A whole-genome scan for association with invasion success in the fruit fly Drosophila suzukii using contrasts of allele frequencies corrected for population structure

Laure Olazcuaga, Anne Loiseau, Hugues Parrinello, Mathilde Paris, Antoine Fraimout, Christelle Guedot, Lauren M Diepenbrock, Marc Kenis, Jinping Zhang, Xiao Chen, et al.

A whole-genome scan for association with invasion success in the fruit fly *Drosophila suzukii* using contrasts of allele frequencies corrected for population structure

Laure Olazcuaga[1], Anne Loiseau[1], Hugues Parrinello[2], Mathilde Paris[3], Antoine Fraimout[1], Christelle Guedot[4], Lauren M. Diepenbrock[5], Marc Kenis[6], Jinping Zhang[7], Xiao Chen[8], Nicolas Borowiec[9], Benoit Facon[10], Heidrun Vogt[11], Donald K. Price[12], Heiko Vogel[13], Benjamin Prud'homme[3], Arnaud Estoup[*,1,14] and Mathieu Gautier[*,1,14]

[1]INRA, UMR CBGP (INRA – IRD – Cirad – Montpellier SupAgro), Montferrier-sur-Lez, France

[2]MGX, Biocampus Montpellier, CNRS, INSERM, Universite de Montpellier, Montpellier, France

[3]Aix Marseille Université, CNRS, IBDM, Marseille, France

[4]Department of Entomology, University of Wisconsin, Madison, WI

[5]Department of Entomology and Plant Pathology, NC State University

[6]CABI, Delémont, Switzerland

[7]MoA-CABI Joint Laboratory for Bio-safety, Chinese Academy of Agricultural Sciences, BeiXiaGuan, Haidian Qu, China

[8]College of Plant Protection, Yunnan Agricultural University, Kunming 650201, Yunnan Province, China

[9]UMR INRA-CNRS-Université Côte d'Azur Sophia Agrobiotech Institute, Sophia Antipolis, France

[10]UMR Peuplements Végétaux et Bioagresseurs en Milieu Tropical, INRA, Saint-Pierre, La Réunion, France

[11]Julius Kühn-Institut (JKI), Federal Research Centre for Cultivated Plants, Institute for Plant Protection in Fruit Crops and Viticulture, Dossenheim, Germany

[12]School of Life Sciences, University of Nevada, Las Vegas, Las Vegas, NV

[13]Department of Entomology, Max Planck Institute for Chemical Ecology, Jena, Germany

[14]These authors are joint senior authors on this work

***Corresponding author:** E-mail: mathieu.gautier@inrae.fr and arnaud.estoup@inrae.fr

Abstract

Evidence is accumulating that evolutionary changes are not only common during biological invasions but may also contribute directly to invasion success. The genomic basis of such changes is still largely unexplored. Yet, understanding the genomic response to invasion may help to predict the conditions under which invasiveness can be enhanced or suppressed. Here we characterized the genome response of the spotted wing drosophila *Drosophila suzukii* during the worldwide invasion of this pest insect species, by conducting a genome-wide association study to identify genes involved in adaptive processes during invasion. Genomic data from 22 population samples were analyzed to detect genetic variants associated with the status (invasive versus native) of the sampled populations based on a newly developed statistic, we called C_2, that contrasts allele frequencies corrected for population structure. We evaluated this new statistical framework using simulated data sets and implemented it in an upgraded version of the program BayPass. We identified a relatively small set of single nucleotide polymorphisms (SNPs) that show a highly significant association with the invasive status of *D. suzukii* populations. In particular, two genes, *RhoGEF64C* and *cpo*, contained SNPs significantly associated with the invasive status in the two separate main invasion routes of *D. suzukii*. Our methodological approaches can be applied to any other invasive species, and more generally to any evolutionary model for species characterized by non-equilibrium demographic conditions for which binary covariables of interest can be defined at the population level.

Key words: Biological invasions, *Drosophila suzukii*, GWAS, BayPass, Pool–Seq.

Introduction

Managing and controlling introduced species require an understanding of the ecological and evolutionary processes that underlie invasions. Biological invasions are also of more general interest because they constitute natural experiments that allow investigation of evolutionary processes on contemporary timescales. Colonizers are known to experience differences in biotic interactions, climate, availability of resources, and disturbance regimes relative to populations in their native regions, often with opportunities for colonizers to evolve changes in resource allocation which favor their success (Balanya *et al.*, 2006; Dlugosch *et al.*, 2015; Lee and Gelembiuk, 2008). Adaptive evolutionary shifts in response to novel selection regimes may therefore be central to initial establishment and spread of invasive species after introduction (Colautti and Barrett, 2013; Colautti and Lau, 2015). In agreement with this adaptive evolutionary shift hypothesis, experimental evidence is accumulating that evolutionary changes are not only common during invasions but also may contribute directly to invasion success (Bock *et al.*, 2015; Colautti and Lau, 2015; Ellstrand and Schierenbeck, 2000; Facon *et al.*, 2011; Lee, 2002; Ochocki and Miller, 2017; Williams *et al.*, 2016). However, despite an increase in theoretical and empirical studies on the evolutionary biology of invasive species in the past decade, the genetic basis of evolutionary adaptations during invasions is still largely unexplored (Barrett, 2015; Reznick *et al.*, 2019; Welles and Dlugosch, 2018).

The spotted wing drosophila, *Drosophila suzukii*, represents an attractive biological model to study invasion processes. This pest species, native to South East Asia, initially invaded North America and Europe, simultaneously in 2008, and subsequently La Réunion Island (Indian Ocean) and South America, in 2013. Unlike most Drosophilids, this species lays eggs in unripe fruits by means of its sclerotized ovipositor. In agricultural areas, it causes dramatic losses in fruit production, with a yearly cost exceeding one billion euros worldwide (e.g., Asplen *et al.*, 2015; Cini *et al.*, 2012). The rapid spreading of *D. suzukii* in America and Europe suggests its remarkable ability to adapt or to acclimate to new environments and host plants. Using evolutionarily neutral molecular markers, Adrion *et al.* (2014) and Fraimout *et al.* (2017) finely deciphered the routes taken by *D. suzukii* in its invasion worldwide. Interestingly, both studies showed that North American (plus Brazil) and European (plus La Réunion Island) populations globally represent separate invasion routes, with different native source populations and multiple introduction events in both invaded regions (Fraimout *et al.*, 2017). These two major and separate invasion pathways provide the

opportunity to evaluate replicate evolutionary trajectories. Finally, *D. suzukii* is a good model species for finely interpreting genomic signals of interest due to the availability of genome assemblies for this species (Chiu *et al.*, 2013; Ometto *et al.*, 2013; Paris *et al.*, 2020) along with the large amount of genomic and gene annotation resources available in its closely related model species *D. melanogaster* (Hoskins *et al.*, 2015).

In this context, advances in high-throughput sequencing technologies together with population genomics statistical methods offer novel opportunities to disentangle responses to selection from other forms of evolution. These advances are thus expected to provide insights into the genomic changes that might have contributed to the success in a new environment (reviewed in Bock *et al.*, 2015; Welles and Dlugosch, 2018). Hence, comparing the structuring of genetic diversity on a whole genome scale among invasive populations and their source populations might allow the characterization of the types of genetic variation involved in adaptation during invasion of new areas and their potential ecological functions. For example, Puzey and Vallejo-Marin (2014) used whole genome resequencing data to scan for shifts in site frequency spectra to detect positive selection in introduced populations of monkey-flower (*Mimulus guttatus*). Regions putatively under selection were associated with flowering time and abiotic and biotic stress tolerance and included regions associated with a chromosomal inversion polymorphism between the native and introduced range.

Identifying loci underlying invasion success can be considered in the context of whole-genome scan for association with population-specific covariate. These approaches, also known as Environmental Association Analysis (EAA), have received considerable attention in recent years (e.g., Coop *et al.*, 2010; de Villemereuil and Gaggiotti, 2015; Frichot *et al.*, 2013; Gautier, 2015). Most of the methodological developments have focused on properly accounting for the covariance structure among population allele frequencies that is due to the shared demographic history of the populations. This neutral covariance structure may indeed confound the relationship between the across population variation in allele frequencies and the covariates of interest (Coop *et al.*, 2010; Frichot *et al.*, 2013, 2015; Gautier, 2015). Yet, defining relevant environmental characteristics or traits as proxy for invasion success remains challenging and might even be viewed as the key aim. Therefore, we propose to simply summarize invasion success into a binary variable corresponding to the population's historical status (i.e., invasive or native) based on previous studies. By extension, functional annotation of the associated variants identified may provide insights into candidate traits underlying invasion success (Estoup *et al.*, 2016; Li *et al.*, 2008; Wu *et al.*, 2019).

The Bayesian hierarchical model initially proposed by Coop *et al.* (2010), later extended in Gautier (2015) and implemented in the software BayPass, represents one of the most flexible and powerful frameworks to carry out EAA since it efficiently accounts for the correlation structure among allele frequencies in the sampled populations. Although association analyses may be carried out with categorical or binary covariables (see the example of *Littorina* population ecotypes in Gautier, 2015), the assumed linear relationship with allele frequencies is not entirely satisfactory and may even be problematic when dealing with small data sets or if one wishes to disregard some populations.

In the present study, we developed a non-parametric counterpart for the association model implemented in BayPass (Gautier, 2015). This new approach relies on a contrast statistic, we named C_2, that compares the standardized population allele frequencies (i.e., the allele frequencies corrected for the population structure) between the two groups of populations specified by the binary covariable of interest. We evaluated the performance of this statistic on simulated data and used it to characterize the genome response of *D. suzukii* during its worldwide invasion. To that end, we generated Pool-Seq data (e.g., Gautier *et al.*, 2013; Schlotterer *et al.*, 2014) consisting of whole-genome sequences of pools of individual DNA (from n=50 to n=100 individuals per pool) representative of 22 worldwide populations sampled in both the invaded (n=16 populations) and native (n=6 populations) ranges of the species. We then estimated the C_2 statistics associated with the invasive vs. native status of the populations on a worldwide scale or considering separately each of the two invasion routes (European and American) as characterized by Fraimout *et al.* (2017). Our aim was to identify genomic regions and genes involved in adaptive processes underlying the invasion success of *D. suzukii*.

New Approaches

To identify single nucleotide polymorphisms (SNPs) associated with a population-specific binary trait, such as the invasive versus native status of *D. suzukii* populations, we developed a new statistic, we called C_2. The C_2 statistic was designed to contrast SNP allele frequencies between the two groups of populations specified by the binary trait while accounting for the possibly complex evolutionary history of the different populations. Indeed, the shared population history is a major (neutral) contributor to allele frequency differentiation across populations (e.g. Bonhomme *et al.*, 2010; Gunther and Coop, 2013) that may confound association signals (e.g. Coop *et al.*, 2010; Gautier, 2015).

We here relied on the multivariate normal approximation introduced by Coop *et al.* (2010) and further extended by Gautier (2015) to model population allele frequencies and to define the C_2 contrast statistic. More precisely, consider a

sample made of J populations (each with a label $j=1,\ldots,J$) that have been characterized for I bi-allelic SNPs (each with a label $i=1,\ldots,I$), with the reference allele arbitrarily defined (e.g., by randomly drawing the ancestral or the derived state). Let α_{ij} represent the (unobserved) allele frequency of the reference allele at SNP i in population j. As previously defined and discussed (Coop *et al.*, 2010; Gautier, 2015), we introduced an instrumental allele frequency $\alpha^{\star}_{ij}$ (for each SNP i and population j) taking values on the real line such that $\alpha_{ij}=\min\left(1,\max\left(0,\alpha^{\star}_{ij}\right)\right)$.

Following Coop *et al.* (2010) and Gautier (2015), a multivariate Gaussian (prior) distribution of the vector $\boldsymbol{\alpha^{\star}_{i}}=\left\{\alpha^{\star}_{ij}\right\}_{1\ldots J}$ is then assumed for each SNP i:

$$\boldsymbol{\alpha^{\star}_{i}}\,|\,\boldsymbol{\Omega},\pi_i \sim \mathrm{N}_J\left(\pi_i\mathbf{1}_J;\pi_i(1-\pi_i)\boldsymbol{\Omega}\right) \qquad (1)$$

where $\mathbf{1}_J$ is the all-one vector of length J; $\boldsymbol{\Omega}$ is the (scaled) covariance matrix of the population allele frequencies which captures information about their shared demographic history; and π_i is the weighted mean frequency of the SNP i reference allele. If $\boldsymbol{\Omega}$ is used to build a tree or an admixture graph (Pickrell and Pritchard, 2012), π_i corresponds to the root allele frequency. We further define for each SNP i the vector $\ddot{\boldsymbol{\alpha}}_i$ of standardized (instrumental) allele frequencies in the J populations as:

$$\ddot{\boldsymbol{\alpha}}_i = \Gamma_{\Omega}^{-1}\left\{\frac{\alpha_{ij}-\pi_i}{\sqrt{\pi_i(1-\pi_i)}}\right\}_{(1..J)} \qquad (2)$$

where Γ_{Ω} results from the Cholesky decomposition of Ω (i.e., $\Omega=\Gamma_{\Omega}{}^{t}\Gamma_{\Omega}$). The vector $\ddot{\boldsymbol{\alpha}}_i$ thus contains scaled allele frequencies that are corrected for both the population structure (summarized by $\boldsymbol{\Omega}$) and the across-population (e.g., ancestral) allele frequency (π_i).

