Individual-based simulations of genome evolution with ancestry

Janzen, Thijs; Diaz, Fernando

Published in:
Methods in ecology and evolution

DOI:
10.1111/2041-210X.13612

Publication date:
2021

Received: 17 October 2020 | Accepted: 22 March 2021

DOI: 10.1111/2041-210X.13612

APPLICATION

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

Individual-based simulations of genome evolution with ancestry: The GenomeAdmixR R package

Thijs Janzen[1,2] | Fernando Diaz[3]

[1]Groningen Institute for Evolutionary Life Sciences, University of Groningen, Groningen, The Netherlands

[2]Carl von Ossietzky University, Oldenburg, Germany

[3]Department of Entomology, University of Arizona, Tucson, AZ, USA

Correspondence
Thijs Janzen
Email: t.janzen@rug.nl

Handling Editor: Giovanni Strona

Abstract

1. Hybridization between populations or species results in a mosaic of the two parental genomes. This and other types of genome admixture have received increasing attention for their implications in speciation, human evolution, Evolve and Resequence (E&R) and genetic mapping. However, a thorough understanding of how local ancestry changes after admixture and how selection affects patterns of local ancestry remains elusive. The complexity of these questions limits analytical treatment, but these scenarios are specifically suitable for simulation.
2. Here, we present the R package GenomeAdmixR, which uses an individual-based model to simulate genomic patterns following admixture forward in time. GenomeAdmixR provides user-friendly functions to set up and analyse simulations under evolutionary scenarios with selection, linkage and migration.
3. We show the flexible functionality of the GenomeAdmixR workflow by demonstrating (a) how to design an E&R simulation using GenomeAdmixR and (b) how to use GenomeAdmixR to verify analytical expectations following from the theory of junctions.
4. GenomeAdmixR provides a mechanistic approach to explore expected genome responses to realistic admixture scenarios. With this package, we aim to aid researchers in testing specific hypotheses based on empirical findings involving admixing populations.

KEYWORDS

admixture, ancestry, Evolve and Resequence, genome evolution, individual-based modelling

1 | INTRODUCTION

Genetic exchange has long been recognized as an important driver of genetic diversity, from the recombination of conspecific genomes and the evolution of speciation with migration (Abbott et al., 2013) to introgressive hybridization (Janzen et al., 2018; Lavretsky et al., 2019) and horizontal gene transfer (Keeling & Palmer, 2008). With recent advances in sequencing technologies, the view of genetic exchange has now moved from information at specific loci (e.g. alleles, genes) to the level of whole-genome admixture (Chafin & Douglas, 2020; Lavretsky et al., 2019; Leitwein et al., 2018). The focus is now on understanding how major drivers of evolution (i.e. recombination, selection and migration) shape the linear connection of loci along the genome (i.e. synteny and linkage) and how these processes evolve under particular scenarios of genome divergence. Understanding these dynamics has helped to elucidate the impact of migration on human evolution (Hellenthal et al., 2014; Payseur & Rieseberg, 2016), the impact of hybridization on speciation (Schumer,

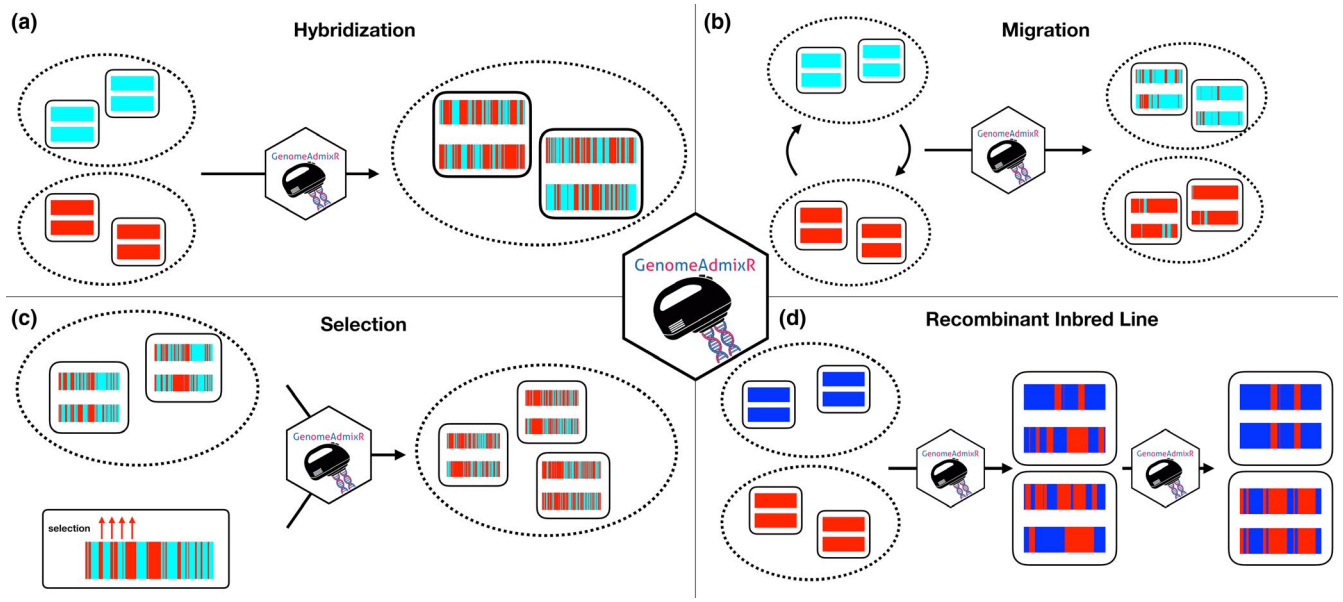

FIGURE 1 Overview of applications of the GenomeAdmixR package. The GenomeAdmixR package can be used to (a) simulate admixture of genomes (rectangular, coloured) over time, passed on between individuals (rounded squares) when two populations (dotted lines) meet. (b) Simulate ongoing genetic exchange under migration, (c) explore the impact of selection for alleles from the red ancestor in the first section of the genome and (d) Simulate the formation of Recombinant Inbred Lines, where continued inbreeding generates fully homozygous

Rosenthal, et al., 2014; Schumer et al., 2018), and helped as well in designing experiments relying on admixture, for instance Evolve and Resequence (E&R) experiments (Barghi & Schlötterer, 2019; Franssen et al., 2017; Otte & Schlötterer, 2021).

Recombination is one of the main drivers shaping macro-genomic patterns after admixture, where contiguous blocks of ancestry are broken down into smaller blocks over time (Fisher, 1954, 1959). As time progresses, genomes transform into macro-genomic mosaics consisting of many small contiguous ancestry blocks. Our understanding of how these mosaics form is relatively robust, but analytical treatments are only known for boundary cases, such as admixture between two distant populations, or for restricted backcrossing (Fisher, 1954, 1959; Janzen et al., 2018; Lavretsky et al., 2019; Macleod et al., 2005; Stam, 1980).

