

## Peer Review File

---

Population suppression by release of insects carrying a dominant sterile homing gene drive targeting doublesex in *Drosophila*



**Open Access** This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

## Reviewers' Comments:

### Reviewer #1:

#### Remarks to the Author:

In this paper, the authors transfer an existing gene drive technology in *Anopheles* mosquitoes to *Drosophila melanogaster*, targeting the developmental sex determination gene *doublesex* (*dsx*). They find that unlike in *Anopheles* (but similar to *D. suzukii*), *dsx* knockouts produce a dominant lethal phenotype in females that limits its utility of *dsx* as a target for standard homing drives in *Drosophila*. Homing drives are also less efficient in *Drosophila* spp., so a high frequency of potential resistance genes are created. The authors found that these all resistance genes generated to date are also dominant lethal in females. As a consequence, when used as a population suppression tool, this approach is similar to genetic control methods such as daughterless, except that they are more effective because males carrying the female lethal construct will continue to be produced as long as there are wild type females present in the target population. Like other female-lethal approaches, multiple releases are required. The authors promote this approach as more effective than existing non-driving approaches, while also being safer than standard gene drive approaches because it is self-limiting.

The paper is well written, but perhaps a bit longer than it needs to be. I was initially confused by the Abstract, which described the approach as "combining features of homing drive and fsRIDL" while also naming the approach RIDD (Release of Insects carrying a Dominant sterile Drive). Because they use a name so close to RIDL, I immediately and incorrectly interpreted that some aspects of the molecular approach from RIDL were used in RIDD. I was halfway through the paper before it became apparent that the approach resembled RIDL only in the dynamics of suppression. Perhaps the authors could use the "combining characteristics", because features can refer to components of a construct. I also think that naming the technology so similar to an existing, but quite different approach will lead to more product confusion - the authors may want to consider a different name or acronym. Daughterless technologies (eg. Thresher 2008) pre-dated fsRIDL, so perhaps these should be included in the discussion and analysis - and considered for the nomenclature.

In general, the methods are sound and were described sufficiently to be repeated. My only concern in the methods was the scoring of drive male mating success in the mating competitive assays. The authors describe (p. 4, para 2) that "if the percentage of individuals with red fluorescent eyes was higher than approximately 60-70%, we scored this as a drive male mating success. First, setting a threshold as "approximate" is worrying. Secondly, I can see that the authors might want some leeway from the expected conversion efficiency (80% +), but some statistical basis for selecting a

threshold would be reassuring.

My final concern is the throwaway commentary that this approach would be self-limiting, without including any modelling. I believe the paper should include some modelling on how much spread of the construct (and effects) would occur into neighbouring populations with different levels and repetitive rates of release.

Reviewer #2:

Remarks to the Author:

Summary

The authors have created a split-drive DsX drive construct, using a multiplexed approach to target and disrupt a female-specific splice site. The use of overlapping generation cage trials is very responsible and the implementation of a potential novel method for calculating population size should be expanded upon and shared with the community. The methods need expanded, some experimental choices need justified (het. vs hom. Cas9), and the modeling needs completely redone.

Page 3, Methods, Plasmid construction

Sentence 1 needs a reference. "...which were constructed previously."

Page 3, Methods, Phenotypes and Morphological Analysis

I didn't realize this was a split drive design until the 3rd sentence of this section. Since it is a split drive, I think the background needs to include some discussion on split drive designs, not just the classic CRISPR design. It also needs to be clarified that RIDD is a split drive design, as that's neither how the original CRISPR designs worked nor how fsRIDL is designed.

The Akbari and Bier groups have done a lot with split drives, anything specifically relating to dominant effect drives or female specific drives that should be noted here? These would be the proper comparison for RIDD. (I can't think of anything that is female specific, but dominant effectors are not a new idea.)

Page 3, Methods, Phenotypes and Morphological Analysis

Fix capitalization in this header.

"Most lines were tested in drive heterozygotes..." - why? Any untested hypotheses on how this

impacted transmission?