The C_2 contrast statistic is then simply defined as the mean squared difference of the sum of standardized allele frequencies of the two groups of populations defined according to the binary trait modalities:

$$C_2(i)=\frac{1}{\boldsymbol{c}^t\boldsymbol{c}}\left(\ddot{\boldsymbol{\alpha}}_i{}^t\boldsymbol{c}\right)^2 \qquad (3)$$

where $\boldsymbol{c}=c_{j\,(1..J)}$ is a vector of the trait values observed for each population j such that $c_j=1$ (respectively $c_j=-1$) if population j displays the first (respectively second) trait modality. One may also define $c_j=0$ to exclude a given population j from the comparison. According to our model, the J elements of $\ddot{\boldsymbol{\alpha}}_i$ are independent and identically distributed as a standard Gaussian distribution under the null hypothesis of only neutral marker differentiation. The C_2 statistic is thus expected to follow a χ^2 distribution with one degree of freedom.

The estimation of the C_2 statistic was here performed using the Markov-Chain Monte Carlo (MCMC) algorithm implemented in the BayPass software (Gautier, 2015). Due to the hierarchical structure of the underlying Bayesian model, the SNP population allele frequencies (in the vectors $\boldsymbol{\alpha^{\star}_{i}}$'s) tend to be pulled closer together (i.e., shrunk) because they share the same overarching prior multivariate Gaussian

distribution (equation 1) (see e.g., Kruschke, 2014, pp. 245-249, for a general presentation of shrinkage in Bayesian hierarchical models). This leads to a shrinkage of the estimated C_2 posterior means and the estimates of the SNP-specific XtX differentiation statistic, as already noticed in Gautier (2015). The XtX is indeed defined as the variance of the standardized allele frequencies of the SNP across the populations ($XtX = \ddot{\boldsymbol{\alpha}}_i{}^t \ddot{\boldsymbol{\alpha}}_i$) and is thus analogous to a SNP-specific F_{ST} that would account for the overall covariance structure of the population allele frequencies (Gunther and Coop, 2013). As a matter of expedience, to ensure proper calibration of both the C_2 and XtX estimates (see below) we decided to rescale the posterior means of the $\ddot{\alpha}_{ij}$'s as:

$$\widehat{\underline{\ddot{\boldsymbol{\alpha}}}_i} = \left\{ \frac{\widehat{\ddot{\alpha}}_{ij} - \mu_{\ddot{\alpha}}}{\sigma_{\ddot{\alpha}}} \right\}_{(1\ldots J)} \tag{4}$$

where $\widehat{\ddot{\alpha}}_{ij}$ is the posterior means of $\ddot{\alpha}_{ij}$ and $\mu_{\ddot{\alpha}}$ (respectively $\sigma_{\ddot{\alpha}}$) is the mean (respectively standard deviation) of the $I \times J$ $\widehat{\ddot{\alpha}}_{ij}$'s ($\mu_{\ddot{\alpha}} \simeq 0$ usually). The following estimators of XtX and C_2, denoted for each SNP i as $\widehat{XtX^\star}(i)$ and $\widehat{C_2}(i)$ respectively, were then obtained as:

$$\begin{aligned} \widehat{XtX^\star}(i) &= \widehat{\underline{\ddot{\boldsymbol{\alpha}}}_i}^t \widehat{\underline{\ddot{\boldsymbol{\alpha}}}_i} \\ \widehat{C_2}(i) &= \frac{1}{\boldsymbol{c}^t \boldsymbol{c}} \left(\widehat{\underline{\ddot{\boldsymbol{\alpha}}}_i}^t \boldsymbol{c} \right)^2 \end{aligned} \tag{5}$$

Under the null hypothesis, $\widehat{XtX^\star}(i) \sim \chi^2_J$ and $\widehat{C_2}(i) \sim \chi^2_1$ allowing one to rely on standard decision making procedures, e.g. based on p-values or more preferably on q-values to control for multiple-testing issues (Storey and Tibshirani, 2003).

Results

Simulation-based evaluation of the performance of our novel statistical framework

To evaluate the performances of the C_2 contrast statistic for the identification of SNPs associated with binary population-specific covariables, we simulated 100 data sets under the evolutionary scenario depicted in Figure 1A. This scenario was adapted from the so-called HsIMM model proposed by de Villemereuil *et al.* (2014) to deal with binary environmental constraints rather than environmental gradient. This scenario choice was motivated by the use of the same underlying HsIMM demographic history in previous studies to carry out in-depth evaluations of a wide range of popular methods for genome-wide selection scans and EAA in realistic situations (de Villemereuil *et al.*, 2014; Gautier, 2015). In these studies, the XtX statistic (for genome-wide selection scan approaches) and the Bayes Factor (BF) for EAA, as computed with BayPass, were found to be among the best performing approaches in their respective categories under various scenarios, including HsIMM. Each simulated data set consisted of 5,000 SNPs genotyped for 320 individuals belonging to 16 differentiated populations subjected to two different contrasting environmental constraints, denoted *ec1* and *ec2* in Figure 1A. The *ec1* constraint was aimed

at mimicking adaptation of eight pairs of geographically differentiated populations to two different ecotypes (e.g., host plant) replicated in different geographic areas. Conversely, the *ec2* might be viewed as replicated local adaptive constraints with a first type *a* specifying a large native area with several geographically differentiated populations (here six), and a second type *b* specifying invasive areas with differentiated populations originating from various regions of the native area (i.e., not related to the same extent to their contemporary native populations). It should be noted that the two *ec1* types were evenly distributed in the population tree while for *ec2*, the type *b* was over-represented in 10 populations (Figure 1A). During the adaptive phase, the fitness of individuals in the environment of their population of origin was determined by their genotypes at 25 SNPs for *ec1* and 25 SNPs for *ec2* constraints (hereafter referred to as *ec1* and *ec2* selected SNPs, respectively). Overall, the realized F_{ST} (Weir and Cockerham, 1984) ranged from 0.110 to 0.122 (0.116 on average) across the data sets, a level of differentiation similar to that observed in our worldwide *D. suzukii* sample (see below).

We further estimated with BayPass (Gautier, 2015) the C_2 statistics for each *ec1* or *ec2* contrasting environmental constraints together with the corresponding BF as an alternative measure of the support for association representative of state-of-the-art EAA approaches. For comparison with standard genome-wide selection scan approaches, we also estimated the SNP XtX differentiation statistic, using both the posterior mean estimator (Gautier, 2015) and the $\widehat{XtX^{\star}}$ estimator described above. Note however that because selection scan approaches rely on (covariate-free) differentiation statistics (here, the XtX), they do not allow to distinguish among the outlier SNPs those responding to the *ec1* constraint from those responding to the *ec2* constraint.

Based on the status of each simulated SNP (i.e., neutral, and *ec1* or *ec2* selected) and combining results in the 100 simulated data sets, standard receiver operating curves (ROCs) were computed (Grau *et al.*, 2015) and plotted in Figure 1B (respectively 1C) for the six statistics. This allowed comparing for various thresholds covering the range of variation of the different statistics, the power to detect *ec1* (respectively *ec2*) selected SNPs (i.e., the proportion of true positives among the corresponding selected SNPs) as a function of the false positive rates (FPR, i.e., the proportion of positives among neutral SNPs). The C_2 statistic was found to efficiently detect SNPs affected by *ec1* and *ec2* environmental constraints, the area under the ROC curve (AUC) being equal to 0.977 (Figure 1B) and 0.943 (Figure 1C), respectively. The unbalanced population representation of the two *ec2* types had a limited impact on the performance of the C_2 statistic to identify the underlying selected SNPs. In addition, the

C_2 statistics clearly discriminated the selected SNPs according to their underlying environmental constraint. In other words, no selection signal was identified by the C_2 statistic computed for the *ec2* (respectively *ec1*) contrast on *ec1* (respectively *ec2*) selected SNPs, resulting in ROC AUC close to the value of 0.5 obtained with a random classifier.

The ROC curves displayed in Figures 1B and 1C also revealed nearly identical performance of the C_2 statistic and the BF. Accordingly, the correlation between both statistics were fairly high (Pearson's r equal to 0.983 and 0.923 for *ec1* and *ec2*, respectively). Yet, one practical advantage of the C_2 statistic was its good calibration with respect to the null hypothesis of no association, the corresponding p-values (assuming a χ^2 distribution with 1 degree of freedom) being close to uniform (Figure S1).

Similarly, the two XtX estimators were found highly correlated (Pearson's $r = 0.998$) with almost confounded ROC curves, but only the $\widehat{XtX^{\star}}$ was properly calibrated (Figure S2). Their performances were however clearly worse than those obtained with the C_2 (and BF) statistics. This was in part explained by their inability to discriminate between the two types of selected SNPs, i.e. selected SNPs overly differentiated in *ec2* generating false positives in the identification of *ec1* SNPs (Figure 1B) and vice versa. Accordingly, ROC AUC in Figure 1B for the XtX were also smaller than in Figure 1C, *ec1* selected SNPs being more differentiated than those in *ec2* due to the simulated design. Yet, the power of the XtX statistic to detect *ec1* or *ec2* selected SNPs remained substantially smaller than that of the corresponding C_2 contrast statistics. For instance, at the 1% p-value significance threshold, the power to detect ec1 (respectively ec2) selected SNPs was equal to 72.6% (respectively 59.1%) with the C_2 statistic and only 17.1% (respectively 10.4%) with the $\widehat{XtX^{\star}}$ estimator, even when considering for the latter, a unilateral test to only target overly differentiated SNPs. Note that, as expected from the good calibration of the $\widehat{XtX^{\star}}$ statistic, similar results were obtained when considering empirical p–value thresholds computed from the distribution of the XtX statistics estimated from neutral SNPs.

Robustness of the different approaches to demographic events typical of biological invasions

The HsIMM scenario used for the above simulations may imperfectly capture some characteristics of the demographic history of invasive species. Indeed, an invasion may be triggered by a relatively small number of colonizers leading to population bottlenecks (e.g, Estoup *et al.*, 2016). Moreover, invasive species, particularly those with low dispersal capabilities (which is not the case of *D. suzukii*), may be prone to allele surfing, wherein a variant rises to high frequency by chance as the expansion wave advances due to repeated bottlenecks