Linkage Disequilibrium blocks can also be generated by selection through genetic hitchhiking or selection of epistatic interactions, which can alter genome-wide patterns of genetic variation, population dynamics and evolvability (Arnold & Kunte, 2017; Gerrish et al., 2007; Zhou et al., 2017). Thus, macro-genomic patterns can be driven both by recombination and selection alike. However, a mathematical treatment of the interaction between selection, recombination and admixture is currently lacking. In contrast, individual-based simulations are very suitable to explore these interactions, for instance using packages such as MimicrEE2 (Vlachos & Kofler, 2018; written in Java) and SLiM (Haller & Messer, 2017, 2018; written in C++). Unfortunately, these do not explicitly track local ancestry, and information on haplotypes is often only indirectly available or requires extensive post-processing. forqs (Kessner & Novembre, 2014) and SELAM (Corbett-Detig & Jones, 2016; both written in C++) remedy this issue and are specifically focused on tracking local ancestry and inheritance of ancestry blocks; however, these lack ease of use (requiring, e.g., local compilation of C++ code and configuration of input files) and are not easily applied cross-platform. The R package plmgg (Cottin et al., 2020) is easily used cross-platform, but focuses mainly on plant-like admixture, including selfing.

Here, we present GenomeAdmixR, a software package written in the R language. The R language is readily used within biology and can easily be used cross-platform. GenomeAdmixR provides routines to simulate admixture in a host of scenarios (Figure 1) and includes a wide range of routines available to analyse and visualize simulation results.

2 | MATERIALS AND METHODS

2.1 | Description

The interface of GenomeAdmixR is written in the R programming language (R Core Team, 2020), with the underlying simulation code using C++, integrated using Rcpp (Eddelbuettel & Francois, 2011). Similar to SliM (Haller & Messer, 2017, 2018) and SELAM (Corbett-Detig & Jones, 2016), GenomeAdmixR simulates a population forward in time using a Wright–Fisher model with non-overlapping generations and a constant population size. GenomeAdmixR simulates diploid individuals that sexually reproduce. For computational tractability, only one pair of chromosomes is simulated, and all individuals are assumed to be hermaphroditic. Recombination is modelled in the same way as in, for instance, SLiM and SELAM (Corbett-Detig & Jones, 2016; Haller & Messer, 2018), that is, as a Poisson process, with the number of crossovers Poisson distributed with the size of the chromosome in Morgan as rate parameter. The location of crossovers is drawn from a uniform distribution, without interference.

The package offers two different implementations of this simulation scheme. First, the user can use the ***Ancestry module*** of GenomeAdmixR to perform the simulations using known local ancestry, where the simulation tracks the locations of changes in local ancestry (like SELAM and forqs) but does not explicitly model nucleotides or mutation (as mutation between ancestries complicates the simplifying assumptions of junctions' theory used to propagate the simulation). Second, the ***Sequence module*** can be used to perform simulations starting with sequencing data. Recombination is provided in cM/Mb and simulations can therefore be performed on a section rather than the entire chromosome. When using the *Sequence module*, mutation options are included as well. Sequencing data can be loaded from VCF or PLINK format (using the function read_input_data). To demonstrate the functionality of the package, we have included sample data from the *Drosophila melanogaster* Reference Panel (Huang et al., 2014; MacKay et al., 2012). The sample data consist of 5,000 SNPs with minimal allele frequency of 0.05, located along the 3R chromosome arm.

2.2 | Usage

The GenomeAdmixR package can be installed from CRAN:

```
> install.packages("GenomeAdmixR")
```

The core of GenomeAdmixR is formed by the function simulate_admixture, customizable with two modules: the ***Ancestry module*** and the ***Sequence module*** (Table 1). Here we show how to implement simulations using both modules side-by-side. To simulate a simple admixture scenario where two distinct populations hybridize (e.g. a population of admixed individuals resulting from a single mating event between two unrelated individuals, resulting in an exactly 50/50 mixing of ancestral genomes in the first generation) to form a population, which continues to admix for 100 generations, we write:

Ancestry module	*Sequence module*
> simulated_population <-	> data("dgrp2.3R.5k.data")
simulate_admixture(module =	> simulated_population <-
ancestry_module(number_of_ founders =	simulate_admixture_ data(module =
2), pop_size = 1000, total_ runtime =	sequence_module(input_data =
100)	dgrp2.3R.5k.data), pop_size = 1000,
	total_runtime = 100)

2.3 | Molecular markers and statistic estimations

GenomeAdmixR includes the functionality to track molecular markers located along the genome, in line with molecular methods (Dennenmoser et al., 2019; Lavretsky et al., 2019; Schumer, Cui, et al. 2014). For the *ancestry module*, markers also serve to simulate the uncertainty in tracking local ancestry due to limited coverage. The markers are assumed to be on a fixed position (in Morgan for the *ancestry module*, and in bp for the *sequence module*) and are tracked over time. Resulting local information is returned from the function in long format, facilitating downstream analyses. Repeating the previous scenario, but now tracking 1,000 markers at fixed positions along the genome:

Ancestry module	*Sequence module*
> simulated_population <-	> simulated_population <-
simulate_admixture(module =	simulate_admixture_data(module =
ancestry_module(number_of_ founders = 2,	sequence_module(input_data =
markers = seq(0, 1, length. out =	dgrp2.3R.5k.data, markers =
1000)), pop_size = 1000, total_runtime	dgrp2.3R.5k. data$markers[1:1000]),
= 100)	pop_size = 1000, total_runtime = 100)

Based on these markers, a set of genetic diversity statistics can be calculated, including allele frequency (calculate_allele_frequencies, Table 1), heterozygosity (calculate_heterozygosity, Table 1), linkage disequilibrium (calculate_ld, Table 1) and F_{ST} (F_{ST} is estimated following Weir & Cockerham, 1984, using the R package hierfstat, Goudet, 2005 for the implementation, see calculate_fst, Table 1), as well as changes of these over generations for better assessing significance of simulated evolution. When using the *ancestry module*, caution should be taken when using more than four distinct ancestries, because most genetic diversity statistics assume a maximum nucleotide number of four.

2.4 | Demonstration of genetic diversity statistics

We estimate linkage disequilibrium (LD) and average heterozygosity at different points in time. Our expectation is for LD and heterozygosity to decrease over time. In the code below, we either use 100 regularly spaced markers in Morgan (*ancestry module*) or we draw 100 random marker positions from the data (*sequence module*).

