Model description and fitting method description

Full model description

Model summary

We simulated the experiment using an individual based stochastic model in which mosquitoes are grouped by life-stage (larvae or adult), age, sex, and genotype. Adult females are further grouped by whether or not they have mated (it is assumed that females mate only once) and, if they have, by the genotype of their mate. Finally, females are grouped by whether either parent, or both parents, carried the drive allele. We assume there are three possible allele types at the dsx locus: wildtype (W), Ag(QFS)1 drive construct (D), and non-functional alleles that are resistant to homing (R).

The model simulates the daily changes to the population that result from egg laying, deaths, and matings. Each of these processes is stochastic, with probabilities that depend on the mosquito groupings. For example, the per day survival probability of adult mosquitoes reduces with age, while the egg production of a female will depend on her genotype and possibly her parental genotypes. The model replicates the experimental design with respect to twice-weekly egg-laying, the initiation phase, the transgene introductions, and the subsequent monitoring phase. The model parameters and their default values are listed in table 1.

Egg laying

Mated females lay eggs twice each week (on Mondays and Thursdays), and the probability a given female lays is p_{Lay}. For those that do, the number of eggs is Poisson distributed with expectation $\theta \times \omega_i^q$ where θ is the fertility of wildtype females, and ω_i^q is the relative fertility of females whose genotype is i and the superscript q is either f, m, n or b to indicate, respectively, that the mother had the drive allele, the father had it, neither had it, or both had it. We assume females need at least one copy of the wildtype allele to produce eggs, and so $\omega_{DD}^q = \omega_{DR}^q = \omega_{RR}^q = 0 \quad \forall q$. Drive heterozygous females have a fitness cost of ξ, and we set the maternal and paternal drive allele fitness costs to be ρ_f and ρ_m. This gives

$$
\begin{aligned}
\omega^n_{WW} &= \omega^n_{WR} = 1 \\
\omega^f_{WW} &= \omega^f_{WR} = 1-\rho_f \\
\omega^m_{WW} &= \omega^m_{WR} = 1-\rho_m \\
\omega^b_{WW} &= \omega^b_{WR} = (1-\rho_f)(1-\rho_m) \\
\omega^n_{WD} &= 1-\xi \\
\omega^f_{WD} &= (1-\xi)(1-\rho_f) \\
\omega^m_{WD} &= (1-\xi)(1-\rho_m) \\
\omega^b_{WD} &= (1-\xi)(1-\rho_f)(1-\rho_m)
\end{aligned}
$$

The genotype of each egg is randomised depending on the parental genotypes. The inheritance probabilities are Mendelian except in the case of W/D parents for which we assume sex-specific homing rates as measured in small cages. Non-homed gametes from W/D individuals become R alleles with probability γ. A random sample of 400 larvae (or all the larvae if there are less than 400) are kept from each twice-weekly batch of eggs and added to the adult population after their development which takes ten days.

Transgenic female fertility costs

Note that the final three terms above relate the overall fitness costs of transgenic females depending on parentage, ω^f_{WD}, ω^m_{WD}, and ω^b_{WD}, to the mechanistic parameters ξ, ρ_f and ρ_m. We used these equations to transform the posterior distributions of the mechanistic parameters to posterior estimates of the fitness costs shown in Fig. 2i of the main text.

Deaths

We used the cage survival assays that were performed both before and after the population dynamic experiments to parameterise adult survival, as follows. For both the start and end survival assay, we fitted Weibull distributions to the survivorship curves separately for females and males. For each sex, we transformed the Weibull parameters linearly from the start to the end of the experiment. On day d, we thus have

$$
\begin{aligned}
k_i(d) &= k_i(0) + d \times \frac{k_i(T) - k_i(0)}{T}, \\
\lambda_i(d) &= \lambda_i(0) + d \times \frac{\lambda_i(T) - \lambda_i(0)}{T}.
\end{aligned}
$$