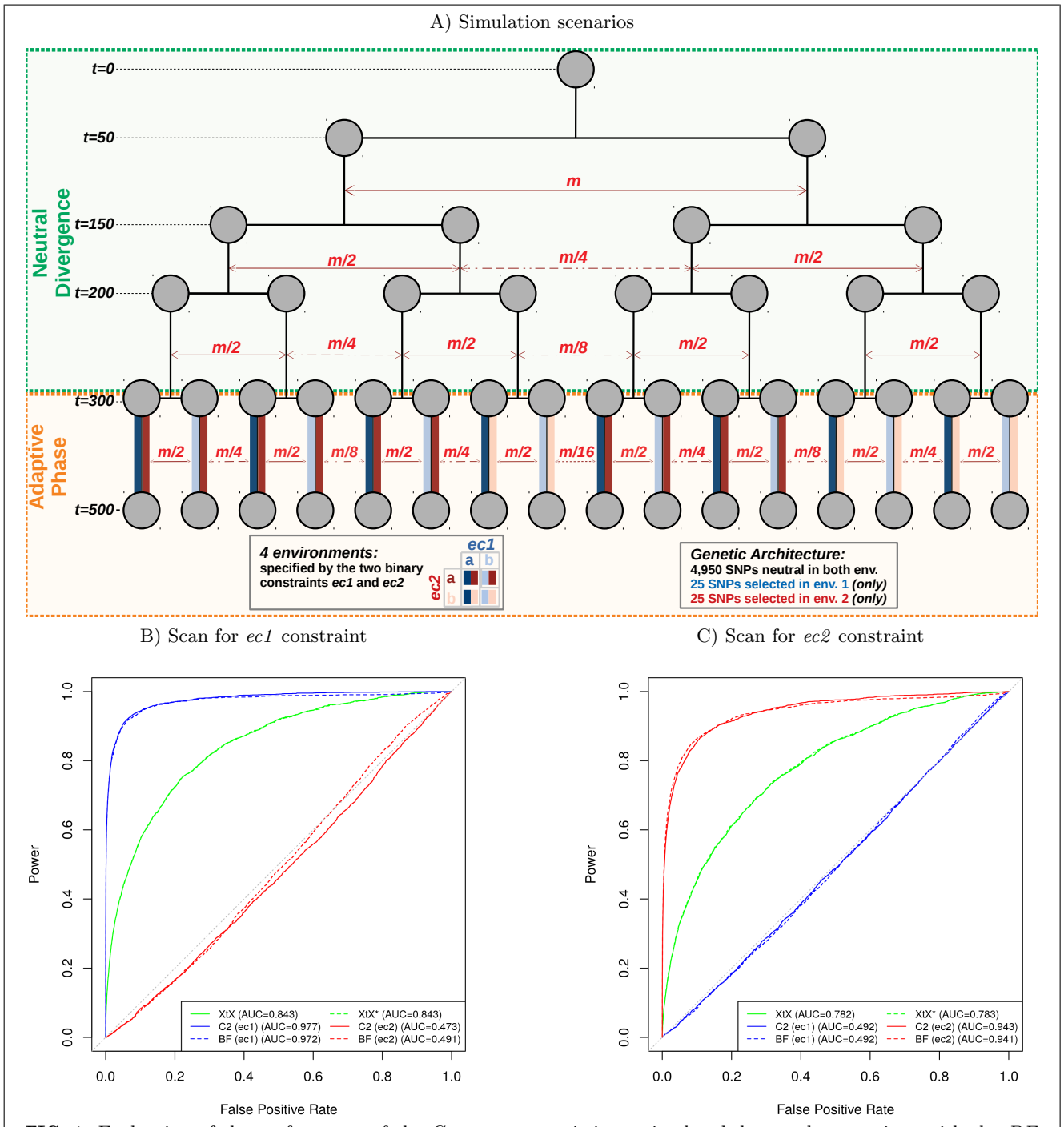

FIG. 1. Evaluation of the performance of the C_2 contrast statistic on simulated data and comparison with the BF for association and two XtX SNP-specific differentiation estimators. A) Schematic representation of the demographic scenario used for the simulation. It consists of two successive phases: (i) a neutral divergence phase with migration (only some illustrative migration combinations being represented) leading to the differentiation of an ancestral population into 16 populations after four successive fission events (at generations $t=50$, $t=150$, $t=200$ and $t=300$); and (ii) an adaptive phase (lasting 200 generations) during which individuals were subjected to selective pressures exerted by two environmental constraints (*ec1* and *ec2*) each having two possible modalities (a or b) according to their population of origin (i.e., eight possible environments in total). Out of the 5,000 simulated SNPs, the fitness of individuals in the environment of their population of origin was determined by their genotypes at 25 SNPs for *ec1* and 25 SNPs for *ec2* constraints. In total 100 data sets were simulated. B) and C) The ROC curves associated to the *ec1* and *ec2* C_2 contrasts and the two corresponding BF for association are plotted together with those associated with the two XtX estimators (i.e., posterior mean estimator XtX, and the new calibrated estimator $XtX^{\star}$). The FPR's associated to each statistic were obtained from the corresponding neutral SNP estimates combined over the 100 simulated data sets ($n=4{,}950\times100=495{,}000$ values in total). Similarly, the TPR's were estimated from either the $n=2{,}500$ combined *ec1* (B) or *ec2* (C) selected SNPs. ROC AUC values are given between parentheses.

(Excoffier and Ray, 2008). At the genomic level, such bottleneck events may lead to large but correlated random fluctuations of some variant frequencies in the invasive populations deriving from the founders of the primary introduction, up to the fixation of the same variant in all of them. To evaluate to which extent such demographic events may result in spurious signals of association of some variants with the invasive status of populations, we simulated data sets (with 165,020 and 152,321 evolutionary neutral SNPs respectively) under two invasion scenarios: i) the scenario *Inv1* (Figure 2A) in which each invasive population of an area derived from the same primarily introduced population with a bottleneck occurring at different time in the past; and ii) the scenario *Inv2* (Figure 2B) in which the invasive populations of an area are successively founded one after the other with a bottleneck event at each foundation, a process likely to favor allele surfing during geographic range expansion. As for the *D. suzukii* case study detailed below, the two scenarios consisted of six native populations with a moderate level of genetic structuring (realized F_{ST} equal to 4.95% and 4.93% respectively) and two groups of eight invasive populations, each group originating from one of the native populations and corresponding to a given invaded area. For the scenario *Inv1*, the chosen simulation parameters resulted in a realized F_{ST} within each group of 3.49% and 3.48% and an overall realized F_{ST} (among the 22 simulated native and invasive populations) of 9.60%. In the scenario *Inv2*, the succession of bottlenecks led to an increased level of differentiation among the invasive populations, compared to the scenario *Inv1*, with a realized F_{ST} within each of the two groups equal to 5.76% and 5.77% and an overall realized F_{ST} of 14.5%.

We ran BayPass on the two data sets simulated under the invasion scenarios *Inv1* and *Inv2* to identify outlying differentiated SNPs (based on the $XtX^{\star}$ statistic) and to evaluate the support for association of each SNP with the invasive population status (based on both the C_2 statistic and the BF criterion). Three different association tests were used to compare the six native populations with i) all the 16 invasive populations (C_2^{all} and BF^{all}); ii) the eight invasive populations from the first group (C_2^{G1} and BF^{G1}); or iii) the eight invasive populations from the second group (C_2^{G2} and BF^{G2}). As shown in Figures 2C and 2D, the estimated scaled covariance matrices $\mathbf{\Omega}$ provided an overall structuring of genetic diversity across the 22 simulated populations that was consistent with the simulated histories. The Singular Value Decomposition (SVD) of the two $\mathbf{\Omega}$ matrices separated the populations according to their invasive or native origins in the first axis of variation while the second axis separated the two groups of invasive populations. Under the *Inv2* scenario, the successive bottlenecks at the origin of the different invasive populations also resulted in their clear separation on the

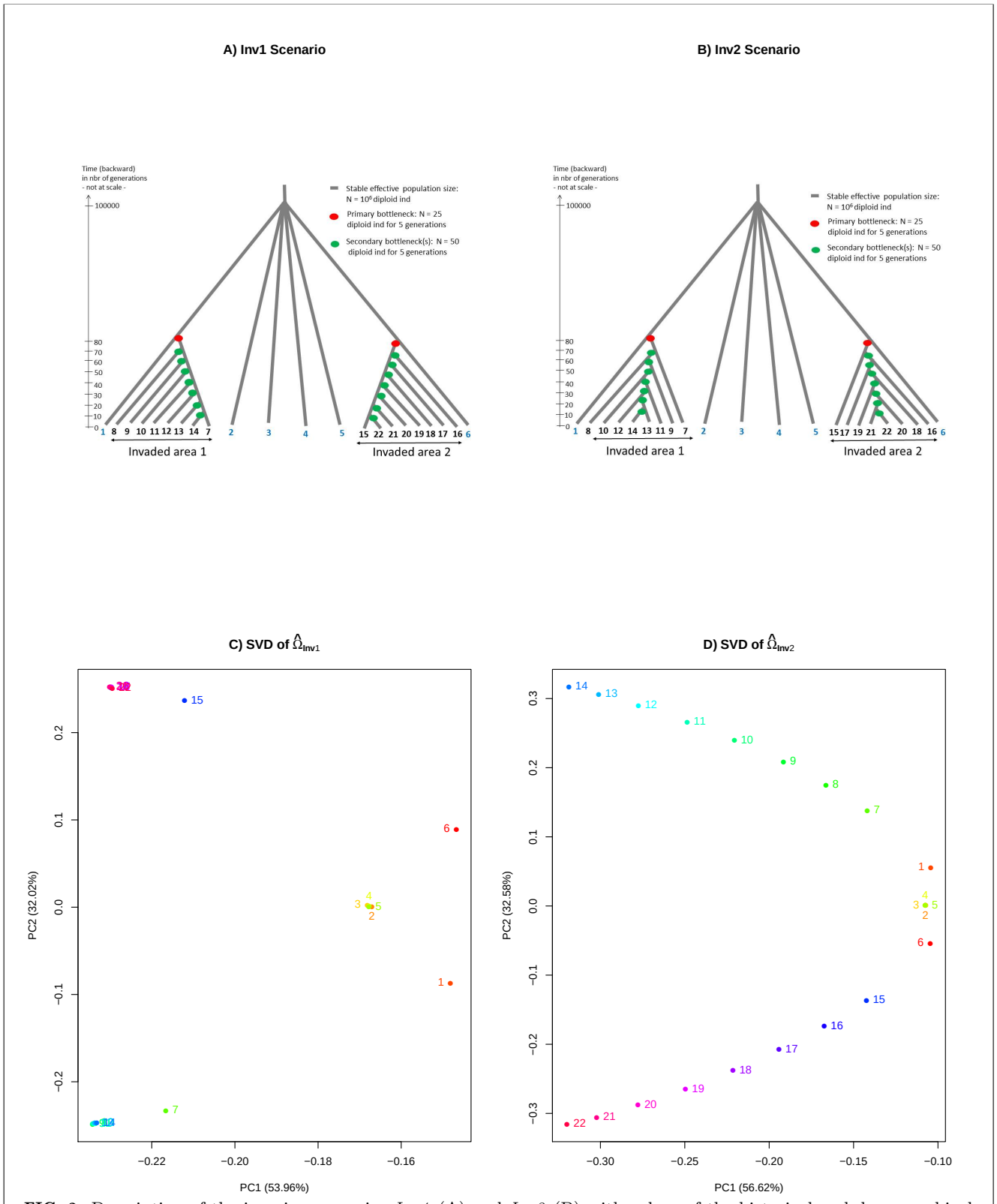

FIG. 2. Description of the invasion scenarios *Inv1* (A) and *Inv2* (B) with values of the historical and demographical parameters used to simulated evolutionary neutral SNP data sets. The coordinates of the 22 populations on the first two axes of variation (following Singular Value Decomposition with the plot.omega function available in the BayPass package) of the scaled covariance matrix of allele frequencies $\widehat{\mathbf{\Omega}}_{Inv1}$ (resp. $\widehat{\mathbf{\Omega}}_{Inv2}$) estimated using BayPass from the SNP data simulated under *Inv1* (resp. *Inv2*) are plotted in C (resp. D).

	scenario *Inv1*		Scenario *Inv2*	
	all	G1 and G2	all	G1 and G2
BF>20 (>15)	1.9×10^{-4} (9.2×10^{-4})	2.4×10^{-4} (1.0×10^{-3})	2.6×10^{-4} (1.1×10^{-3})	3.6×10^{-4} (1.3×10^{-3})
C_2 q-value<0.01 (<0.05)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	4.6×10^{-5} (1.9×10^{-3})
$XtX^\star$ q-value<0.01 (<0.05)	0.0 (0.0)		1.3×10^{-5} (4.6×10^{-5})	

Table 1. Proportion of SNPs (False Positive Rate) simulated under the invasion scenarios *Inv1* (n=165,020 SNPs) and *Inv2* (n=152,321 SNPs) displaying outlying differentiation (based on the $XtX^\star$ statistic) or showing a signal of association with the population invasive status (based on the BF criterion or the C_2 statistic). Support for association was evaluated using the BayPass regression models (BF criterion) or the q-value derived from the estimated contrast statistics C_2 for the three different tests comparing the six native populations allele frequencies to i) all 16 invasive populations ("all"); ii) the eight invasive populations from the first group ("G1"); or iii) the eight invasive populations from the second group ("G2") (Figure S3). Results from the two latter tests were combined to compute False Positive Rates (columns "G1 and G2" in the Table). Note that the BF threshold of 20 dB (resp. 15 dB) corresponds to decisive (resp. very strong) evidence in favor of association according to the Jeffreys' rule (Jeffreys, 1961). To account for the bilateral nature of the underlying test (SNPs might be over or under-differentiated if under directional or balancing selection), the p-values derived from the $XtX^\star$ statistic were computed as $p=1-2\,|\,\Phi_{\chi^2(J)}\left(\widehat{\mathrm{XtX}}\right)-0.5\,|$, where $\Phi_{\chi^2(J)}$ represents the cumulative density function of the χ^2 distribution with J degrees of freedom (here $J=22$).

first axis (Figure 2D). These results suggest that the shared population history may be globally well accounted for by the model. Accordingly, the distribution of the p-values derived from the $XtX^\star$ was close to uniform for the scenario *Inv1* (Figure S3A) as expected given that all the analyzed SNPs evolved neutrally. This resulted in a desired null FPR at both the 1% and 5% q–value threshold (Table 1). Yet, for the *Inv2* scenario, we observed an (undesirable) excess of small p–values derived from the $XtX^\star$. However, this feature only resulted in an almost null FPR after correcting for multiple testing, the FPR being equal to only 4.6×10^{-5} at the 5% q–value threshold (Table 1).