Ancestry module	*Sequence module*
> markers <- seq(from = 0, to = 1, length.out =	> markers <- sort(sample(dgrp2. 3R.5k.data$markers,
100)	100))
> pop_10 <- simulate_ admixture(module =	> pop_10 <- simulate_admixture_ data(module =
ancestry_module(number_ of_founders = 2, markers =	sequence_module(input_data = dgrp2.3R.5k.data,

TABLE 1 Overview of available functions in the package

Simulation	Input	Output	Summary
simulate_admixture	- chosen module (ancestry_module or sequence_module) - population size - number of generations to simulate - migration settings (generated with function migration_settings) - selection matrix (see text) - number of threads	- Admixed population - Tracked allele frequencies through time	Forward simulate a population
ancestry_module	- input population - number of unique ancestors (if no input population was given) - initial frequency of each unique ancestor - size of chromosome in Morgan - a vector with locations of molecular markers	Ancestry module object	
sequence_module	- molecular data - initial frequencies - size of chromosome in morgan - recombination rate (if not an entire chromosome is modelled) - a vector with locations of molecular markers - mutation rate - substitution matrix	Sequence module object	
migration_settings	- migration rate - option to stop at critical FST value - critical FST value - population size vector for both populations - initial frequencies (for ancestry_module simulation) - generations between measure of FST - number of individuals sampled to calculate FST - number of markers used to calculate FST	Migration settings object	
Statistical methods			
calculate_heterozygosity	- input population - vector with locations of markers	Population average heterozygosity	Calculates average heterozygosity
calculate_fst	- two input populations - number of individuals sampled to calculate the statistic - number of markers sampled to calculate the statistic	Fst value	Calculates the Weir and Cockerham Fst for a set of markers
calculate_LD	- input population - number of individuals sampled to calculate the statistic - number of markers sampled to calculate the statistic	- Pairwise linkage disequilibrium values for all markers - Pairwise R^2 values for all markers	Calculates all pairwise LD and R^2 values for a set of markers
calculate_marker_frequency	- input population - locations of markers	A table containing the frequency of each ancestor at the provided marker locations	Estimates the frequency of ancestors at a marker location within the population
calculate_allele_frequencies			Estimates the frequency of ancestors in the population for sliding windows across the genome

(Continues)

TABLE 1 (Continued)

Simulation	Input	Output	Summary
Visualization			
plot.individual	- genome of an individual (result of simulate_admixture)	Visualization of ancestry along both chromosomes, with different ancestry represented by different colours	Plots both chromosomes of an individual, where colours represent ancestors
plot_chromosome	- chromosome of an individual	Visualization of ancestry along one chromosome, with different ancestry represented by different colours	Visualize ancestry along a chromosome using different colours for each ancestor
plot_joyplot_frequencies	- tracked frequencies over time (output object of simulate_admixture) - vector of timepoints to include in the plot - selection of ancestors to plot	The so-called 'joyplot' or 'ridge plot', which visualizes density distributions over time	Plots the distribution of genomic frequencies within a region as a joyplot, where the change over time in the distribution is visualized
plot_start_end	- output of simulate_admixture - selection of ancestors to plot	Plots both the initial and final frequency of each ancestor along the genome	Plot the distribution of genomic frequencies at the start and at the end of a simulation together in one plot
plot_difference_frequencies			Plots the difference in frequency, compared between the start and the end
plot_frequencies	- output of simulate_admixture - a vector with marker locations	Plots the frequency of each ancestor at the end of the simulation	Plot the frequency of each ancestor along the genome
plot_over_time	- output of simulate_admixture - focal marker to focus on	Plot ancestor frequencies over time	Plot the frequency of each ancestor at a specific marker, over time
Utilities			
save_population	- population object - file name		Save a population to file
load_population	- file name	Population object	Load a population from file
read_input_data	- file names - type of data (vcf or plink) - chosen chromosome - number of snps	Genomeadmixr_data object	Read sequencing data from file
write_plink	- population object - marker locations - file name prefix - recombination rate		Writes output in PLINK format to file

Ancestry module	*Sequence module*
markers), total_runtime = 10, pop_size = 1000)	markers = markers), pop_size = 100, total_runtime = 10)
> ld_10 <- calculate_ld(pop_10, markers)	> ld_10 <- calculate_ld(pop_10, markers)
> het_10 <- calculate_heterozygosity(pop_10$population, locations = markers)	> het_10 <- calculate_heterozygosity(pop_10$population, locations = markers)
> pop_100 <- simulate_admixture(module = ancestry_module(input_population = pop_10, markers	> pop_100 <- simulate_admixture_data(module = sequence_module(input_data = pop_10, markers =

Ancestry module	*Sequence module*
= markers), pop_size = 100, total_runtime = 90)	markers), pop_size = 100, total_runtime = 90)
> ld_100 <- calculate_ld(pop_100, markers)	> ld_100 <- calculate_ld(pop_100, markers)
> het_100 <- calculate_heterozygosity(pop_100$population, locations = markers)	> het_100 <- calculate_heterozygosity(pop_100$population, locations = markers)

The results (Figure 2) show that LD (here plotted as the correlation (R^2) between loci) initially is relatively high, especially for closely located loci. However, as time increases, LD disappears. Similarly, initially heterozygosity is high (as expected), but decreases over time across the genome, until most markers are fixed. Patterns are similar

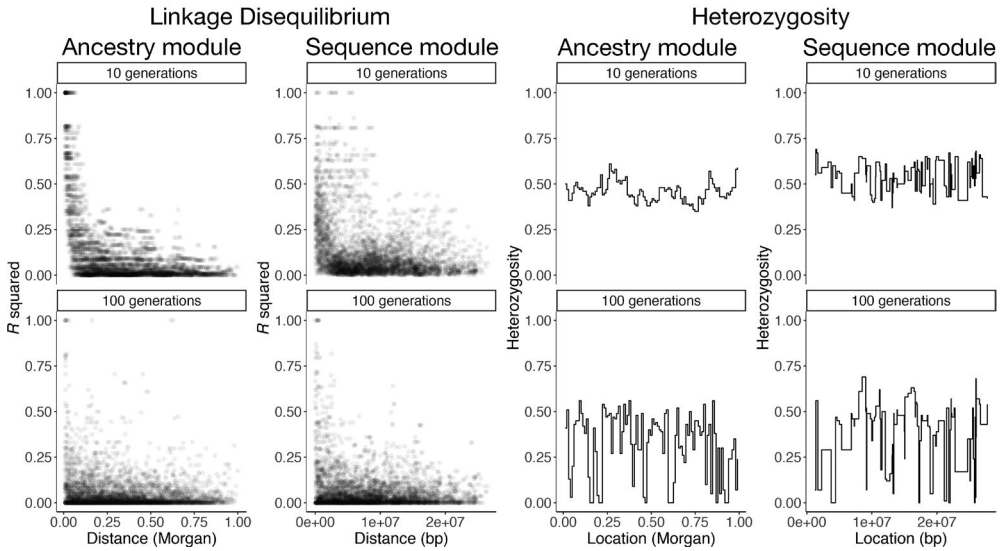

FIGURE 2 Plot showing Linkage Disequilibrium (R^2; left) and average heterozygosity (right) for a population 10 and 100 generations after admixture, where the initial admixture event involved two unrelated populations. Shown are results using the *ancestry module* (left columns) and the *sequence module* (right columns)

across the *ancestry* and *sequence* simulations, although initially, LD is slightly higher for the *sequence* simulations.

2.5 | Visualization

simulate_admixture by default returns the frequencies of all ancestors at the start and the end of the simulation, if markers are specified. Allele frequency changes can be visualized using plot_difference_frequencies_ (Table 1; Figure 3) or plot_start_end (Table 1; Figure 3) and using plot_frequencies (Table 1; Figure 3) and the final frequencies are plotted. The functions plot_over_time (Table 1; Figure 3) and plot_joyplot_frequencies (Table 1; Figure 3) provide functionality to show allele frequency changes over time. All plotting functions return ggplot2 (Wickham, 2009) objects, which can be further customized by the user.

2.6 | Migration

simulate_admixture can be extended further by including migration, where the specific settings for migration can be specified using the function 'migration_settings' (Table 1). Migration is simulated by evolving two independent populations in parallel, given a fraction of migrants each generation.