Page 4, Methods, Male mating competitiveness assay

Can we specify that the males were also nanos-?

Why were competitiveness assays not performed for any other promoters, or using heterozygotes?

Page 4, Methods, Cage study

It sounds like the population sizes for the cages were estimates, but I don't see that noted in the data supplement. Can I also assume that the population size estimate was not used during the statistical analyses? Did you take multiple pictures to perform the population size estimates or just a single one – and if so, what's the standard error on the population size estimates? Was there any validation done on the accuracy of this method for population estimates? Maybe a reference demonstrating design (computer vision or segmentation approach?) and accuracy.

In the data supplement, Data Set S1, what is it? I can't find a description/reference anywhere. I'm guessing it's the drive-inheritance/conversion metrics, and maybe includes the data for the male mating competitiveness?

Data Set S2 – only 2 cages? I don't see that noted in the methods, and I'm a little curious where there isn't at least 3? 5 would be amazing, but that's also a ton of work.

Data Set S3 – were the "Weekly Release size" estimates as well, or exact counts? Did I miss a note explaining the "NA" in week 2? The control cage was stopped at week 10 – why?

Page 5, Methods, Population models

The population model is discrete generation, but I didn't get the impression from the "Cage study" that it was discrete generation. Was it discrete generation, and I missed, or does your model not match your experiment?

I'm really confused what your first model is good for – it's not a split drive, it's a population frequency model, it's discrete generation. It neither matches your experiments nor reality (which I think it is used to describe later). I think we could claim it matches the experiments b/c they were performed in a homozygous Cas9 background – that should be stated with the "unlike our experiments.." as a simplifying assumption. The model is assumed to be "...a complete drive..." which is unsupported by anything in the methods, and not supported by the "Drive performance ..." section in the results (looks like max 85%?). If the purpose of this model is to fit data from the experiments, I think it needs to reflect the experiments. If it's to perform predictive simulations, I think it needs to reflect reality. At the moment, it's does neither.

SLiM – “...single panmictic population of diploids...” – so not split-drive?

“Remaining wild-type alleles are all converted to resistance alleles.” – So the SLiM simulations also assume “a complete drive”?

Are there 2 sets of simulations using SLiM? It seems like there was one set, with discrete generations (“Stochastic simulations were performed in SLiM 4.0, similar to previous ...”), which I can’t tell where/how this was used, and then another set of simulations (“A SLiM framework was also used to simulate the dynamics...”) using weekly steps and some ecological models. Would it be possible to clarify this and specify what each model was used for? Why did the first set of simulations not use density-dependent death on the eggs? Why the different distributions on eggs (Binomial vs Poisson? At  $E[x]=2$ , they are significantly different)?

“100 drive males were released into ...” – Why? The cages trials didn’t do this, correct? Why the increase after those 10 weeks?

“The low-density growth rate was 10, roughly estimated ...” – how was this estimate performed? A short note (<1 page) in the supplement of how that was done would be plenty.

“Because the effective population size...” – Interesting, I’ve never seen this issue noted in experimental settings before. Can you explain how your experimental setup matches the Wright-Fisher model used to demonstrate this effect? If this result applies here, how come that model wasn’t used for parameter estimation and simulations here?

Page 6, Results, Drive performance and ...

I don’t see any note in here that some lines were homozygous Cas9 and some heterozygous. That’s an important caveat that needs to be repeated with the results, possibly with some discussion on how it impacted said results.

Page 7, Results, Figure 1

Following on from the Methods, Phenotypes ... section – how come some lines were done as Cas9 heterozygotes, some as homozygotes, and the cage trials as homozygotes? It seems like an inconsistency, which might be part of the discontinuities with the models.

Looking at 1a, this drive uses multiplexing? That should be noted in the intro, with appropriate references.

Page 8, Results, Figure 2

Panel “a” uses “drive”, while panel “c” uses “D”, and the description uses “D”.

I appreciate the note on homozygous Cas9 line used here.

Section “b” needs to have “dsx” italicized – “female-specific dsx transcripts.”