Similar patterns were observed for the distribution of the p–values derived from the different C_2 statistics (Figures S3C and S3D). Note that for the *Inv1* scenario, we observed a smaller proportion of small C_2 p-values than expected under uniform expectation that might originate from an imperfect deshrinking of the standardized allele frequencies. Overall, the FPR associated to the C_2 statistics was null at the 1% q–value threshold for both the scenarios *Inv1* and *Inv2* except, for the latter, when considering C_2^{G1} and C_2^{G2} group-specific contrasts for which the FPR was equal to 4.6×10^{-5} (Table 1). At the stringent decisive evidence threshold of 20 dB (Jeffreys, 1961) on BF, the FPR was always one order of magnitude higher ranging from 1.9×10^{-4} to 3.6×10^{-4} across the different analyses. This may result in a substantial amount of false positives on (real) data sets of millions SNPs.

Interestingly, Figures S3E and S3F showed that the SNPs with the highest BF were clearly different from those with the highest C_2 suggesting that C_2 and BF may actually be sensitive to different confounding structuring of SNP genetic diversity. As detailed in Table S1 for the simulated data set *Inv2*, the median of the SNP-specific F_{ST} computed within all the invasive populations was equal to 0.12 across the 162 top BF^{all} SNPs (with

a $BF^{\text{all}} > 15$), close to the median computed over all the simulated SNPs, but far lower than the median computed across the 74 top C_2^{all} SNPs (with a C_2^{all} derived p–value $> 10^{-4}$). Accordingly, the size of the allele frequency range within the invasive populations was larger for the top C_2^{all} SNPs than for the top BF^{all} SNPs (0.745 against 0.450); but the reverse was observed within native populations. In addition, the average allele frequencies of the top C_2^{all} SNPs were far smaller (median of 0.017) with a smaller range of variation (median of 0.070) when compared to top BF^{all} SNPs (median of 0.251) characterized by a wider range of variation (median of 0.215), whereas the same values computed over all the SNPs were intermediary between these two extremes. Overall, these results suggest that, in such invasion scenarios, the regression analysis based on BF may be more sensitive to neutral variants common in the native populations and displaying rather homogeneous allele frequencies within the invasive populations. In contrast, the C_2 statistic appears more sensitive to variants rare in the native populations and raising to high allele frequencies with high heterogeneity in the invasive populations. Yet, correction for multiple testing on the C_2 derived p–value turned out to more efficiently control for false positive rates than standard thresholds on the BF, and this for both simulated invasion scenarios.

Genome-wide scan for association with invasion success in *D. suzukii*

To identify genomic regions associated with the invasion success of *D. suzukii*, we carried out a genome scan, based on the C_2 statistic, to contrast the patterns of genetic diversity among 22 populations originating from either the native (n=6 populations) or invaded areas (n=16 populations) (Figure 3A). To that end we sequenced pools of 50 to 100 individuals representative of each population (Table S2) and mapped the resulting sequencing reads onto the newly released WT3-2.0 *D. suzukii* genome assembly (Paris *et al.*, 2020). These Pool-Seq data allowed the characterization of 11,564,472 autosomal and 1,966,184 X–linked SNPs segregating in the 22 populations that were sub-sampled into 154 autosomal and 26 X–linked data sets (of ca. 75,000 SNPs each) for further analyses.

The overall differentiation was estimated using the recently developed F_{ST} estimator for Pool-Seq data (Hivert *et al.*, 2018). It ranged from 8.86% to 9.02% (8.95% on average) for the autosomal data sets and from 17.6% to 17.8% (17.8% on average) for the X–chromosome data sets. Although a higher genetic differentiation is expected for the X-chromosome even under equal contribution of males and females to demography, the almost twice higher overall differentiation observed for the X chromosome compared to autosomes might have been accentuated by unbalanced sex-ratio

(e.g., polyandry), male–biased dispersal or a higher impact of selection on the X–chromosome (Clemente *et al.*, 2018). Inferring sex-specific demography was beyond the scope of the present study, but for our purposes, this finding justified to perform separate genome scans on autosomal and X–linked SNPs.

We ran BayPass on the different data sets to estimate, for every SNPs, the C_2 statistic that contrasts the allele frequencies of native and invasive populations, while accounting for their shared population history as summarized in the scaled covariance matrix $\mathbf{\Omega}$. Interestingly, the estimated $\mathbf{\Omega}$ matrices for autosomal and X–linked SNPs resulted in a similar structuring of the genetic diversity across the 22 populations (Figure S4), which may rule out selective forces as the main driver of the differences of global differentiation levels observed between the two chromosome types. Note also that the representation of the two major axes of variation of $\mathbf{\Omega}$ (Figure S4) resulted in a pattern intermediary between the ones obtained when analyzing the data simulated above under the invasion scenarios *Inv1* and *Inv2* (Figure 2). The distribution of the p-values derived from the C_2 statistics was well-behaved, being close to uniform for higher p–values (Figure S5A). To account for multiple testing issues, we used the *qvalue* R package (Storey and Tibshirani, 2003) to compute the individual SNP q–values plotted in Figure 3B.

A striking feature of the resulting Manhattan plot was the lack of clustering of SNPs with high q–values which might be related to a small extent of linkage disequilibrium (LD) across the *D. suzukii* populations, as expected from their large effective populations sizes (Fraimout *et al.*, 2017). We identified 101 SNPs (including three X–linked) that were significant at the 1% q–value threshold (i.e., 1% of these 101 SNPs are expected to be false positives). As a matter of comparison, we also estimated the BF for association of the (standardized) population allele frequencies with the native or invasive status of the population, i.e., under a parametric regression model (Gautier, 2015) (Figure S6A). Out of the 101 significant SNPs previously identified, 80 displayed a BF$>$20 db, the threshold for decisive evidence according to the Jeffreys' rule (Jeffreys, 1961). However, in total, 6,406 SNPs displayed a BF$>$20 db probably due to an increase of false positives at this threshold (see above simulations under invasion scenarios). We also compared the C_2 statistic to the XtX measure of overall differentiation. The (two-sided) p–values derived from the latter were also well behaved (Figure S5B) and allowed the computation of q–values to control for multiple testing. As shown in Figure S6B, at the same 1% q–value threshold for XtX, 71 out of the 101 C_2 significant SNPs were significantly differentiated but they represented only a small proportion of the 35,546 significantly differentiated SNPs. This is not surprising since invasion success is obviously

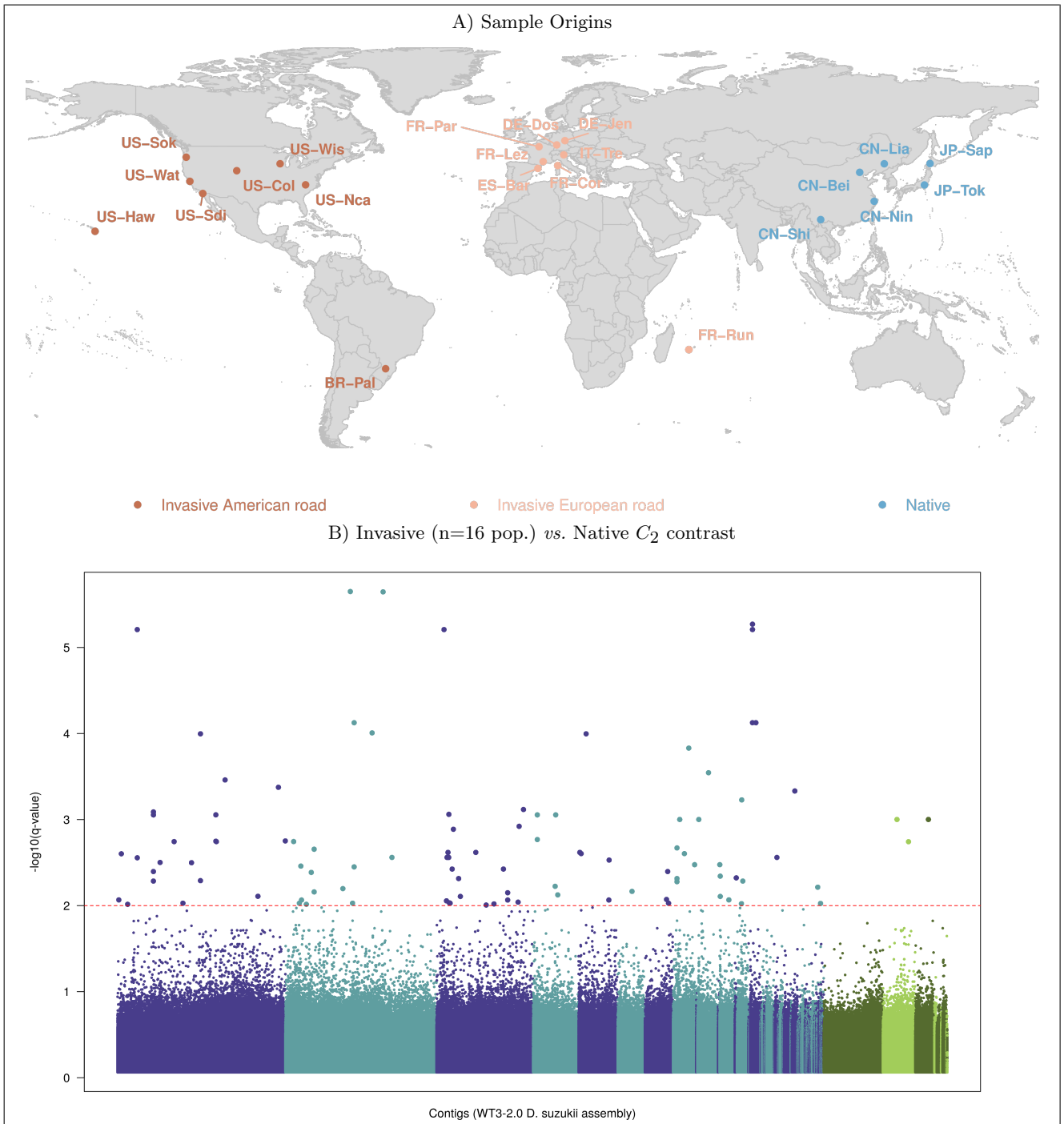

FIG. 3. Whole-genome scan for association with invasion success in *D. suzukii*. A) Geographic location of the 22 *D. suzukii* population samples genotyped using a pool-sequencing methodology. Population samples from the native range are in blue and those from the invaded range are in red (American invasion route) or light red (European invasion route) (Fraimout *et al.*, 2017). See Table S2 for details on each population sample. B) Manhattan plot of the SNP q-values on a $-\log_{10}$ scale derived from the estimated C_2 statistics for the native *vs.* invasive status contrast of the 22 worldwide *D. suzukii* populations. SNPs are ordered by their position on their contig of origin displayed with alternating dark blue and light blue color when autosomal and dark green and light green when X–linked. The horizontal dashed line indicates the 1% q-value threshold (here corresponding to a p-value threshold of 8.49×10^{-8}) which gives the expected FDR (False Discovery Rate), i.e., the expected proportion of false positives among the 110 SNPs (highlighted in the plot) above this threshold.

not the only selective constraint exerted on the 22 worldwide populations considered here.

The North-American (plus Brazil) and European (plus La Réunion Island) populations globally represent separate invasion routes that can be considered as two independent invasion replicates (Figure 3A). Interestingly enough, this feature of historical invasion fits well with the overall pattern of structuring of genetic diversity inferred from the $\mathbf{\Omega}$ matrix estimated with our Pool-Seq data (see above and Figure S4). To identify signals common or specific to each invasion routes, we estimated the C_2 statistic associated with the invasive vs. native status focusing either on the native and invasive populations of the European invasion route (C_2^{EU}), or native and invasive populations of the American invasion route (C_2^{AM}). Note that the two invasion routes were both represented by eight invasive populations, suggesting similar power for the two C_2^{EU} and C_2^{AM} statistics. As observed above, the distribution of p–values derived from C_2^{EU} and C_2^{AM} were found well behaved (Figures S5C and S5D, respectively) and hence q–values to control for multiple testing could be confidently computed. The cross-comparisons of the C_2 statistics considering the 22 worldwide populations (hereafter denoted C_2^{WW}), the C_2^{EU} and the C_2^{AM} are plotted in Figures 4A (C_2^{EU} versus C_2^{WW}), 4B (C_2^{AM} versus C_2^{WW}) and 4C (C_2^{AM} versus C_2^{EU}).

In total, 204 SNPs (detailed in Table S3) were significant in at least one of the three contrasts at the 1% q–value threshold. The overlap among the three different sets of significant SNPs was summarized in the Venn diagram displayed in Figure 4D. Among the 68 SNPs significant for the C_2^{EU}, 15 were also significant for C_2^{WW} and 49 were not significant in the other tests. Likewise, among the 72 SNPs found significant for the C_2^{AM}, 14 were also significant for C_2^{WW} and 54 were not significant in the other tests. Hence, the majority of the significant SNPs identified with either the C_2^{EU} or the C_2^{AM} contrasts might be viewed as specific to one of the two invasion routes, the signal being lost in the global worldwide comparison for a substantial proportion of them. This is presumably due to a reduced power resulting from the addition of non-informative populations when computing the C_2^{WW} statistic. Conversely, 68 SNPs found significant with C_2^{WW} were neither significant with C_2^{EU} nor C_2^{AM} contrasts. These SNPs might correspond to partially convergent signals among the two invasion routes (i.e., the informative populations are distributed among the two routes). Most interestingly, four SNPs were found significant at the 1% q–value threshold in the three contrast analyses (C_2^{EU}, C_2^{AM} and C_2^{WW}) and might thus be viewed as strong candidates for association with the global worldwide invasion success of *D. suzukii*.