Ancestry module	*Sequence module*
> migr_pop <- simulate_admixture(	> migr_pop <- simulate_admixture
module = ancestry_module(),	(module =
total_runtime = 100,	sequence_module(input_data_population_1
migration = migration_settings(pop_size	= dgrp2.3R.5k.data,
= c(100, 100), migration_rate = 0.01))	input_data_population_2 =
	dgrp2.3R.5k.data), total_runtime = 100,
	migration = migration_settings
	(pop_size = c(100, 100), migration_rate
	= 0.01))

Then, we can analyse genetic divergence across these two populations using the F_{ST} statistic:

Ancestry module	*Sequence module*
> calculate_fst(migr_pop$population_1,	> calculate_fst(migr_pop$population_1,
migr_pop$population_2)	migr_pop$population_2)

Functionality is included in the package to stop the simulation once a threshold of genetic differentiation (i.e. F_{ST}) between the two populations is reached, which facilitates scenarios where two related but genetically different populations are required.

2.7 | Selection

simulate_admixture provides options to impose fitness benefits upon individuals that contain a marker under selection (where selection is interpreted as favouring alleles from a specific ancestor [*ancestry module*] or SNP [*sequence module*]). We follow

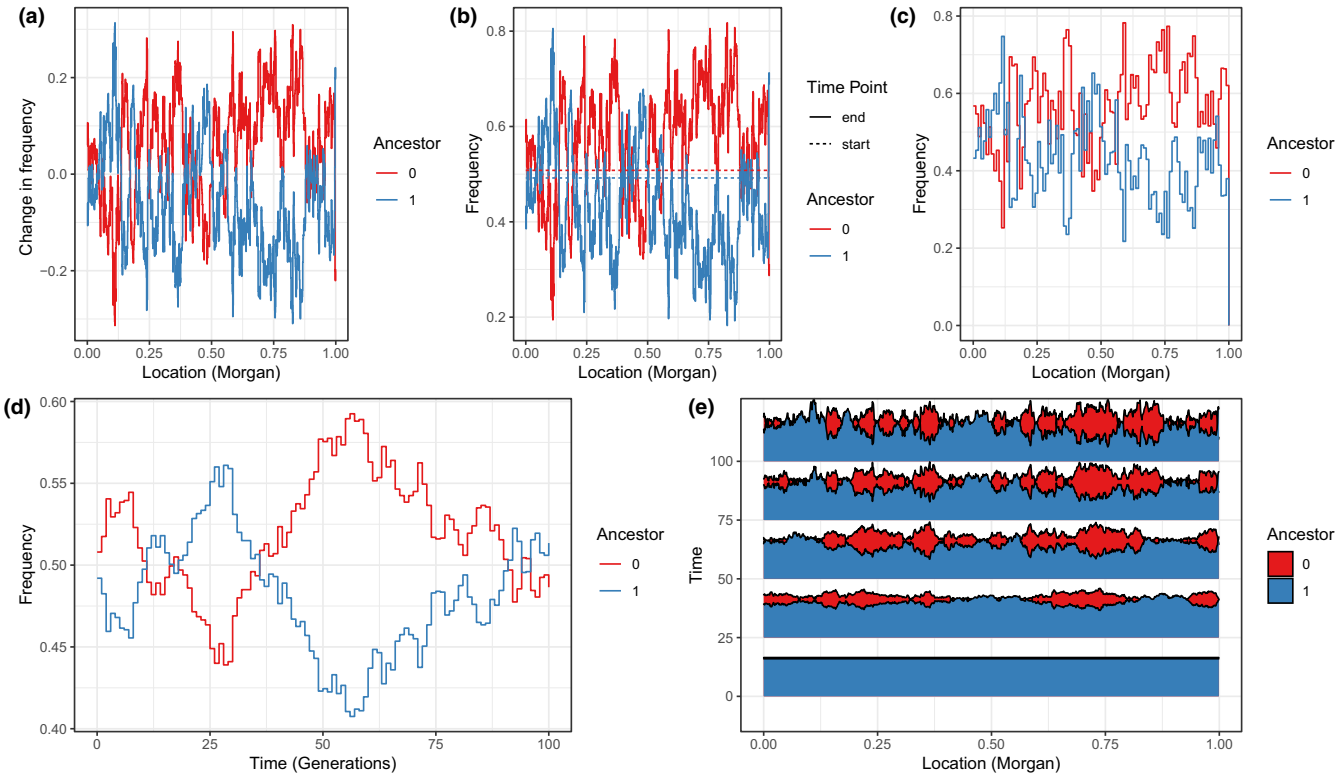

FIGURE 3 Example visualizations available in GenomeAdmixR. Shown are example plots after a hybridization scenario using the *ancestry module* between two distinct ancestral populations, parameter values used are: population size = 10,000, total runtime = 100. 1,000 markers evenly spaced in [0, 1] were used. (a) plot_difference_frequencies, which visualizes the average change in frequency between the start and end of the simulation (b) plot_start_end, which plots the average frequencies per ancestor, both at the start and the end of the simulation. (c) plot_frequencies, which plots the average frequency of each ancestor across the genome (d) plot_over_time, which plots the average frequency of a single marker (here a marker at 50 cm) in the population, over time. (e) plot_joyplot_frequencies, which visualizes the average frequency over time

conventional fitness notation (Crow & Kimura, 1970) and denote $[w_{aa}, w_{Aa}, w_{AA}] = [1, 1 + hs, 1 + s]$, where w indicates the fitness, s indicates the selective benefit of allele A and h indicates the degree of dominance. Parents of offspring in the next generation are drawn from the parental population proportional to fitness.

The selection matrix (location, w_{aa}, w_{Aa}, w_{AA}, ancestor) includes information on the selected marker, the different fitness weights and the origin of the allele under selection. The weights can be freely provided by the user, allowing for the implementation of a wide range of potential selective benefits upon receiving a marker, including overdominance and epistasis.

Furthermore, the user is not restricted to providing selection on a single marker and can expand the selection matrix by adding extra rows for each marker. This opens up the possibility for the user to explore polygenic selection (as for instance for a Quantitative Trait Locus), by specifying several loci with defined distances, where each locus provides a small fitness benefit. Hence, the focal trait under selection here is translated directly in its resulting fitness effect. If multiple loci are under selection, the fitness of an individual is the product of the finesses at each locus. An example of adding selection on two markers at locations of 50 and 60 cM to the previously selected hybridization scenario is:

Ancestry module	*Sequence module*
> s = 0.1	> s <- 0.1> s <- 0.1
> ancestor_under_ selection = 0	> allele_under_selection <- 1
> markers <- c(0.5, 0.6)	> markers <- sort(sample(dgrp2.3R.5k. data$markers, 2))
> s_matrix <- matrix(nrow = 2, ncol = 5)	> s_matrix <- matrix(nrow = 2, ncol = 5)
> s_matrix[1,] <- c(markers[1], 1, 1 + 0.5	> s_matrix[1,] <- c(markers[1], 1, 1 + 0.5
* s, 1 + s, ancestor_under_selection)	* s, 1 + s, allele_under_selection)
> s_matrix[2,] <- c(markers[2], 1, 1 + 0.5	> s_matrix[2,] <- c(markers[2], 1, 1 + 0.5
* s, 1 + s, ancestor_under_selection)	* s, 1 + s, allele_under_selection)
> pop <- simulate_ admixture(module = ancestry_module(number_ of_founders = 2, markers = markers), pop_ size = 1000,	> pop <- simulate_admixture(module = sequence_module(input_data = dgrp2.3R.5k.data, markers = markers),

Ancestry module	*Sequence module*
total_runtime = 100,select_matrix = s_matrix)	pop_size = 1000, total_runtime = 100, , select_matrix = s_matrix)

3 | DEMONSTRATION OF THE PACKAGE

To demonstrate the functionality of GenomeAdmixR, we first show a series of simulations to compare alternative strategies in Evolve & Resequence (*E&R*) experiments. Second, we show how simulations of GenomeAdmixR match analytical predictions in junction theory.

