.

Rhoné, B., Defrance, D., Berthouly-Salazar, C., Mariac, C., Cubry, P., Couderc, M., Dequincey, A., Assoumanne, A., Kane, N. A., Sultan, B., Barnaud, A., & Vigouroux, Y. (2020). Pearl millet genomic vulnerability to climate change in West Africa highlights the need for regional collaboration. *Nature Communications*, *11*(1), 5274.

Rius, M., & Darling, J. A. (2014). How important is intraspecific genetic admixture to the success of colonising populations? *Trends in Ecology & Evolution*, *29*(4), 233–242.

Robinson, J., Kyriazis, C. C., Yuan, S. C., & Lohmueller, K. E. (2023). Deleterious variation in natural populations and implications for conservation genetics. *Annual Review of Animal Biosciences*, *11*(1), 93–114.

Ruegg, K., Bay, R. A., Anderson, E. C., Saracco, J. F., Harrigan, R. J., Whitfield, M., Paxton, E. H., & Smith, T. B. (2018). Ecological genomics predicts climate vulnerability in an endangered southwestern songbird. *Ecology Letters*, *21*(7), 1085–1096.

Sakai, A. K., Allendorf, F. W., Holt, J. S., Lodge, D. M., Molofsky, J., With, K. A., Baughman, S., Cabin, R. J., Cohen, J. E., Ellstrand, N. C., McCauley, D. E., O'Neil, P., Parker, I. M., Thompson, J. N., & Weller, S. G. (2001). The population biology of invasive species. *Annual Review of Ecology and Systematics*, *32*(1), 305–332.

Seebens, H., Bacher, S., Blackburn, T. M., Capinha, C., Dawson, W., Dullinger, S., Genovesi, P., Hulme, P. E., van Kleunen, M., Kühn, I., Jeschke, J. M., Lenzner, B., Liebhold, A. M., Pattison, Z., Pergl, J., Pyšek, P., Winter, M., & Essl, F. (2021). Projecting the continental accumulation of alien species through to 2050. *Global Change Biology*, *27*(5), 970–982. https://doi.org/10.1111/gcb.15333

Simberloff, D. (2009). The role of propagule pressure in biological invasions. *Annual Review of Ecology, Evolution, and Systematics*, *40*(1), 81–102.

Sutherst, R. W., & Yonow, T. (1998). The geographical distribution of the Queensland fruit fly, *Bactrocera* (*Dacus*) *tryoni*, in relation to climate. *Australian Journal of Agricultural Research*, *49*(6), 935–954.

Tayeh, A., Estoup, A., Hufbauer, R. A., Ravigne, V., Goryacheva, I., Zakharov, I. A., Lombaert, E., & Facon, B. (2013). Investigating the genetic load of an emblematic invasive species: The case of the invasive harlequin ladybird *Harmonia axyridis*. *Ecology and Evolution*, *3*(4), 864–871.

Wadgymar, S. M., DeMarche, M. L., Josephs, E. B., Sheth, S. N., & Anderson, J. T. (2022). Local adaptation: Causal agents of selection and adaptive trait divergence. *Annual Review of Ecology, Evolution, and Systematics*, *53*(1), 87–111.

Weldon, C. W., & Taylor, P. W. (2010). Desiccation resistance of adult Queensland fruit flies *Bactrocera tryoni* decreases with age. *Physiological Entomology*, *35*(4), 385–390.

Whiteley, A. R., Fitzpatrick, S. W., Funk, W. C., & Tallmon, D. A. (2015). Genetic rescue to the rescue. *Trends in Ecology & Evolution*, *30*(1), 42–49.

Wittmann, M. J., Metzler, D., Gabriel, W., & Jeschke, J. M. (2014). Decomposing propagule pressure: The effects of propagule size and propagule frequency on invasion success. *Oikos*, *123*(4), 441–450.

Yeaman, S. (2022). Evolution of polygenic traits under global vs local adaptation. *Genetics*, *220*(1), iyab134.

Zayed, A., Constantin, C. A., & Packer, L. (2007). Successful Biological Invasion despite a severe genetic load. *PLoS One*, *2*(9), e868.

Zhang, L.-W., Chen, J.-Q., Zhao, R.-M., Zhong, J., Lin, L.-H., Li, H., Ji, X., & Qu, Y.-F. (2023). Genomic insights into local adaptation in the Asiatic toad Bufo gargarizans, and its genomic offset to climate warming. *Evolutionary Applications*, *16*(5), 1071–1083.

Zhang, X., Guo, R., Shen, R., Landis, J. B., Jiang, Q., Liu, F., Wang, H., & Yao, X. (2023). The genomic and epigenetic footprint of local adaptation to variable climates in kiwifruit. *Horticulture Research*, *10*(4), uhad031.

Zhao, Z., Hui, C., Peng, S., Yi, S., Li, Z., Reddy, G. V. P., & van Kleunen, M. (2023). The world's 100 worst invasive alien insect species differ in their characteristics from related non-invasive species. *Journal of Applied Ecology*, *60*(9), 1929–1938.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Camus, L., Gautier, M., & Boitard, S. (2024). Predicting species invasiveness with genomic data: Is genomic offset related to establishment probability? *Evolutionary Applications*, *17*, e13709. https://doi.org/10.1111/eva.13709

17524571, 2024, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/eva.13709 by Cochrane France, Wiley Online Library on [24/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License