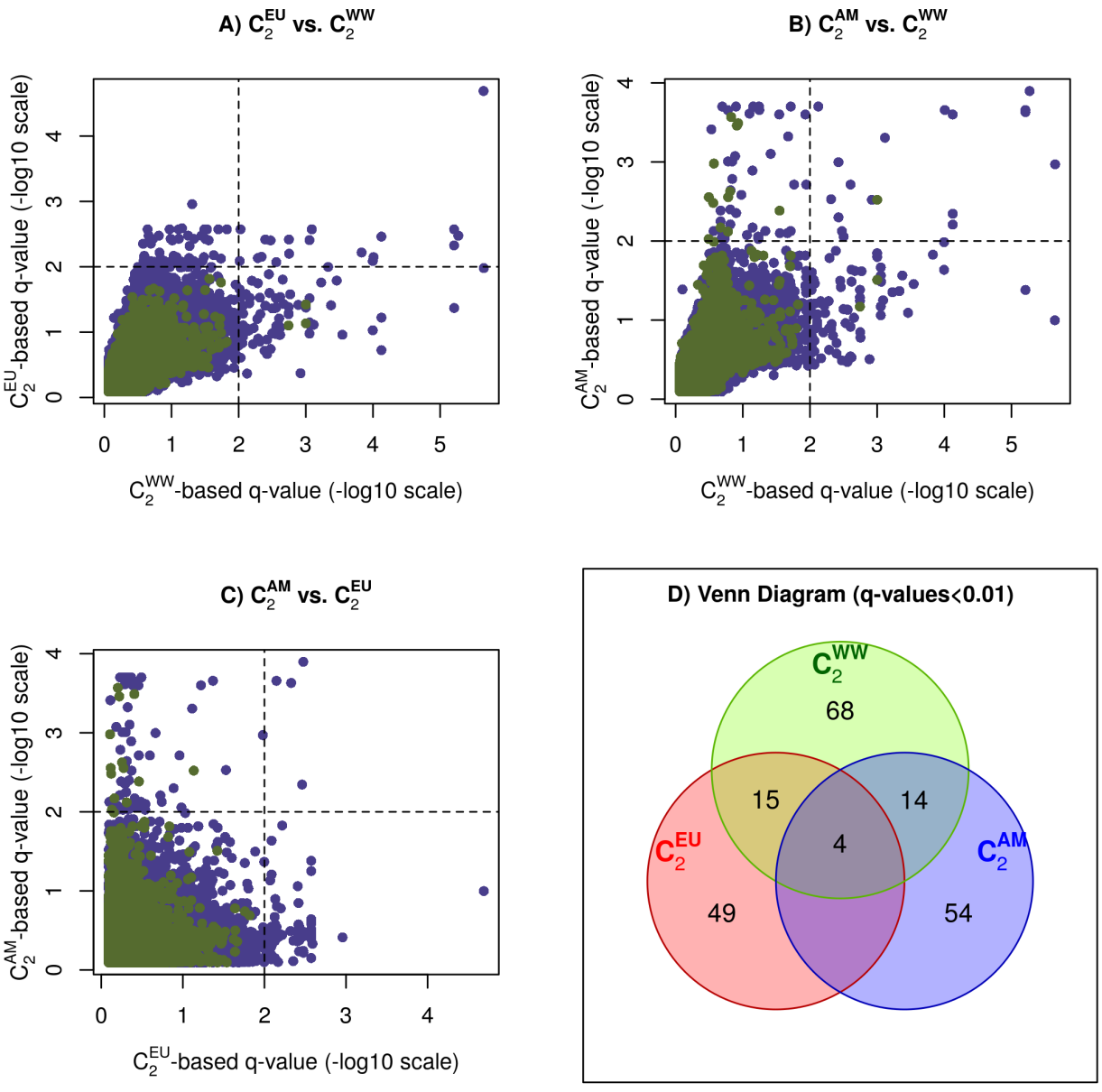

FIG. 4. Pairwise comparison of the q–values derived from the C_2^{EU} (native vs. invasive *D. suzuki* populations of the European invasion route) versus the C_2^{WW} (native vs. worldwide invasive populations) statistics A), the C_2^{AM} (native vs. invasive populations of the American invasion route) versus the C_2^{WW} statistic B), and the C_2^{AM} versus the C_2^{EU} statistics C). In A), B) and C) the dashed vertical and horizontal lines indicate the 1% q–value threshold for the C_2 derived q–values. D) Venn diagram of the number of SNPs significant at the 1% q–values among the three contrast analyses (C_2^{WW}, C_2^{EU} and C_2^{AM}). Values for the autosomal (X–linked) SNPs are plotted in purple (green).

Annotation of candidate SNPs

For annotation purposes, we relied on genomic resources available in *D. melanogaster*, a model species closely related to *D. suzukii*. More specifically we extracted from the WT3-2.0 *D. suzukii* genome assembly 5 kb long genomic

sequences surrounding each of the 204 SNPs identified above and aligned them onto the *dmel6* reference genome (Hoskins *et al.*, 2015) using the BLAT algorithm implemented in the program *pblat* (Wang and Kong, 2019). The gene annotation available from the UCSC genome browser allowed us to map 169 SNPs out of the 204 SNPs onto 130 different *D. melanogaster* genes, 145 SNPs lying within the gene sequences and 24 less than 2.5 kb apart (our predefined threshold; Table S3). Only one of the four SNPs significant for the three contrasts (C_2^{WW}, C_2^{EU} and C_2^{AM}) could not be assigned to a *D. melanogaster* gene, because its derived 5 kb long sequences aligned onto a *D. melanogaster* sequence located 10 kb away from the closest annotated gene.

Most of the 130 identified genes (80%) were represented by a single SNP, a feature in agreement with the visual lack of clustering of SNPs with strong signal already observed in the Manhattan plot (Figure 3B). It should be noticed that 14 of the 130 genes (ca. 11%) were long non-coding RNA. We however decided to focus on the 26 genes that were represented by at least two SNPs significant in one of the three contrast analyses; see Table 2 for details. The significant SNPs underlying the different genes tended to be very close, spanning a few bp (span $>$1kb for only five genes). In particular, we observed doublet variants (i.e., adjacent SNPs in complete LD) within three genes (*cpo*, *ome* and *lnc:CR45759*).

Among these 26 candidate genes, 10 and 12 might be considered as specific to the European and American invasion routes, respectively, since they did not contain any SNP significant for the alternative contrasts. Only two genes contained SNPs significant in all three contrast analyses: *RhoGEF64C* with one SNP and *cpo* with two SNPs. Such convergent signals of association with invasive status in the two independent invasion routes were particularly convincing. The median allele frequencies (computed from raw read counts) for the reference allele underlying the corresponding *RhoGEF64C* significant SNP was 0.09 (from 0.00 to 0.44) in the native populations compared to 0.93 (from 0.90 to 0.98) and 0.87 (from 0.59 to 1.00) in the invasive populations of the European and American invasion routes, respectively (Table S3). Similarly, the two SNPs significant for the three contrast analyses in the *cpo* gene actually formed a doublet with a median reference allele frequency of 0.20 (from 0.02 to 0.33) in the native populations compared to 0.99 (from 0.91 to 1.00, excluding the outlying Hawaiian population) in the invasive populations of the European and American invasion routes, respectively (Table S3). Finally, for both the genes *RhoGEF64C* and *cpo*, all *D. suzukii* extended sequences underlying the corresponding SNPs aligned within potentially rapidly evolving intronic sequences. These sequences nevertheless displayed substantial similarities with other

D. melanogaster Gene (Full Name)	Position on *dmel6* (in kb)	Number of significant SNPs			
		All C_2 (dist. in bp)	C_2^{WW}	C_2^{EU}	C_2^{AM}
Der-1 (Derlin-1)	chr2L:1,974-1,975	2 (236)	1	-	1
Gdi (GDP dissociation inhibitor)	chr2L:9,492-9,495	4 (342)	4	4	-
lncRNA:CR45693 (long non-coding RNA)	chr2L:14,51-14,512	2 (14)	2	1	-
Tpr2 (tetratricopeptide repeat protein 2)	chr2L:16,492-16,507	2 (8)	-	2	-
Ret (Ret oncogene)	chr2L:21,182-21,199	2 (70)	2	-	-
tou (toutatis)	chr2R:11,579-11,616	2 (18)	1	-	2
jeb (jelly belly)	chr2R:12,091-12,119	2 (14)	2	-	-
CG5065	chr2R:16,608-16,625	2 (13)	-	2	-
bab2 (bric a brac 2)	chr3L:1,140-1,177	2 (11189)	1	-	1
axo (axotactin)	chr3L:4,630-4,687	2 (25886)	-	1	1
RhoGEF64C (ρ guanine nucl. exch. fact. at 64C)	chr3L:4,693-4,796	2 (8)	2	1	1
CG7509	chr3L:4,803-4,805	2 (5)	-	2	-
Con (connectin)	chr3L:4,938-4,976	2 (616)	1	1	-
Ets65A (Ets at 65A)	chr3L:6,098-6,124	2 (27998)	1	1	-
lncRNA:CR45759 (long non-coding RNA)	chr3L:6,787-6,787	4 (106)	-	-	4
ome (omega)	chr3L:14,673-14,748	2 (1)	2	-	-
sa (spermatocyte arrest)	chr3L:21,405-21,407	2 (61)	1	1	-
yellow-e (yellow-e)	chr3R:13,410-13,415	3 (33)	3	-	1
cv-c (crossveinless c)	chr3R:14,392-14,482	4 (2737)	1	-	3
osa (osa)	chr3R:17,688-17,718	2 (29)	-	-	2
cpo (couch potato)	chr3R:17,944-18,016	3 (193)	3	2	3
Rh3 (rhodopsin 3)	chr3R:20,081-20,082	2 (5709)	2	1	-
Ctl2 (choline transporter-like 2)	chr3R:29,123-29,128	2 (3)	-	-	2
Syt12 (synaptotagmin 12)	chrX:13,359-13,368	3 (65)	1	-	2
Ac13E (adenylyl cyclase 13E)	chrX:15,511-15,554	4 (19)	-	-	4
Axs (abnormal X segregation)	chrX:16,680-16,684	2 (11)	-	-	2

Table 2. Description of the 26 orthologous *D. melanogaster* genes represented by at least two of the 204 SNPs found significant for one of the three contrast analyses, C_2^{WW} (6 native *vs.* 16 invasive populations), C_2^{EU} (6 native *vs.* 8 invasive populations of the European invasion route) and C_2^{AM} (6 native vs. 8 invasive populations of the American invasion route). The third column gives the overall number of significant SNPs (at the 1% q–value threshold) and their maximal spacing in bp (on the *D. suzukii* assembly). Columns 4 to 6 gives the number of significant SNPs for each of the three contrast analyses.

related drosophila species, as shown in Figure S7 for the gene *cpo*.

Discussion

We characterized the genome response of *D. suzukii* during its worldwide invasion by conducting a genome-wide scan for association with the invasive or native status of the sampled populations. To that end, we relied on the newly developed C_2 statistic that was aimed at identifying significant allele frequencies differences between two contrasting groups of populations while accounting for their overall correlation structure due to the shared population history. Our approach identified genomic regions and candidate genes most likely involved in adaptive processes underlying the invasion success of *D. suzukii*.

Overall, we found that a relatively small number of SNPs were significantly associated with the invasive status of *D. suzukii* populations. This may seem surprising since the binary trait under study (invasive versus native) is complex in the sense that numerous biological differences may characterize invasive and native populations. The invasion process itself, including the associated

selective pressures and the genetic composition of the source populations, may actually differ depending on the considered invaded areas. Hence the small number of SNPs showing strong signals of association with the invasive status may stem from the integrative nature of our analysis over a large number of invasive populations from different invasion routes. The genomic features that may be identified under this evolutionary configuration are expected to correspond to major genetic changes instrumental to invasions shared by a majority of populations. Accordingly, it is worth noting that the independent contrast analyses of the two main invasion routes (i.e. the American and the European routes) point to substantially different subsets of SNPs significantly associated with the invasive status of the populations. This suggests that the source populations and some aspects of the invasion process differ in the two invaded areas. This could also reflect the presumably polygenic nature of the traits underlying invasion success since the evolutionary trajectories of complex traits may rely on different combination of favorable genetic variants.