3.1 | E&R experiment

One of the goals in *E&R* experiments is to identify underlying genetic loci driving responses to selection (Burke & Rose, 2009; Schumer, et al., 2014). The most common sampling method starts with a 'well mixed population' founded by individuals sampled from the same natural population (*Model 1*). This method minimizes Linkage Disequilibrium (*LD*) while trying to compensate for genetic variability by performing large sampling (Kofler & Schlötterer, 2014). Alternatively, genetic variability can be maximized by sampling from different populations, but this can generate additional *LD* (Kawecki et al., 2012; Kofler & Schlötterer, 2014; Schumer, et al., 2014). However, it has been proposed that genetic bases of population' or species' differences can be investigated by using mixing lines from different populations or species (Parts et al., 2011). This method (*Model 2*) has been used in yeast to fine-tune Quantitative Trait Locus (*QTL)* studies from phenotypically extreme inbred lines using experimental selection (Koide et al., 2012; Parts et al., 2011).

We used GenomeAdmixR to simulate the entire *E&R* pipeline and explore the impact of both sampling schemes (Figure S1 demonstrates the simulation setup). Figure 4 (for a single simulation) indicates that the resolution to detect the marker under selection in the experiment was much higher in *Model 1*. Also, *Model 1* seems to experience less genetic hitchhiking compared to *Model 2*. We performed 10 independent simulations with the same results, and replicates can be used for further analysis and statistical tests. Overall, the obtained results indicate that *Model 1* is for this pipeline the preferred approach. We implemented the mentioned pipeline using both the *ancestry module* (Figure 4a) and the *sequence module* (Figure 4b, using *D. melanogaster* Reference Panel data of the 3R arm chromosome) rendering largely congruent results.

We further expanded *Model 1*, using individuals instead of isofemale lines (*Models 3* and 4) revealing similar trends (see Figure S2).

3.2 | Theory of junctions

From the extended theory of junctions (Janzen et al., 2018), we know that the expected number of junctions (where a junction delineates the end of one contiguous stretch of genomic content from the same ancestor, and the start of another) depends on the initial heterozygosity H_0 (where heterozygosity here reflects a locus with two alleles stemming from two different ancestors, e.g. hetero-ancestry). Janzen's extended theory of junctions focuses specifically on the scenario with two ancestors, whereas GenomeAdmixR allows for an arbitrary number of ancestors. We simulate the accumulation of junctions using the *ancestry module* of GenomeAdmixR for two population sizes (100 and 1,000 individuals) and vary the number of ancestors n in [2, 4, 8]. We find that when we average the number of accumulated junctions over 1,000 replicates, the average number of junctions closely follows that of our analytical expectation (Figure 5; See Supplementary Information).

4 | DISCUSSION

We have presented here a new R package that uses individual-based simulations to study genomic dynamics following admixture. We have shown how both the *ancestry* and *sequence module* can be used for a general approach and more specifically, we have shown how GenomeAdmixR can be used within an experimental evolution framework and how results obtained using GenomeAdmixR are in line with theoretical expectations.

We expect GenomeAdmixR to be applicable in many fields. First, we expect GenomeAdmixR to allow for improved ease of use of individual-based simulations to study patterns in the formation of contiguous tracts of ancestry, and their breakdown due to recombination. The study of the breakdown of ancestry blocks has recently attracted renewed interest, with studies expanding the theory of junctions towards potential marker effects (Janzen et al., 2018), studies utilizing patterns in the distribution of ancestry tracts to infer the onset of hybridization (Buerkle & Rieseberg, 2008; Corbett-Detig & Nielsen, 2017; Gravel, 2012; Liang & Nielsen, 2014; Medina et al., 2018; Pool & Nielsen, 2009; Schumer et al., 2020; Shchur et al., 2020; Ungerer et al., 1998) and the inclusion of ancestral tract information in the analysis of human history (Hellenthal et al., 2014; Payseur & Rieseberg, 2016). GenomeAdmixR complements these analyses by providing a framework to verify findings through simulation. Furthermore, preliminary sequencing data can be used to verify findings using the *sequence module*, providing an additional tool to explore the efficacy of ancestry tract information.

Second, we expect GenomeAdmixR to be an important tool in investigating population genomics and designing *E&R* experiments, where GenomeAdmixR can be used to test whether the setup of an experiment is sufficient to detect signatures of selection on a background of existing LD. The functionality of the package allows simulating the effect of (a) previous population dynamics that occurred before sampling (*natural populations*), (b) initial parameters when obtaining the starting population for selection (*mixed populations*) as well as (c) variation in genomic and *selection* parameters. Here, we have demonstrated how to use GenomeAdmixR to estimate the