Suggestion – It would be nice to have panel “b” on top, relabeled as panel “a”, with panels “a”

and “c” moved to the bottom and labels updated. That would make it consistent with figure 1 and would

fix the strange “a”/”c” panel combination on top, so that it flows from top-to-bottom, then left-to-right.

Page 9, Results, Resistance allele formation and classification

“...and targeted by three gRNAs.” – This notes that it is a multiplexed drive, which hasn’t been stated up to this point. That literature needs included in the intro, as part of the ethos behind the construct design, and what the expected impact is (why complicate something, unless you get a benefit?). Were any tests done on single or double gRNA constructs, to determine that 3 gRNAs was necessary for the best conversion rate?

Page 9, Results, New pest control strategy RIDD system

Need to move the header to page 10 (might be a note for printing later).

“...high total germline cut rates in males...” – where was this measured? Figure 1 only shows transmission with no discussion of cleavage. Is this why your models assume “a complete drive”?

“It is more powerful and efficient than fsRIDL...” – based on what? Your intro simply states that fsRIDL requires mass-rearing facilities. The current paragraph states “...which are released continuously in large number into a population...” – which is exactly what fsRIDL requires. The effect on mass-rearing needs to be clarified, as well as the expected real-world performance difference between RIDD and fsRIDL. Are smaller releases required, and if so, by how much? Is RIDD easier to rear, why? I think the actual improvement is in transmission rates from males, which is kinda alluded to in the intro “...it has been proposed to combine fsRIDL with drive”.

“...allowing for highly confined population suppression...” – theoretically, I see what you’re trying to say. Any proof that this is real? Are there any requirements on migration rate to make this true? Any restrictions on the size/frequency/total amount of releases to prevent spillage into nearby populations? Any caveats a local government should be aware of before attempting to apply this method?

Page 11, Figure 5

“...fly populations with overlapping generations” – This needs to be clarified in the methods and then which simulations does it match? At least 2 of the simulations use discrete generations, as well as the Wright-Fisher model.

“For the first 10 weeks...” – okay, so here is the 10-weeks of releases I questioned in the model earlier. Why aren’t these noted in the methods for the cage trials? Why the range of 3-7%? B/c of the population size being an estimate, and if so, why the massive variance when you can more accurately hit 20% later? Given the release sizes here (3-7%, then 20%), why the difference with model releases (100 flies, then 600 flies)? Given an actual simulated population of 2500 (right? Or is that pre-reduction? And if so, did you reduce release sizes accordingly?) (I’m assuming the 3 set of sims, the second SLiM ones), 100 flies is 4% and 600 is 24%, not exactly in-line with the experimental setup.

Panel “a” – what’s the error on the population sizes? That should be plotted.

Panel “c” – “This data was in agreement with ...” – is that a measurable/statistical agreement or you looked at the plot? T0 doesn’t match, T8/9 is questionable, the drop at T11/12 is not matched but the sims eventually catch up around T20. It’d be nice to have a distributional measure (or time-series measure) to quantify the similarity between sims and cages.

#### Page 11, Results, Assessment of the RIDD in cage populations

“...all flies were homozygous for the split...” – was it ever clarified why you picked this line to do cage trials? Were any other lines tested in the mating competitiveness assay?

“...was substantially increased to approximately 500 per cage...” – why don’t your models reflect this?

“We observed the sex ratio biased ...” – none of the models mention reflecting this.

“...which was approximately 1200 flies.” – why was the control cage so much smaller? It’s nearly 1/2 the size of the test cages. Somewhere else I already asked why it was only kept for ~1/3 the time of the cage trials.

“...using estimates for several parameters related to the ...” – why only simulate 1 resistant allele, when you have population measures for 3 of them?

This section also assumes 100% cleavage – need the data shown/noted somewhere.

“The trajectories ... closely matched ...” – “closely” needs quantified. KL-divergence or earth-movers distance for distributions, or a time-series approach to quantify this. There may be better methods out there. A binomial test at each datapoint is inappropriate for this – it fails to account for the correlated nature of each datapoint.