The availability of a high quality genome assembly of *D. suzukii* (Paris *et al.*, 2020) and a large amount of genomic resources for its sister model species *D. melanogaster* allowed identifying a set of genes associated with the invasive status of populations. A subset of those genes was associated with physiological functions and traits previously documented in *D. melanogaster*, but for most of them, functional and phenotypic studies turned out to be limited. Their putative role in explaining the invasion success thus remained largely elusive. To avoid too speculative interpretations (Pavlidis *et al.*, 2012), we will not elaborate further on the candidate genes. Yet, we did notice that long non–coding RNAs represent more than 10% (14 out of 130) of our candidate genes, a feature which may underline a critical role of variants involved in gene regulation to promote short-term response to adaptive constraints during invasion. Also, two genes *RhoGEF64C* and *cpo* contained SNPs that were found to be highly significantly associated with the invasive status in both the European and American invasion routes. While the function of the *RhoGEF64C* gene has so far not been extensively studied, several functional and phenotypic studies in other Drosophila species identified genetic variation in the *cpo* gene associated with traits possibly important for invasion success. For instance, *cpo* genetic variation was found to contribute to natural variation in diapause in *D. melanogaster* populations of a North American cline and in populations from the more distantly related species *Drosophila montana* (Kankare *et al.*, 2010; Schmidt *et al.*, 2008). Moreover, indirect action of selection on diapause, by means of genetic correlations involving *cpo* genetic variation, was found on numerous other life-history traits in *D.*

melanogaster (Schmidt and Paaby, 2008; Schmidt *et al.*, 2005). Specifically, compared to diapausing populations, non-diapausing populations had a shorter development time and higher early fecundity, but also lower rates of larval and adult survival and lower levels of cold resistance.

Both theoretical (Roughgarden, 1971) and experimental (Mueller and Ayala, 1981) evidence show that traits typical for colonization (i.e., the so-called r-traits; Charlesworth, 1994), such as a non-diapausing phenotype, are selected when a population evolves in a new habitat with low densities and low levels of competition. Common garden studies are needed to assess potential differences in key life history traits (including diapause induction and correlated traits) between native and invasive populations of *D. suzukii* and to evaluate to which extent these are related to the identified variants (including those within the *cpo* gene) differentiating the native and invasive populations of this species.

The C_2 statistic we developed in the present study appears particularly well suited to search for association with population-specific binary traits. Apart from the invasive vs. native status we studied in *D. suzukii*, numerous examples can be found where adaptive constraints may be formulated in terms of contrasting binary population features, including individual resistance or sensibility to pathogens or host-defense systems (e.g., Eoche-Bosy *et al.*, 2017), high vs. low altitude adaptation (e.g., Foll *et al.*, 2014), ecotypes of origin (e.g., Roesti *et al.*, 2015; Westram *et al.*, 2014), or domesticated vs. wild status (e.g., Alberto *et al.*, 2018). In our simulation study using the HsIMM model, the power of the C_2 statistic was similar to that of a standard BF obtained after assuming a linear relationship between the (standardized) population allele frequencies and their corresponding binary status. Yet, we found that the robustness of both statistics strongly differed according to the structuring of genetic diversity of the neutral variants. In the analyses of association with the invasion status of the neutral SNPs simulated under the two invasion scenarios we investigated, the BF was the highest for SNPs displaying homogeneous variant allele frequencies in the invasive populations and that were also common in all the populations. In this case, the assumed linear relationship of the population allele frequencies with their invasive or native group membership may result in significantly non-null regression coefficients, while the difference of the mean allele frequencies of both groups (as measured after standardization by the C_2 statistic) is not outlying. Conversely the C_2 contrast statistic was the highest for neutral variants that were rare in the native populations and for which allele frequencies were high on average in the invasive populations, but still displayed high heterogeneity in invasive populations (hence the absence of linear relationship among populations with

invasive or native group memberships). It should be noticed that these patterns were observable on raw allele frequencies since we simulated balanced population invasion histories. Because of the different behaviors of the BF criterion and the C_2 statistic, combining both metrics may globally help reducing false positive rates. Doing so may however substantially reduce power since the two metrics are similarly expected, at least to some extent, to be sensitive to different association signal at truly causal variants. We therefore chose to focus only on the C_2 statistic to identify our candidate SNPs associated to invasive status in the *D. suzukii* populations.

It is worth stressing that C_2 has several critical advantages over BF, as well as over any other decision criterion that may be derived from a parametric modeling. From a practical point of view, the C_2 estimation does not require inclusion of any other model parameters making it more robust when dealing with data sets including a small number of populations (e.g., <8 populations), the later type of data sets often leading to unstable estimates of BF (unpublished results). In addition, it may easily be derived from only a subset of the populations under study (as we did here when computing the C_2^{EU} and C_2^{AM} contrasts specific to each of the two invasion routes), while using the complete design to capture more accurate information about the shared population history. Last, the χ^2 calibration of the C_2 under the null hypothesis represents an attractive property in the context of large data sets since it allows to deal with multiple testing issues by controlling for FDR (Francois *et al.*, 2016), via, e.g., the estimation of q–values (Storey and Tibshirani, 2003).

To estimate the C_2 statistic, we needed to correct allele frequencies for population structure. To that end, we relied on the Bayesian hierarchical model implemented in the software BayPass that has several valuable properties including (i) the accurate estimation of the scaled covariance matrix of population allele frequencies ($\mathbf{\Omega}$), (ii) the integration over the uncertainty of the across population allele frequencies (π parameter), and (iii) the inclusion of additional layers of complexities such as the sampling of reads from (unobserved) allele counts in Pool-Seq data (Gautier, 2015). However, as previously mentioned, the Bayesian hierarchical modeling results in shrinking the posterior means of the (lower-level) model parameters (Kruschke, 2014) and also related statistics such as, here, the C_2 and XtX differentiation statistics. To ensure proper calibration of the two corresponding estimates, we hence needed to rely on the rescaled posterior means of the standardized allele frequencies. This empirical procedure proved efficient in providing well behaved p–values, while avoiding computationally intensive calibration procedure based on the analysis of pseudo-observed data sets simulated under the generative model (Gautier, 2015). Still, this did not allow accounting for the

uncertainty of the allele frequencies estimation (i.e., their full marginal distribution) and more importantly, it implicitly assumes exchangeability of SNPs both across the populations and along the genome. Such an assumption, which pertains to the null hypothesis of neutral differentiation only (and consequently of no association with binary population–specific covariable), might actually be viewed as conservative even in the presence of background LD across the populations, providing that a reasonably large number of SNPs is analyzed. Interestingly, the almost absence of clustering of associated SNPs we observed in the *D. suzukii* genome suggested a very limited extent of across-population LD, presumably resulting from large effective population sizes. This conversely led to a high mapping resolution. In practice, when dealing with large data sets, a sub-sampling strategy consisting in analyzing data sets thinned by marker position also allows further reduction of across-population LD (Gautier *et al.*, 2018). Finally, it should be noticed that information from LD might be at least partially recovered by combining C_2 or XtX derived p–values into local scores (Fariello *et al.*, 2017).

Other less computationally intensive (but less flexible and versatile) approaches may be considered to estimate the C_2 statistic. For instance, the C_2 statistic is closely related to the S_B statistic recently proposed by Refoyo-Martinez *et al.* (2019) to identify footprints of selection in admixture graphs. However, while the C_2 statistic relies on the full scaled covariance matrix of population allele frequencies ($\mathbf{\Omega}$), the S_B statistic relies on a covariance matrix called $\boldsymbol{F}$ (Refoyo-Martinez *et al.*, 2019) that specifies an a priori inferred admixture graph summarizing the history of the sampled populations. The covariance matrix $\boldsymbol{F}$ thus represents a simplified version of $\mathbf{\Omega}$ that may only partially capture the covariance structure of the population allele frequencies. In addition, to compute S_B, the graph root allele frequencies are estimated as the average of allele frequencies across the sampled population, which might result in biased estimates, particularly when the graph is unbalanced. Deriving the matrix $\boldsymbol{F}$ from $\mathbf{\Omega}$ (e.g., Pickrell and Pritchard, 2012) might actually allow interpreting C_2 as a Bayesian counterpart of the S_B statistic, thereby benefiting from the aforementioned advantages regarding the estimation of the parameters Ω and π and allowing proper analysis of Pool-Seq data.

Conclusion and perspectives

Our genome-wide association approach allowed identifying genomic regions and genes most likely involved in adaptive processes underlying the invasion success of *D. suzukii*. The approach can be transposed to any other invasive species, and more generally to any species models for which binary traits of interest can be defined at the population level. The major advantage of our approach is that it does not require a preliminary, often extremely laborious,

phenotypic characterization of the populations considered (for example using common garden experiments) in order to inform candidate traits for which genomic associations are sought. As a matter of fact, in our association study the populations analyzed are simply classified into two categories: invasive or native.

The functional and phenotypic interpretation of the signals obtained by our genome scan methods remains challenging. Such interpretation requires a good functional characterization of the genome of the studied species or, failing that, of a closely related species (i.e. *D. melanogaster* in our study). Following a strategy sometimes referred to as "reverse ecology" since it goes from gene(s) to phenotype(s) (Li *et al.*, 2008), it is then necessary to test and validate via quantitative genetic experiments whether the inferred candidate traits show significant differences between native and invasive populations. The functional interpretation of the statistical association results can also benefit from experimental validation approaches based on techniques using RNA interference (RNA-silencing, e.g. Janitz *et al.*, 2006) and/or genome editing approaches (e.g., Karageorgi *et al.*, 2017) targeting the identified candidate variants. Hopefully, such a combination of statistical, molecular and quantitative approaches will provide useful insights into the genomic and phenotypic responses to invasion, and by the same, will help better predict the conditions under which invasiveness can be enhanced or suppressed.

Materials and Methods

Simulation study

We used computer simulations to evaluate the performance of the novel statistical framework described in the section *New Approach*. A first set of simulated data sets were generated under the SimuPOP environment (Peng and Kimmel, 2005) using individual–based forward–in–time simulations implemented on a modified version of the code developed by de Villemereuil *et al.* (2014) for the so-called *HsIMM-C* demographic scenario. This corresponded to an highly structured isolation with migration demographic model (Figure 1A) that was divided in two successive periods: (i) a neutral divergence phase leading to the differentiation of an ancestral population into 16 populations after four successive fission events (at generations $t=50$, $t=150$, $t=200$ and $t=300$); and (ii) an adaptive phase (lasting 200 generations) during which individuals of the 16 populations were subjected to selective pressures exerted by two environmental constraints (*ec1* and *ec2*), each constraint having two possible modalities (a or b). We thus had a total of four possible environments in our simulation setting (Figure 1A).

All the simulated populations consisted of 500 diploid individuals reproducing under random-mating with non-overlapping generations. From generation $t=150$ (with four populations), the

migration rate $m_{jj\prime}$ between two populations j and $j\prime$ was set to $m_{jj\prime}=\frac{m}{2^{p-1}}$ where p is the number of populations in the path connecting k to $k\prime$ in the population tree. The migration rate between the two ancestral populations from generation $t=50$ to $t=150$ was set to $m=0.005$. For illustration purposes, some of the migration edges were displayed in Figure 1A.

Following de Villemereuil *et al.* (2014), a simulated genotyping data set consisted of 320 individuals (20 per populations) that were genotyped for 5,000 bi-allelic SNPs regularly spread along ten chromosomes of one Morgan length and with a frequency of 0.5 for the reference allele (randomly chosen) in the root population. Polygenic selection acting during the adaptive phase was simulated by choosing 50 randomly distributed SNPs (among the previous 5,000 ones) that influenced individual fitness according to either the *ec1* or *ec2* environmental constraints (with 25 SNPs for *ec1* and 25 SNPs for *ec2*).

The fitness of each individual, given its genotype, can be defined at each generation. Let $p(o)=j$ $(j=1,\ldots,16)$ denote the population of origin of individual o $(o=1,\ldots,16\times500)$, and $e_k(j)=1$ (respectively $e_k(j)=-1$) if the environmental constraint *eck* $(k=1,2)$ of population j is of type a (respectively b). Let further denote $s_i(k)$ the local selective coefficient of SNP i such that $s_i(k)=0$ if the SNP is neutral with respect to *eck* and $s_i(k)=0.01$ otherwise. The fitness of each individual o (at each generation) given its genotypes at all the SNPs is then defined using a cumulative multiplicative fitness function as:

$$w(o)=\prod_{i=1}^{I}\prod_{k=1}^{2}(1+e_k(p(o))(1-g_i(o))s_i(k)) \quad (6)$$

where $g_i(o)$ is the genotype of individual o at marker i coded as the number of the reference allele (0, 1 or 2).