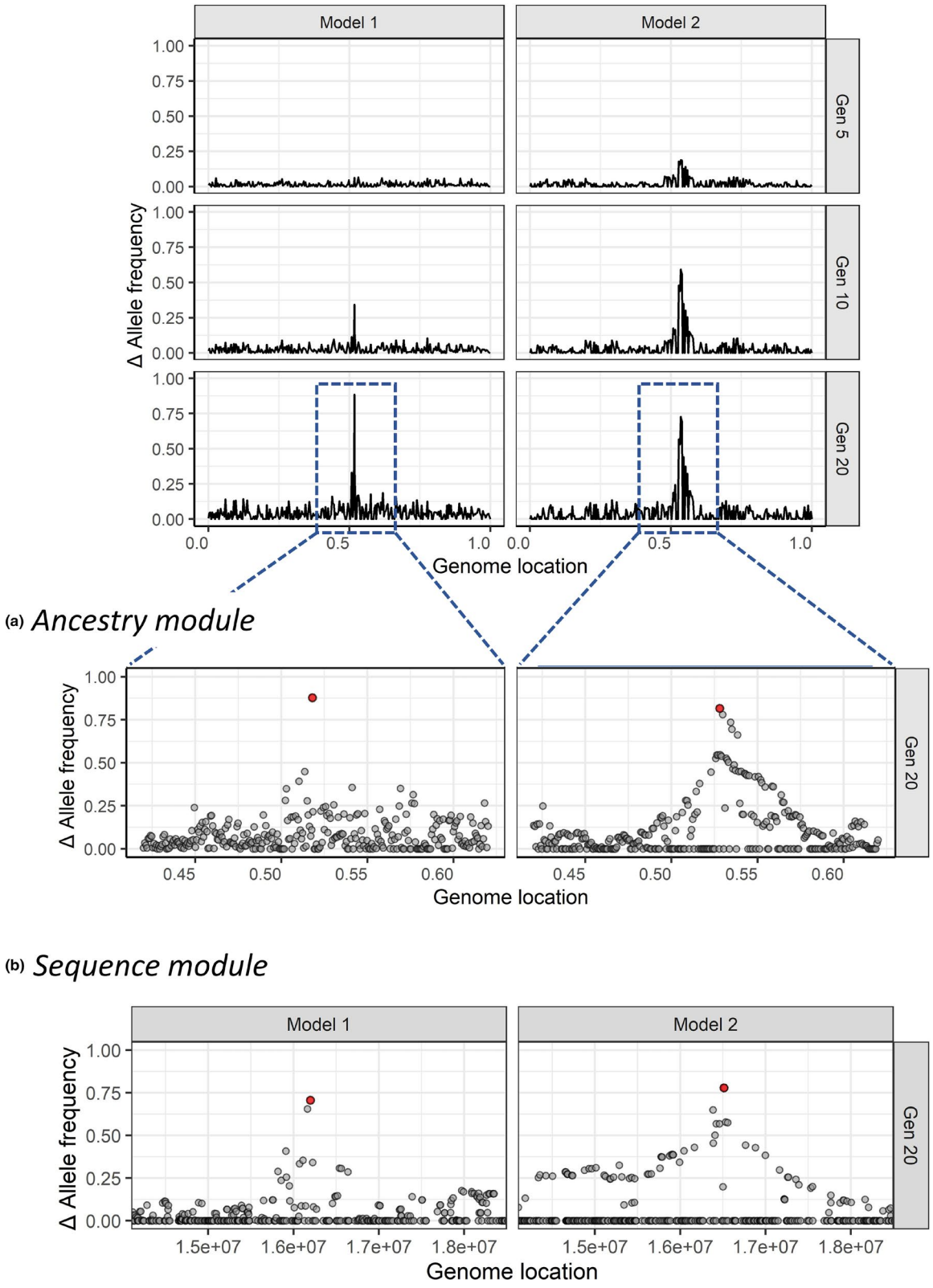

FIGURE 4 Summary of results obtained after simulating two scenarios for an *E&R* experiment using the GENOMEADMIXR package. Genome-wide allele frequencies after 5, 10 and 20 generations of selection are compared between the two models. The models differ in their starting populations as they were founded from a single or diverging populations (e.g. *Models 1* vs. *Model 2*, respectively). *Model 1*, founded by 100 isofemales from a single population; *Model 2*, founded by two isofemales from diverging populations. Selection was simulated for only one location (red point in the lower zoomed panel) in a 0.2 Morgan window, and tracked in 1,000 markers, while the rest of the genome evolves neutrally. The results are shown for simulations using the (a) *Ancestry module* and (b) the *Sequence module* starting with data from the *Drosophila melanogaster* Reference Panel

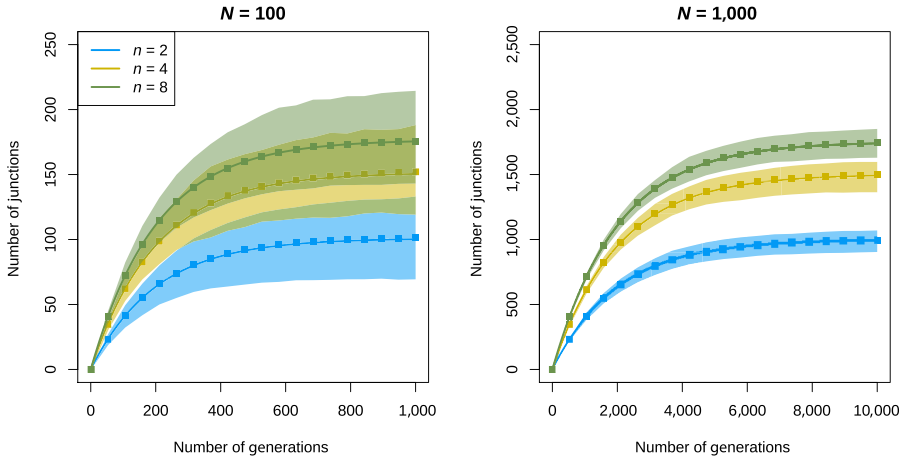

FIGURE 5 Accumulation of junctions for different numbers of ancestors (2, 4 and 8 unique ancestors), for a small population ($N = 100$) and a population of intermediate size ($N = 1{,}000$). Square dots indicate the mean number of junctions observed across 1,000 replicate simulations with the GenomeAdmixR package. Shaded areas indicate the 95% confidence interval across 1,000 replicates. The solid lines indicate the analytical prediction following Equation 1 in the Supplementary Material. Mean simulation dynamics follow the analytical prediction very closely

level of resolution obtained when combining inbred lines from different populations in the starting population of an E&R experiment (Koide et al., 2012; Parts et al., 2011). We found that although this model narrowed down the haplotype blocks around the selected marker substantially, this still lacks resolution when compared with traditional sampling (Kofler & Schlötterer, 2014; Schumer, et al., 2014) from a single population (even after 30 admixing generations). However, the obtained resolution seems deep enough to investigate questions involving variation between evolutionarily independent entities if these are not possible to address by sampling a single population. For example, Comeault and Matute (2018) used this model to study genomic trajectories following species hybridization.

Third, we expect GenomeAdmixR to function as a useful teaching tool, where it can be used to demonstrate the impact of hybridization on linkage patterns, the subsequent interaction between selection and drift and more generally to provide an easy toolkit to explore the interaction between recombination and hybridization.

GenomeAdmixR shares many similarities with its predecessors SliM (Haller & Messer, 2017, 2018) and SELAM (Corbett-Detig & Jones, 2016), both are powerful admixture simulation programs. However, SLiM is mainly focused on more advanced population demographic scenarios and is not directly suited for tracking ancestry. In contrast, SELAM tracks local ancestry, but requires local compilation of C++ code (which can be difficult) and uses complicated input tables to parameterize the simulations. GenomeAdmixR aims to alleviate both these issues, first GenomeAdmixR is specifically focused on tracking ancestry and its admixture over time. Second, because GenomeAdmixR is an R package, parameterization can be done through the use of the R scripting language, which facilitates sharing of code between platforms, replication of approaches and ease of use.

By providing both an *ancestry* and *sequence module*, we think that GenomeAdmixR provides a very complete package that can be used both to explore theoretical expectations (*ancestry module*) and verify these expectations using sequencing data (*sequence module*). Furthermore, because GenomeAdmixR readily accepts sequencing data in popular formats, it can be readily incorporated in existing pipelines. Thus, we expect that GenomeAdmixR has a bright future ahead, with a myriad of potential implementations across different fields of population genetics and molecular ecology.

ACKNOWLEDGEMENTS

We thank Omer Markovitch, Pratik R. Gupte and Cyrus Mallon and two anonymous refereers for helpful comments on an earlier version of the manuscript. The authors indicate that they have no conflicts of interest.