The mortality probability of an adult dying on day d, $\mu_i(d,a)$, is then determined by a function that depends on d as well as the individual's age, a, and sex, $i \in \{m, f\}$. This has the form:

$$
\mu_i(d,a) = 1 - e^{\left(\frac{a}{\lambda_i(d)}\right)^{k_i(d)}} - e^{\left(\frac{1+a}{\lambda_i(d)}\right)^{k_i(d)}}
$$

Parameter	interpretation	default value	source
e_m	Homing rate in males	0.9635	Kyrou et al. (2018)
e_f	Homing rate in females	0.9985	Kyrou et al. (2018)
β	Number of males when females mate with probability 1/2 per day	100	Expert judgement
$(k_f(0), k_f(T), k_m(0), k_m(T))$	Weibull shape parameters, for females $\cdot_f$ and males $\cdot_m$, at the start (0) or end (T) of the experiment	$(2.33, 2.39, 2.69, 1.79)$	Survival assays
$(\lambda_f(0), \lambda_f(T), \lambda_m(0), \lambda_m(T))$	Weibull scale parameters, for females $\cdot_f$ and males $\cdot_m$, at the start (0) or end (T) of the experiment	$(6.50, 15.91, 5.21, 10.38)$	Survival assays
ρ_f	Fitness cost of paternal cas9 deposition	$0.67^{*\dagger}$	Kyrou et al. (2018)
ρ_m	Fitness cost of maternal cas9 deposition	$0^{*\dagger}$	Kyrou et al. (2018)
ξ	Fitness cost of somatic cas9 expression	$0.35^{*\dagger}$	Kyrou et al. (2018)
γ	Proportion of non-homed alleles that form R alleles	0.5	
p_{Lay}	Proportion of females that lay eggs at a given opportunity	0.5^*	Hammond et al. (2020)
θ	Number of eggs per batch from a fully fit female	20^*	Expert judgement

Table 1: Model parameters. *: we also inferred this parameter from the cage data; †: the default values for these parameters are derived from the fertility of heterozygous females as measured by Kyrou et al. (2018) for the different parental combinations of genotypes.

Mating

A virgin female will mate on a given day with a probability $\frac{M_{\text{Tot}}}{M_{\text{Tot}}+\beta}$ where M_{Tot} is the total number of adult males and β is a parameter that controls the extent to which mating becomes limited when there are few males in the population. We set $\beta = 100$, which means that most virgin females will mate on the first day after their emergence (mating probability $> 1/2$ per day if there are > 100 males in the cage). The male a female mates with is randomised from all the males in the cage (neither genotype nor age affects mating ability).

ABC Model fitting

We attempt to fit six model parameters from the model to data comparison: $\xi, \rho_f, \rho_m, \gamma, p_{\text{Lay}}$, and θ.

Fitting method

We fit the these parameters using a simple Monte-Carlo algorithm based on 'approximate Bayesian computation' (ABC). In short:

1. Parameters are drawn at random from a prior distribution. The parameter vector is called a *particle*.
2. A number of simulations are run with this particle.
3. The simulations are scored for their closeness to the cage data according to a number of distance measures (below).
4. Steps 1-3 are repeated lots of times.
5. The particles are filtered using the distance scores. E.g. we retain all particles where each distance score is in the lowest q quantile (for some q).
6. The retained particles are the posterior distribution.

Distance metrics

Time-series smoothing: transgene frequency and male ratio We first smooth each time-series in the data $\{a_d\}_{d=1..T}$, where T is the length of the time series, by transfroming it into a 2 week moving average time-series $\{S(a)_d\}_{d=3..T-2}$. Note that the smoothed time series is four data points fewer than the original series because there are two data points per week each original series. We compute the sum of square differences between a_d and $S(a)$, $R(a) = \sum_{d=3}^{T-2}(a_d - S(a)_d)^2$. We apply the same smoothing and residual functions to the simulated data, to get corresponding variables $S(b)$ and $R(b)$. By applying these transformations to the two sets of treatment time series, we obtain four distance measures between data and simulated data.