In a second time, we evaluated the robustness of the new C_2 criterion as well as the BF statistic to the occurrence of bottleneck events, as the latter are expected (especially in the context of biological invasion) to strongly impact the genetic variation of populations within invaded areas and between invasive and native areas (e.g, Estoup *et al.*, 2016). To this aim, we used the software DIYABC v2.1.0 (Cornuet *et al.*, 2014) to simulate data sets composed of selectively neutral and independent SNP loci under two invasion scenarios depicted in Figure 2. The two scenarios roughly mimic the situation of the worldwide *D. suzukii* invasion (Fraimout *et al.*, 2017) by considering two invaded areas (with eight populations sampled in each area) with two independent primary bottlenecked introductions from two different native populations (among a set of six sampled native populations). The two scenarios differ by the relationships among the invasive populations within each invaded area: i) in the scenario *Inv1* each invasive population of an area derived from the same primarily introduced population with a bottleneck occurring

at different time in the past, and ii) in the scenario *Inv2* the invasive populations of an area are successively founded one after the other with a bottleneck event at each foundation, a process likely to favor allele surfing during geographic range expansion (Excoffier and Ray, 2008). A detailed description of the two invasion scenarios with values of the historical and demographical parameters used to simulate the two SNP data sets is provided in Figures 2A and 2B. We used the ‘‘`Simulate dataset + SNP option`’’ of DIYABC v2.1.0 to generate autosomal SNP genotypes (n = 50 diploid individuals sampled per population) at 250,000 loci under both the scenarios *Inv1* and *Inv2*, following the algorithm by Hudson (2002) which is equivalent to applying a default MAF threshold on the simulated data sets. As a matter of fact, each locus will be characterized by the presence of at least a single copy of a variant over all genes sampled from all studied populations (i.e. pooling all genes genotyped at the locus). We further applied a 1% MAF threshold on the simulated data sets (i.e., similar to that used on real data) before analyzing them using BayPass (hence a total of 165,020 and 152,321 SNPs analyzed with BayPass under the scenarios *Inv1* and *Inv2*, respectively).

Sampling of *D. suzukii* populations and DNA extraction

Adult *D. suzukii* flies were sampled in the field at a total of 22 localities (hereafter termed sample sites) distributed throughout most of the native and invasive range of the species (Fig 3A and Table S2). Samples were collected between 2013 and 2016 using baited traps (with a vinegar-alcohol-sugar mixture) and sweep nets, and stored in ethanol. Only four of the 22 samples were composed of flies which directly emerged in the lab from fruits collected in the field (Table S2). Native Asian samples consisted of a total of six sample sites including four Chinese and two Japanese localities. Samples from the invasive range were collected in Hawaii (1 sample site), Continental US (6 sites), Brazil (1 site), Europe (7 sites) and the French island of La Réunion (1 site). The continental US (plus Brazil) and European (plus La Réunion Island) populations are representative of two separate invasion routes (the American and European routes, respectively), with different native source populations and multiple introduction events in both invaded areas (Fraimout et al. 2017; see Table S2).

Pool sequencing

For each of the 22 sampling sites, the thoraxes of 50 to 100 representative adult flies (Table S2) were pooled for DNA extraction using the EZ-10 spin column genomic DNA minipreps kit (Bio basic inc.). Barcoded DNA PE libraries with insert size of ca. 550 bp were further prepared using the Illumina Truseq DNA Library Preparation Kit following manufacturer protocols using the 22 DNA pools samples. The DNA libraries were then validated on a DNA1000

chip on a Fragment Analyzer (Agilent) to determine size and quantified by qPCR using the Kapa library quantification kit to determine concentration. The cluster generation process was performed on cBot (Illumina) using the Paired-End Clustering kit (Illumina). Each pool DNA library was further paired-end sequenced on a HiSeq 2500 (Illumina) using the Sequence by Synthesis technique (providing 2x125 bp reads, respectively) with base calling achieved by the RTA software (Illumina). The Pool-Seq data were deposited in the Sequence Read Archive (SRA) repository under the BioProject accession number PRJNA576997.

Raw paired-end reads were filtered using *fastp* 0.19.4 (Chen *et al.*, 2018) run with default options to remove contaminant adapter sequences and trim for poor quality bases (i.e., with a phred-quality score <15). Read pairs with either one read with a proportion of low quality bases over 40% or containing more than 5 N bases were removed. Filtered reads were then mapped onto the newly released WT3-2.0 *D. suzukii* genome assembly (Paris *et al.*, 2020), using default options of the *mem* program from the *bwa* 0.7.17 software (Li, 2013; Li and Durbin, 2009). Read alignments with a mapping quality Phred-score <20 or PCR duplicates were further removed using the *view* (option -q 20) and *markdup* programs from the *SAMtools* 1.9 software (Li *et al.*, 2009), respectively.

Variant calling was then performed on the resulting *mpileup* file using *VarScan mpileup2cns* v2.3.4 (Koboldt *et al.*, 2012) (options *–min–coverage* 50 *–min–avg–qual* 20 *–min-var-freq* 0.001 *–variants-output-vcf* 1). The resulting *vcf* file was processed with the *vcf2pooldata* function from the R package *poolfstats* v1.1 (Hivert *et al.*, 2018) retaining only bi-allelic SNPs covered by >4 reads, <99.9th overall coverage percentile in each pool and with an overall MAF>0.01 (computed from read counts). In total, n=11,564,472 SNPs (respectively n=1,966,184 SNPs) SNPs mapping to the autosomal contigs (respectively X–chromosome contigs) were used for genome-wide association analysis. The median coverage per pool ranged from 58X to 88X and from 34X to 84X for autosomal and X chromosomes, respectively (Table S2). As previously described (Gautier *et al.*, 2018), the autosomal and X-chromosome data sets were divided into sub-data sets of ca. 75,000 SNPs each (by taking one SNP every 154 SNPs and one SNPs every 26 SNPs along the underlying autosomal and X–chromosome contigs, respectively).

Genome scan analyses

All genome-wide scans were performed using an upgraded version (2.2) of BayPass (Gautier, 2015) (available from `http://www1.montpellier.inra.fr/CBGP/software/baypass/`), that includes the new C_2 and XtX statistics estimated as described in the above section *New Approach*. We always used

the BayPass core model with default options for the MCMC algorithm to obtain estimates of four items: (i) the scaled covariance matrix ($\mathbf{\Omega}$); (ii) the SNP-specific XtX overall differentiation statistic in the form of both $\widehat{XtX}$, the posterior mean of XtX (Gautier, 2015) and $\widehat{XtX^{\star}}$, our newly described calibrated estimator; (iii) our novel C_2 statistic in the form of the calibrated estimator described above; and (iv) Bayes Factor reported in deciban units (db) as a measure of support for association with contrasts of each SNP based on a linear regression model (Coop *et al.*, 2010; Gautier, 2015). For BF, a value >15 db (respectively >20 db) provides very strong (respectively decisive) evidence in favor of association according to the Jeffreys' rule (Jeffreys, 1961).

For the *D. suzukii* data sets, we specified the pool haploid sample sizes, for either autosomes or the X–chromosome (Table S2), to activate the Pool-Seq mode of BayPass. The C_2^{WW} statistic for the contrast of the six native and 16 worldwide invasive populations was estimated jointly with the C_2^{EU} and C_2^{AM} statistics for the contrast of the six native and eight invasive populations of the European and American invasion routes, respectively. For these two latter estimates, this simply amounted to setting $c_j = 0$ (see eq. 3) for all population j not considered in the corresponding contrast analysis. Finally, two additional independent runs (using option *-seed*) were performed to assess reproducibility of the MCMC estimates. We found a fairly high correlation across the different independent runs (Pearson's $r > 0.92$ for autosomal and $r > 0.87$ and X–chromosome data) for the different estimators and thus only presented results from the first run. Similarly and for each chromosome type (i.e., autosomes or the X chromosome), a near perfect correlation of the posterior means of the estimated $\mathbf{\Omega}$ matrix elements was observed across independent runs as well as within each run across SNP sub-samples, with the corresponding FMD distances (Gautier, 2015) being always smaller than 0.4. We thus only reported results regarding the $\mathbf{\Omega}$ matrix that were obtained from a single randomly chosen sub-data set analysed in the first run.

Acknowledgments

We wish to thank our three anonymous reviewers for their very helpful and constructive comments. AE, MG and LO acknowledge financial support from the National Research Fund ANR (France) through the project ANR-16-CE02-0015-01 (SWING), the Languedoc-Roussillon region (France) through the European Union program FEFER FSE IEJ 2014-2020 (project CPADROL) and the INRA scientific department SPE (AAP-SPE 2016 and 2018). MGX acknowledges financial support from France Génomique National infrastructure, funded as part of "Investissement d'avenir" program managed by Agence Nationale pour la Recherche (contract ANR-10-INBS-09). We are grateful to the genotoul bioinformatics

platform Toulouse Midi-Pyrenees for providing computing resources, Nicolas Rode for useful discussions and comments on a previous version of the manuscript and Nicolas Ris, Jon Koch, Masahito Kimura, Simon Fellous, Vincent Debat, Marta Pascual, Ruth Hufbauer, Marindia Depra, Isabel Martinez, Pierre Girod and Maxi Richmond for help in collecting some of the *D. suzukii* samples.

References

Adrion, J. R., Kousathanas, A., Pascual, M., Burrack, H. J., Haddad, N. M., Bergland, A. O., Machado, H., Sackton, T. B., Schlenke, T. A., Watada, M., Wegmann, D., and Singh, N. D. 2014. Drosophila suzukii: the genetic footprint of a recent, worldwide invasion. *Molecular Biology and Evolution*, 31(12): 3148–63.

Alberto, F. J., Boyer, F., Orozco-terWengel, P., Streeter, I., Servin, B., de Villemereuil, P., Benjelloun, B., Librado, P., Biscarini, F., Colli, L., Barbato, M., Zamani, W., Alberti, A., Engelen, S., Stella, A., Joost, S., Ajmone-Marsan, P., Negrini, R., Orlando, L., Rezaei, H. R., Naderi, S., Clarke, L., Flicek, P., Wincker, P., Coissac, E., Kijas, J., Tosser-Klopp, G., Chikhi, A., Bruford, M. W., Taberlet, P., and Pompanon, F. 2018. Convergent genomic signatures of domestication in sheep and goats. *Nature Communications*, 9(1): 813.

Asplen, M. K., Anfora, G., Biondi, A., Choi, D.-S., Chu, D., Daane, K. M., Gibert, P., Gutierrez, A. P., Hoelmer, K. A., Hutchison, W. D., Isaacs, R., Jiang, Z.-L., Kárpáti, Z., Kimura, M. T., Pascual, M., Philips, C. R., Plantamp, C., Ponti, L., Vétek, G., Vogt, H., Walton, V. M., Yu, Y., Zappalà, L., and Desneux, N. 2015. Invasion biology of spotted wing drosophila (drosophila suzukii): a global perspective and future priorities. *Journal of Pest Science*, 88(3): 469–494.

Balanya, J., Oller, J. M., Huey, R. B., Gilchrist, G. W., and Serra, L. 2006. Global genetic change tracks global climate warming in drosophila subobscura. *Science*, 313(5794): 1773–1775.

Barrett, S. C. H. 2015. Foundations of invasion genetics: the baker and stebbins legacy. *Molecular Ecology*, 24(9): 1927–1941.

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

Bonhomme, M., Chevalet, C., Servin, B., Boitard, S., Abdallah, J., Blott, S., and Sancristobal, M. 2010. Detecting selection in population trees: the lewontin and krakauer test extended. *Genetics*, 186(1): 241–262.

Charlesworth, B. 1994. *Evolution in Age-Structured Populations*. Cambridge Studies in Mathematical Biology. Cambridge University Press, 2 edition.

Chen, S., Zhou, Y., Chen, Y., and Gu, J. 2018. fastp: an ultra-fast all-in-one fastq preprocessor. *Bioinformatics*, 34(17): i884–i890.

Chiu, J. C., Jiang, X., Zhao, L., Hamm, C. A., Cridland, J. M., Saelao, P., Hamby, K. A., Lee, E. K., Kwok, R. S., Zhang, G., Zalom, F. G., Walton, V. M., and Begun, D. J. 2013. Genome of drosophila suzukii, the spotted wing drosophila. *G3 (Bethesda)*, 3(12): 2257–71.

Cini, A., Ioriatti, C., and Anfora, G. 2012. A review of the invasion of drosophila suzukii in europe and a draft research agenda for integrated pest management. *Bulletin of Insectology*, 65: 149–160.

Clemente, F., Gautier, M., and Vitalis, R. 2018. Inferring sex-specific demographic history from snp data. *PLoS Genetics*, 14(1): e1007191.

Colautti, R. I. and Barrett, S. C. H. 2013. Rapid adaptation to climate facilitates range expansion of an invasive plant. *Science*, 342(6156): 364–6.

Colautti, R. I. and Lau, J. A. 2015. Contemporary evolution during invasion: evidence for differentiation,

natural selection, and local adaptation. *Molecular Ecology*, 24(9): 1999–2017.

Coop, G., Witonsky, D., Rienzo, A. D., and Pritchard, J. K. 2010. Using environmental correlations to identify loci underlying local adaptation. *Genetics*, 185(4): 1411–1423.