AUTHORS' CONTRIBUTIONS

T.J. and F.D. conceived the model, T.J. developed the R code, and F.D. and T.J. jointly wrote and revised the manuscript.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13612.

DATA AVAILABILITY STATEMENT

The R package is available via CRAN on https://cran.r-project.org/package=GenomeAdmixR. All code used for the paper has been deposited in the Dryad Digital Repository http://datadryad.org/resource/doi:10.5061/dryad.sqv9s4n3q (Janzen & Diaz, 2021).

ORCID

Thijs Janzen https://orcid.org/0000-0002-4162-1140
Fernando Diaz https://orcid.org/0000-0002-8594-7249

REFERENCES

Abbott, R., Albach, D., Ansell, S., Arntzen, J. W., Baird, S. J. E., Bierne, N., Boughman, J., Brelsford, A., Buerkle, C. A., Buggs, R., Butlin, R. K., Dieckmann, U., Eroukhmanoff, F., Grill, A., Cahan, S. H., Hermansen, J. S., Hewitt, G., Hudson, A. G., Jiggins, C., … Zinner, D. (2013). Hybridization and speciation. *Journal of Evolutionary Biology*, *26*, 229–246. https://doi.org/10.1111/j.1420-9101.2012.02599.x

Arnold, M. L., & Kunte, K. (2017). Adaptive genetic exchange: A tangled history of admixture and evolutionary innovation. *Trends in Ecology & Evolution*, *32*, 601–611. https://doi.org/10.1016/j.tree.2017.05.007

Barghi, N., & Schlötterer, C. (2019). Shifting the paradigm in evolve and resequence studies: From analysis of single nucleotide polymorphisms to selected haplotype blocks. *Molecular Ecology*, *28*, 521–524. https://doi.org/10.1111/mec.14992

Buerkle, C. A., & Rieseberg, L. H. (2008). The rate of genome stabilization in homoploid hybrid species. *Evolution*, *62*, 266–275. https://doi.org/10.1111/j.1558-5646.2007.00267.x

Burke, M. K., & Rose, M. R. (2009). Experimental evolution with Drosophila. *American Journal of Physiology: Regulatory, Integrative and Comparative Physiology*, *296*(6), R1847–R1854.

Chafin, T. K., & Douglas, M. R. (2020). Genome-wide local ancestries discriminate homoploid hybrid speciation from secondary introgression in the red wolf (Canidae : Canis rufus). 1–49.

Comeault, A. A., & Matute, D. R. (2018). Genetic divergence and the number of hybridizing species affect the path to homoploid hybrid speciation. *Proceedings of the National Academy of Sciences of the United States of America*, *115*, 9761–9766. https://doi.org/10.1073/pnas.1809685115

Corbett-Detig, R., & Jones, M. (2016). SELAM: Simulation of epistasis and local adaptation during admixture with mate choice. *Bioinformatics*, *32*, 3035–3037. https://doi.org/10.1093/bioinformatics/btw365

Corbett-Detig, R., & Nielsen, R. (2017). A hidden Markov model approach for simultaneously estimating local ancestry and admixture time using next generation sequence data in samples of arbitrary ploidy. *PLoS Genetics*, *13*, 1–40. https://doi.org/10.1371/journal.pgen.1006529

Cottin, A., Penaud, B., Glaszmann, J. C., Yahiaoui, N., & Gautier, M. (2020). Simulation-based evaluation of three methods for local ancestry deconvolution of non-model crop species genomes. *G3: Genes*, *Genomes*, *Genetics*, *10*, 569–579. https://doi.org/10.1534/g3.119.400873

Crow, J. F., & Kimura, M. (1970). *An introduction to population genetics theory*. Harper & Row.

Dennenmoser, S., Schatz, M. C., Sedlazeck, F. J., Zytnicki, M., Nolte, A. W., & Altmüller, J. (2019). Genome-wide patterns of transposon proliferation in an evolutionary young hybrid fish. *Molecular Ecology*, *28*(6), 1491–1505. https://doi.org/10.1111/mec.14969

Eddelbuettel, D., & Francois, R. (2011). Seamless R and C++ integration with Rcpp. *Journal of Statistical Software*, *40*, 1–18.Retrieved from http://www.jstatsoft.org/v40/i08/

Fisher, R. A. (1954). A fuller theory of 'Junctions' in inbreeding. *Heredity*, *8*, 187–197. https://doi.org/10.1038/hdy.1954.17

Fisher, R. A. (1959). An algebraically exact examination of junction formation and transmission in parent-offspring inbreeding. *Heredity*, *13*, 179–186. https://doi.org/10.1038/hdy.1959.21

Franssen, S. U., Barton, N. H., & Schlotterer, C. (2017). Reconstruction of haplotype-blocks selected during experimental evolution. *Molecular Biology and Evolution*, *34*, 174–184. https://doi.org/10.1093/molbev/msw210

Gerrish, P. J., Colato, A., Perelson, A. S., & Sniegowski, P. D. (2007). Complete genetic linkage can subvert natural selection. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 6266–6271. https://doi.org/10.1073/pnas.0607280104

Goudet, J. (2005). HIERFSTAT, a package for R to compute and test hierarchical F -statistics. *Molecular Ecology Notes*, *2*, 184–186. https://doi.org/10.1111/j.1471-8286.2004.00828.x

Gravel, S. (2012). Population genetics models of local ancestry. *Genetics*, *191*, 607–619. https://doi.org/10.1534/genetics.112.139808

Haller, B. C., & Messer, P. W. (2017). SLiM 2: Flexible, interactive forward genetic simulations. *Molecular Biology and Evolution*, *34*, 230–240. https://doi.org/10.1093/molbev/msw211

Haller, B. C., & Messer, P. W. (2018). SLiM 3: Forward genetic simulations beyond the wright-fisher model. *Molecular Biology and Evolution*, *36*, 632–637. https://doi.org/10.1093/molbev/msy228

Hellenthal, G., Busby, G. B. J., Band, G., Wilson, J. F., Capelli, C., Falush, D., & Myers, S. (2014). A genetic atlas of human. *Science*, *343*, 747–751.

Huang, W., Massouras, A., Inoue, Y., Peiffer, J., Ràmia, M., Tarone, A. M., Turlapati, L., Zichner, T., Zhu, D., Lyman, R. F., Magwire, M. M., Blankenburg, K., Carbone, M. A., Chang, K., Ellis, L. L., Fernandez, S., Han, Y., Highnam, G., Hjelmen, C. E., … Mackay, T. F. C. (2014). Natural variation in genome architecture among 205 *Drosophila melanogaster* genetic reference panel lines. *Genome Research*, *24*, 1193–1208. https://doi.org/10.1101/gr.171546.113

Janzen, T., & Diaz, F. (2021). Data from: Individual-based simulations of genome evolution with ancestry: The GENOMEADMIXR R package. *Dryad Digital Repository*, https://doi.org/10.5061/dryad.sqv9s4n3q

Janzen, T., Nolte, A. W., & Traulsen, A. (2018). The breakdown of genomic ancestry blocks in hybrid lineages given a finite number of recombination sites. *Evolution*, *72*, 735–750. https://doi.org/10.1111/evo.13436