1. Transgene frequency trend distance. This distance, d_1, is the sum of square differences between all corresponding pairs of real and simulated transgene frequency time-series:

$$d_1 = \sum_{i=1,2}\sum_{j=1,2}\sum_{k=1}^{N_{sim}}\left(\sum_d (S(a)_d^{(i,j)} - S(b)_d^{(i,k)})^2\right) \quad (1)$$

   where i refers to the experimental treatment (25% or 50%) starting transgene frequency, j refers to the data replicate, k refers to the simulation replicate, and d refers to the day within the (smoothed) time series.

2. Transgene frequency noise distance. This distance, d_2, is the sum of square differences between all corresponding pairs of real and simulated transgene frequency residuals:

$$d_1 = \sum_{i=1,2}\sum_{j=1,2}\sum_{k=1}^{N_{sim}}\left((R(a)^{(i,j)} - R(b)^{(i,k)})^2\right) \quad (2)$$

3. Male and homozygous female proportion trend. This distance, d_3, is computed as d_1 but using the male and homozygous female time series.

4. Male and homozygous female proportion noise (d_4). As d_2 but again using the residuals of the male and homozygous female time series.

Control cage egg number In addition, we obtain two distance measures from the control cage egg number data. To do so we compute the mean and variance of the two time series of egg number in the control cages, from day -53 to $+248$. This gives $M(e^j) = \text{Mean}(e_d^j)$ and $V(e^j) = \text{Variance}(e_d^j)$ for replicate j (out of two replicates), and we obtain similarly $M(e_{sim}^k)$ and $V(e_{sim}^k)$ for the simulated control cages.

5. Egg number mean (d_5). This distance is $\sum_{j=1,2} \sum_{k=1}^{N_{sim}} \left(M(e^j) - M(e_{sim}^k)\right)^2$.

6. Egg number variance (d_6). This distance is $\sum_{j=1,2} \sum_{k=1}^{N_{sim}} \left(V(e^j) - V(e_{sim}^k)\right)^2$.

r2 allele data Finally, we obtain one distance measure, d_7, based on the observed number of r2 alleles at different time points. We denote the data point $c^{(i,j,d)}$ to be the observed frequency of r2 alleles based on $N^{(i,j,d)}$ samples, where i is treatment, j is replicate, and d is day. Similarly, $c_{sim}^{(i,k,d)}$ is the proportion of r2 alleles given in the k^{th} replicate of the simulation of treatment i on day d.

We construct a likelihood function for observing $c^{(i,j,d)}$ if the simulation proportion of r2 alleles is correct:

$$lik(c^{(i,j,d)}) \quad = \quad PDF[\text{Binomial distribution}(N^{(i,j,d)}, c_{sim}^{(i,k,d)}), N^{(i,j,d)} * c^{(i,j,d)}).$$

7. d_7 is then the sum of log-likelihoods over all data points:
$d_7 = -1 * \sum_{i=1,2} \sum_{k=1}^{N_{sim}} \left(Log[lik(c^{(i,j,d)})\right]$. Note we multiply by minus one because smaller distances should correspond to a better correspondence of data and simulation.

References

Hammond, A. M., Karlsson, X., Morianou, I., Kyrou, K., Beaghton, A., Gribble, M., Kranjc, N., Galizi, R., Burt, A., Crisanti, A., et al. (2020). Regulation of gene drive expression increases invasive potential and mitigates resistance. *bioRxiv*, page 360339.

Kyrou, K., Hammond, A. M., Galizi, R., Kranjc, N., Burt, A., Beaghton, A. K., Nolan, T., and Crisanti, A. (2018). A CRISPR-Cas9 gene drive targeting doublesex causes complete population suppression in caged *Anopheles gambiae* mosquitoes. *Nat Biotechnol*, 36:1062.