Cornuet, J.-M., Pudlo, P., Veyssier, J., Dehne-Garcia, A., Gautier, M., Leblois, R., Marin, J.-M., and Estoup, A. 2014. Diyabc v2.0: a software to make approximate bayesian computation inferences about population history using single nucleotide polymorphism, dna sequence and microsatellite data. *Bioinformatics*, 30(8): 1187–1189.

de Villemereuil, P. and Gaggiotti, O. E. 2015. A new FST-based method to uncover local adaptation using environmental variables. *Methods in Ecology and Evolution*, 6(11): 1248–1258.

de Villemereuil, P., Frichot, E., Bazin, E., Francois, O., and Gaggiotti, O. E. 2014. Genome scan methods against more complex models: when and how much should we trust them? *Molecular Ecology*, 23(8): 2006–2019.

Dlugosch, K. M., Anderson, S. R., Braasch, J., Cang, F. A., and Gillette, H. D. 2015. The devil is in the details: genetic variation in introduced populations and its contributions to invasion. *Molecular Ecology*, 24(9): 2095–2111.

Ellstrand, N. C. and Schierenbeck, K. A. 2000. Hybridization as a stimulus for the evolution of invasiveness in plants? *Proceedings of the National Academy of Sciences*, 97(13): 7043–7050.

Eoche-Bosy, D., Gautier, M., Esquibet, M., Legeai, F., Bretaudeau, A., Bouchez, O., Fournet, S., Grenier, E., and Montarry, J. 2017. Genome scans on experimentally evolved populations reveal candidate regions for adaptation to plant resistance in the potato cyst nematode globodera pallida. *Molecular Ecology*, 26(18): 4700–4711.

Estoup, A., Ravigne, V., Hufbauer, R., Vitalis, R., Gautier, M., and Facon, B. 2016. Is there a genetic paradox of biological invasion? *Annual Review of Ecology, Evolution, and Systematics*, 47(1): 51–72.

Excoffier, L. and Ray, N. 2008. Surfing during population expansions promotes genetic revolutions and structuration. *Trends in Ecology and Evolution*, 23(7): 347–51.

Facon, B., Hufbauer, R. A., Tayeh, A., Loiseau, A., Lombaert, E., Vitalis, R., Guillemaud, T., Lundgren, J. G., and Estoup, A. 2011. Inbreeding depression is purged in the invasive insect harmonia axyridis. *Current Biology*, 21(5): 424–7.

Fariello, M. I., Boitard, S., Mercier, S., Robelin, D., Faraut, T., Arnould, C., Recoquillay, J., Bouchez, O., Salin, G., Dehais, P., Gourichon, D., Leroux, S., Pitel, F., Leterrier, C., and SanCristobal, M. 2017. Accounting for linkage disequilibrium in genome scans for selection without individual genotypes: The local score approach. *Molecular Ecology*, 26(14): 3700–3714.

Foll, M., Gaggiotti, O. E., Daub, J. T., Vatsiou, A., and Excoffier, L. 2014. Widespread signals of convergent adaptation to high altitude in asia and america. *American Journal of Human Genetics*, 95(4): 394–407.

Fraimout, A., Debat, V., Fellous, S., Hufbauer, R. A., Foucaud, J., Pudlo, P., Marin, J.-M., Price, D. K., Cattel, J., Chen, X., Depra, M., Duyck, P. F., Guedot, C., Kenis, M., Kimura, M. T., Loeb, G., Loiseau, A., Martinez-Sanudo, I., Pascual, M., Richmond, M. P., Shearer, P., Singh, N., Tamura, K., Xuéreb, A., Zhang, J., and Estoup, A. 2017. Deciphering the routes of invasion of drosophila suzukii by means of abc random forest. *Molecular Biology and Evolution*, 34(4): 980–996.

Francois, O., Martins, H., Caye, K., and Schoville, S. D. 2016. Controlling false discoveries in genome scans for selection. *Molecular Ecology*, 25(2): 454–69.

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

Frichot, E., Schoville, S. D., de Villemereuil, P., Gaggiotti, O. E., and Francois, O. 2015. Detecting adaptive evolution based on association with ecological gradients: Orientation matters! *Heredity*.

Gautier, M. 2015. Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics*, 201(4): 1555–79.

Gautier, M., Foucaud, J., Gharbi, K., Cezard, T., Galan, M., Loiseau, A., Thomson, M., Pudlo, P., Kerdelhue, C., and Estoup, A. 2013. Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. *Molecular Ecology*, 22(14): 3766–3779.

Gautier, M., Yamaguchi, J., Foucaud, J., Loiseau, A., Ausset, A., Facon, B., Gschloessl, B., Lagnel, J., Loire, E., Parrinello, H., Severac, D., Lopez-Roques, C., Donnadieu, C., Manno, M., Berges, H., Gharbi, K., Lawson-Handley, L., Zang, L.-S., Vogel, H., Estoup, A., and Prud'homme, B. 2018. The genomic basis of color pattern polymorphism in the harlequin ladybird. *Current Biology*, 28(20): 3296–3302.e7.

Grau, J., Grosse, I., and Keilwagen, J. 2015. Prroc: computing and visualizing precision-recall and receiver operating characteristic curves in r. *Bioinformatics*, 31(15): 2595–7.

Gunther, T. and Coop, G. 2013. Robust identification of local adaptation from allele frequencies. *Genetics*, 195(1): 205–220.

Hivert, V., Leblois, R., Petit, E. J., Gautier, M., and Vitalis, R. 2018. Measuring genetic differentiation from pool-seq data. *Genetics*, 210(1): 315–330.

Hoskins, R. A., Carlson, J. W., Wan, K. H., Park, S., Mendez, I., Galle, S. E., Booth, B. W., Pfeiffer, B. D., George, R. A., Svirskas, R., Krzywinski, M., Schein, J., Accardo, M. C., Damia, E., Messina, G., Méndez-Lago, M., de Pablos, B., Demakova, O. V., Andreyeva, E. N., Boldyreva, L. V., Marra, M., Carvalho, A. B., Dimitri, P., Villasante, A., Zhimulev, I. F., Rubin, G. M., Karpen, G. H., and Celniker, S. E. 2015. The release 6 reference sequence of the drosophila melanogaster genome. *Genome Research*, 25(3): 445–58.

Hudson, R. R. 2002. Generating samples under a wright-fisher neutral model of genetic variation. *Bioinformatics*, 18(2): 337–338.

Janitz, M., Vanhecke, D., and Lehrach, H. 2006. *High-Throughput RNA Interference in Functional Genomics*, page 97–104. Springer Berlin Heidelberg, Berlin, Heidelberg.

Jeffreys, H. 1961. *Theory of Probability*. Oxford University Press, 3rd edition.

Kankare, M., Salminen, T., Laiho, A., Vesala, L., and Hoikkala, A. 2010. Changes in gene expression linked with adult reproductive diapause in a northern malt fly species: a candidate gene microarray study. *BMC Ecol.*, 10: 3.

Karageorgi, M., Bracker, L. B., Lebreton, S., Minervino, C., Cavey, M., Siju, K. P., Kadow, I. C. G., Gompel, N., and Prud'homme, B. 2017. Evolution of multiple sensory systems drives novel egg-laying behavior in the fruit pest drosophila suzukii. *Curr. Biol.*, 27(6): 847–853.

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L., and Wilson, R. K. 2012. Varscan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, 22(3): 568–76.

Kruschke, J. 2014. *Doing Bayesian Data Analysis, Second Edition: A Tutorial with R, JAGS, and Stan*. Academic Press, Amsterdam, 2 edition edition.

Lee, C. E. 2002. Evolutionary genetics of invasive species. *Trends in Ecology and Evolution*, 17(8): 386–391.

Lee, C. E. and Gelembiuk, G. W. 2008. Evolutionary origins of invasive populations. *Evolutionary Applications*, 1(3): 427–448.

Li, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv*, 1303.3997.

Li, H. and Durbin, R. 2009. Fast and accurate short read alignment with burrows-wheeler transform.

Bioinformatics, 25(14): 1754–60.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. 2009. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16): 2078–9.

Li, Y. F., Costello, J. C., Holloway, A. K., and Hahn, M. W. 2008. "reverse ecology" and the power of population genomics. *Evolution*, 62(12): 2984–2994.

Mueller, L. D. and Ayala, F. J. 1981. Trade-off between r-selection and k-selection in drosophila populations. *Proceedings of the National Academy of Sciences*, 78(2): 1303–5.

Ochocki, B. M. and Miller, T. E. X. 2017. Rapid evolution of dispersal ability makes biological invasions faster and more variable. *Nature Communications*, 8: 14315.

Ometto, L., Cestaro, A., Ramasamy, S., Grassi, A., Revadi, S., Siozios, S., Moretto, M., Fontana, P., Varotto, C., Pisani, D., Dekker, T., Wrobel, N., Viola, R., Pertot, I., Cavalieri, D., Blaxter, M., Anfora, G., and Rota-Stabelli, O. 2013. Linking genomics and ecology to investigate the complex evolution of an invasive drosophila pest. *Genome Biology and Evolution*, 5(4): 745–57.

Paris, M., Boyer, R., Jaenichen, R., Wolf, J., Karageorgi, M., Green, J., Cagnon, M., Parinello, H., Estoup, A., Gautier, M., Gompel, N., and Prud'homme, B. 2020. Near-chromosome level genome assembly of the fruit pest drosophila suzukii using long-read sequencing. *BiorXiv*, 2020.01.02.892844.

Pavlidis, P., Jensen, J. D., Stephan, W., and Stamatakis, A. 2012. A critical assessment of storytelling: gene ontology categories and the importance of validating genomic scans. *Molecular Biology and Evolution*, 29(10): 3237–48.

Peng, B. and Kimmel, M. 2005. simupop: a forward-time population genetics simulation environment. *Bioinformatics*, 21(18): 3686–3687.

Pickrell, J. K. and Pritchard, J. K. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics*, 8(11): e1002967.

Puzey, J. and Vallejo-Marin, M. 2014. Genomics of invasion: diversity and selection in introduced populations of monkeyflowers (mimulus guttatus). *Molecular Ecology*, 23(18): 4472–85.

Refoyo-Martinez, A., da Fonseca, R. R., Halldórsdóttir, K., Arnason, E., Mailund, T., and Racimo, F. 2019. Identifying loci under positive selection in complex population histories. *Genome Research*, 29(9): 1506–1520.

Reznick, D. N., Losos, J., and Travis, J. 2019. From low to high gear: there has been a paradigm shift in our understanding of evolution. *Ecology Letters*, 22(2): 233–244.

Roesti, M., Kueng, B., Moser, D., and Berner, D. 2015. The genomics of ecological vicariance in threespine stickleback fish. *Nature Communications*, 6: 8767.

Roughgarden, J. 1971. Density-dependent natural selection. *Ecology*, 52(3): 453–468.

Schlotterer, C., Tobler, R., Kofler, R., and Nolte, V. 2014. Sequencing pools of individuals - mining genome-wide polymorphism data without big funding. *Nature Review Genetics*, 15(11): 749–763.

Schmidt, P. S. and Paaby, A. B. 2008. Reproductive diapause and life-history clines in north american populations of drosophila melanogaster. *Evolution*, 62(5): 1204–15.

Schmidt, P. S., Matzkin, L., Ippolito, M., and Eanes, W. F. 2005. Geographic variation in diapause incidence, life-history traits, and climatic adaptation in drosophila melanogaster. *Evolution*, 59(8): 1721–32.

Schmidt, P. S., Zhu, C.-T., Das, J., Batavia, M., Yang, L., and Eanes, W. F. 2008. An amino acid polymorphism in the couch potato gene forms the basis for climatic adaptation in drosophila melanogaster. *Proceedings of the National Academy of Sciences*, 105(42): 16207–11.

Storey, J. D. and Tibshirani, R. 2003. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16): 9440–5.

Wang, M. and Kong, L. 2019. pblat: a multithread blat algorithm speeding up aligning sequences to genomes. *BMC Bioinformatics*, 20(1): 28.

Weir, B. S. and Cockerham, C. C. 1984. Estimating f-statistics for the analysis of population structure. *Evolution*, 38(6): 1358–1370.

Welles, S. and Dlugosch, K. 2018. *Population Genomics of Colonization and Invasion*, page 1–29.

Westram, A. M., Galindo, J., Rosenblad, M. A., Grahame, J. W., Panova, M., and Butlin, R. K. 2014. Do the same genes underlie parallel phenotypic divergence in different littorina saxatilis populations? *Molecular Ecology*, 23(18): 4603–16.

Williams, J. L., Kendall, B. E., and Levine, J. M. 2016. Rapid evolution accelerates plant population spread in fragmented experimental landscapes. *Science*, 353(6298): 482–485.

Wu, N., Zhang, S., Li, X., Cao, Y., Liu, X., Wang, Q., Liu, Q., Liu, H., Hu, X., Zhou, X. J., James, A. A., Zhang, Z., Huang, Y., and Zhan, S. 2019. Fall webworm genomes yield insights into rapid adaptation of invasive species. *Nature Ecology and Evolution*, 3(1): 105–115.