Kawecki, T. J., Lenski, R. E., Ebert, D., Hollis, B., Olivieri, I., & Whitlock, M. C. (2012). Experimental evolution. *Trends in Ecology & Evolution*, *27*, 547–560. https://doi.org/10.1016/j.tree.2012.06.001

Keeling, P. J., & Palmer, J. D. (2008). Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics*, *9*, 605–618. https://doi.org/10.1038/nrg2386

Kessner, D., & Novembre, J. (2014). Forqs: Forward-in-time simulation of recombination, quantitative traits and selection. *Bioinformatics*, *30*, 576–577. https://doi.org/10.1093/bioinformatics/btt712

Kofler, R., & Schlötterer, C. (2014). A guide for the design of evolve and resequencing studies. *Molecular Biology and Evolution*, *31*, 474–483. https://doi.org/10.1093/molbev/mst221

Koide, T., Goto, T., & Takano-Shimizu, T. (2012). Genomic mixing to elucidate the genetic system of complex traits. *Experimental Animals*, *61*, 503–509. https://doi.org/10.1538/expanim.61.503

Lavretsky, P., Janzen, T., & McCracken, K. G. (2019). Identifying hybrids & the genomics of hybridization: Mallards & American black ducks of Eastern North America. *Ecology and Evolution*, *9*, 3470–3490.

Leitwein, M., Gagnaire, P. A., Desmarais, E., Berrebi, P., & Guinand, B. (2018). Genomic consequences of a recent three-way admixture in supplemented wild brown trout populations revealed by local ancestry tracts. *Molecular Ecology*, *27*, 3466–3483. https://doi.org/10.1111/mec.14816

Liang, M., & Nielsen, R. (2014). The lengths of admixture tracts. *Genetics*, *197*, 953–967. https://doi.org/10.1534/genetics.114.162362

MacKay, T. F. C., Richards, S., Stone, E. A., Barbadilla, A., Ayroles, J. F., Zhu, D., Casillas, S., Han, Y., Magwire, M. M., Cridland, J. M., Richardson, M. F., Anholt, R. R. H., Barrón, M., Bess, C., Blankenburg, K. P., Carbone, M. A., Castellano, D., Chaboub, L., Duncan, L., … Gibbs, R. A. (2012). The *Drosophila melanogaster* genetic reference panel. *Nature*, *482*, 173–178. https://doi.org/10.1038/nature10811

Macleod, A. K., Haley, C. S., Woolliams, J. A., & Stam, P. (2005). Marker densities and the mapping of ancestral junctions. *Genetical Research*, *85*, 69–79. https://doi.org/10.1017/S0016672305007329

Medina, P., Thornlow, B., Nielsen, R., & Corbett-Detig, R. (2018). Estimating the timing of multiple admixture pulses during local ancestry inference. *Genetics*, *210*, 1089–1107. https://doi.org/10.1534/genetics.118.301411

Otte, K. A., & Schlötterer, C. (2021). Detecting selected haplotype blocks in Evolve and Resequence experiments. *Molecular Ecology Resources*, *21*(1), 93–109.

Parts, L., Cubillos, F. A., Warringer, J., Jain, K., Salinas, F., Bumpstead, S. J., Molin, M., Zia, A., Simpson, J. T., Quail, M. A., Moses, A., Louis, E. J., Durbin, R., & Liti, G. (2011). Revealing the genetic structure of a trait by sequencing a population under selection. *Genome Research*, *21*(7), 1131–1138. https://doi.org/10.1101/gr.116731.110

Payseur, B. A., & Rieseberg, L. H. (2016). A genomic perspective on hybridization and speciation. *Molecular Ecology*, *25*, 2337–2360.

Pool, J. E., & Nielsen, R. (2009). Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics*, *181*, 711–719. https://doi.org/10.1534/genetics.108.098095

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Schlötterer, C., Kofler, R., Versace, E., Tobler, R., & Franssen, S. U. (2015). Combining experimental evolution with next-generation sequencing: A powerful tool to study adaptation from standing genetic variation. *Heredity*, *114*, 431–440. https://doi.org/10.1038/hdy.2014.86

Schumer, M., Cui, R., Powell, D. L., Dresner, R., Rosenthal, G. G., & Andolfatto, P. (2014). High-resolution mapping reveals hundreds of genetic incompatibilities in hybridizing fish species. *eLife*, *2014*, 1–21.

Schumer, M., Powell, D. L., & Corbett-Detig, R. (2020). Versatile simulations of admixture and accurate local ancestry inference with mixnmatch and ancestryinfer. *Molecular Ecology Resources*, *20*, 1141–1151.

Schumer, M., Rosenthal, G. G., & Andolfatto, P. (2014). How common is homoploid hybrid speciation? *Evolution*, *68*, 1553–1560. https://doi.org/10.1111/evo.12399

Schumer, M., Xu, C., Powell, D. L., Durvasula, A., Skov, L., Holland, C., Blazier, C., Sankararaman, S., Andolfatto, P., Rosenthal, G. G., & Przeworski, M. (2018). Natural selection interacts with recombination to shape the evolution of hybrid genomes. *Science*, *360*, 656–660. https://doi.org/10.1126/science.aar3684

Shchur, V., Svedberg, J., Medina, P., Corbett-Detig, R., & Nielsen, R. A. S. M. (2020). On the distribution of tract lengths during adaptive introgression. *G3: Genes*, *Genomes*, *Genetics*, *10*, 3663–3673. https://doi.org/10.1534/g3.120.401616

Stam, P. (1980). The distribution of the fraction of the genome identical by descent in finite random mating populations. *Genetical Research*, *35*, 131. https://doi.org/10.1017/S0016672300014002

Ungerer, M. C., Baird, S. J., Pan, J., & Rieseberg, L. H. (1998). Rapid hybrid speciation in wild sunflowers. *Proceedings of the National Academy of Sciences of the United States of America*, *95*, 11757–11762. https://doi.org/10.1073/pnas.95.20.11757

Vlachos, C., & Kofler, R. (2018). MimicrEE2: Genome-wide forward simulations of Evolve and Resequencing studies. *PLOS Computational Biology*, *14*(8), 1–10. https://doi.org/10.1371/journal.pcbi.1006413

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure author(s): B. S. Weir and C. Clark Cockerham Published by : Society for the Study of Evolution Stable URL : Http://www.jstor.org/stable/2408641. *Evolution*, *38*, 1358–1370.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis - Hadley Wickham - Google Books*. Retrieved from https://books.google.nl/books?hl=nl&lr=&id=XgFkDAAAQBAJ&oi=fnd&pg=PR8&dq=ggplot2&ots=so58bP5WbN&sig=ThS6gEgxaK9XADL_HeG5gDU04Pc#v=onepage&q=ggplot2&f=false

Zhou, Y., Qiu, H., & Xu, S. (2017). Modeling continuous admixture using admixture-induced linkage disequilibrium. *Scientific Reports*, *7*, 1–10. https://doi.org/10.1038/srep43054

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Janzen T, Diaz F. Individual-based simulations of genome evolution with ancestry: The GenomeAdmixR R package. *Methods Ecol Evol*. 2021;12:1346–1357. https://doi.org/10.1111/2041-210X.13612