#### Page 12, Results, Modelling of RIDD system

“For RIDD and similar technologies such as SIT and RIDL...” – RIDD is a split-gene drive, SIT and RIDL are not the most related comparisons, you should compare to other split-drive designs (Akbari has some, Hay has some).

“Here, we introduce a possible method of comparison based on genetic load...” – isn’t that published here - <https://doi.org/10.1111/eva.12583>? I believe there is another paper that introduces this too, written by several Asian authors who used Mathematica to create a unified framework for testing and measuring several different gene drives (I’m really sorry I can’t remember anything more useful about this paper).

“We developed a deterministic, discrete-generation ...the wild-type population does not change...” – Nothing in the remainder of this section matters. These simulations don’t reflect reality and they don’t reflect the drive design presented in this paper. This section needs to be removed or rewritten to be correct and applicable, not a misleading statement on the possible efficacy of some other design in an impossible population.

Page 13, Discussion

“...and homing suppression drives that are dominant-sterile due to somatic expression.” – this is old news, and has been ameliorated through better target choices and the use of germline-specific promoters (such as the nanos promoter implemented here).

“Our overlapping-generation cage study...” – I don’t think I said it earlier, but I really appreciate this cage trial design. Very good choice, I hope it propagates through the field.

“...increasing to approximately 50~60% ...” – this is a huge percentage of the population. Is this even feasible in a real population? It seemed like ~20% was the minimum effective release size to crash the population, how feasible is that?

“RIDD is substantially more efficient...” – how much more efficient in a population w/o Cas9? Then it’s no longer a Cas9 homozygous background, so the sims won’t match.

“If releases are ceased at any time ...” – can you prove this? This is a male-only drive, so it won’t propagate well, but how many generations before the population crashes do we need to stop releases?

How long do we expect it to hang around in the population? Is this impacted by the split-drive nature of your design?

“...highly confined population suppression...” – that is not demonstrated here.

Summary of improvements:

1. We added a drive conversion experiment for drive males carrying heterozygous *nanos*-Cas9-*nanos*. Now for all Cas9 lines, the genotype of the test individuals is the same (*dsxD/+*, Cas9/+).
2. We expand comparisons of theoretical efficiency of different genetic strategies, showing the substantial advantages of RIDD in deterministic models for RIDD, fsRIDL, fsRIDL-drive, and SIT (Fig.6a).
3. We built an overlapping-generation panmictic modeling for RIDD complete drive, split drive, SIT and fsRIDL, which is further supportive for illustrating RIDD's advantages in simulations where target population elimination is the objective.
4. We model a split RIDD system (equivalent to the system we constructed - though the cage experiment is still representative of complete RIDD due to the Cas9 background) and show that it is even more strongly self-limiting than complete RIDD.
5. We built a confinement model for RIDD complete drive in two demes linked by migration. The maximum drive carrier frequency and population suppression in the non-release deme is very low, which support our claims for confinement.
6. Many other smaller improvements in the figures and text throughout the manuscript.

In the following, editor and reviewer comments are in bold, and our responses are in plain text.

#### **Reviewer #1 (Remarks to the Author):**

**In this paper, the authors transfer an existing gene drive technology in *Anopheles* mosquitoes to *Drosophila melanogaster*, targeting the developmental sex determination gene doublesex (*dsx*). They find that unlike in *Anopheles* (but similar to *D. suzukii*), *dsx* knockouts produce a dominant lethal phenotype in females that limits its utility of *dsx* as a target for standard homing drives in *Drosophila*. Homing drives are also less efficient in *Drosophila* spp., so a high frequency of potential resistance genes are created. The authors found that these all resistance genes generated to date are also dominant lethal in females. As a consequence, when used as a population suppression tool, this approach is similar to genetic control methods such as daughterless, except that they are more effective because males carrying the female lethal construct will continue to be produced as long as there are wild type females present in the target population. Like other female-lethal approaches, multiple releases are required. The authors promote this approach as more effective than existing non-driving approaches, while also being safer than standard gene drive approaches because it is self-limiting.**

The paper is well written, but perhaps a bit longer than it needs to be. I was initially confused by the Abstract, which described the approach as "combining features of homing drive and fsRIDL" while also naming the approach RIDD (Release of Insects carrying a Dominant sterile Drive). Because they use a name so close to RIDL, I immediately and

incorrectly interpreted that some aspects of the molecular approach from RIDL were used in RIDD. I was halfway through the paper before it became apparent that the approach resembled RIDL only in the dynamics of suppression. Perhaps the authors could use the "combining characteristics", because features can refer to components of a construct. I also think that naming the technology so similar to an existing, but quite different approach will lead to more product confusion - the authors may want to consider a different name or acronym. **Daughterless technologies** (eg. Thresher 2008) pre-dated fsRIDL, so perhaps these should be included in the discussion and analysis - and considered for the nomenclature.

RIDD combines the characteristics of fsRIDL (self-limiting suppression) and homing drive (efficient) in terms of performance, not construction method. We have made the wording more precise in the abstract. We prefer to keep the name RIDD because this strategy (in the current implementation) is based on *doublesex*, so reference to "insects" makes sense. Further, we want to emphasize the similarity to RIDL. Even though the acronym is similar, the pronunciation is quite different, which should avoid confusion ("riddle" versus "rid").

The daughterless strategy is a carp eradication approach for causing local extinction by shifting the sex ratio of a population. This strategy is similar to fsRIDL. Wild-type alleles are converted to dominant-sterile drive or resistance alleles in RIDD, and the offspring inheriting drive allele can keep conversion in next generation, which is the biggest difference compared to daughterless technology and fsRIDL in terms of performance. Plus, female offspring won't be produced in a daughterless system, while sterile drive/r2 females can take part in juvenile resource competition in RIDD (similar to late-acting fsRIDL). Drive alleles also have advantages in "homing", so RIDD will have better performance than other strategies.

We now reference the daughterless strategy in the introduction, though it gets less emphasis because specific RIDL strategies have already been implemented.

**In general, the methods are sound and were described sufficiently to be repeated. My only concern in the methods was the scoring of drive male mating success in the mating competitive assays. The authors describe (p. 4, para 2) that "if the percentage of individuals with red fluorescent eyes was higher than approximately 60-70%, we scored this as a drive male mating success. First, setting a threshold as "approximate" is worrying. Secondly, I can see that the authors might want some leeway from the expected conversion efficiency (80% +), but some statistical basis for selecting a threshold would be reassuring.**

This is a good point. We can be more precise. For the vials where both males likely mated, the red-eye offspring rates are 53%, 50%, 42%, 28%. According to our drive conversion test data, though the average drive inheritance rate is 84.8%, and the lowest two tests were 64% and 69%, We have thus changed our threshold to 64%. If one or two of our categorizations were incorrect, the generation conclusion of no noticeable effect of mating success would still hold true.

**My final concern is the throwaway commentary that this approach would be self-limiting, without including any modelling. I believe the paper should include some modelling on how much spread of the construct (and effects) would occur into neighbouring populations with different levels and repetitive rates of release.**

Following your suggestion, we added modeling work to better support this statement (Supplementary Fig.6). Though we simulate a very extreme situation (high release ratio and high drive conversion rate and a relatively long release period), the population can still recover with the drive frequency steadily declining, which happens very rapidly for the split drive configuration.

For confinement, we model two demes linked by unidirectional migration. The individuals in the target deme can move to a neighbor deme, maximizing the chance of successful invasion. For the whole parameter range we considered, the neighbor population is safe from suppression, with only a small number of invading males present.

#### **Reviewer #2 (Remarks to the Author):**

**Summary The authors have created a split-drive DsX drive construct, using a multiplexed approach to target and disrupt a female-specific splice site. The use of overlapping generation cage trials is very responsible and the implementation of a potential novel method for calculating population size should be expanded upon and shared with the community. The methods need expanded, some experimental choices need justified (het. vs hom. Cas9), and the modeling needs completely redone.**

Thanks for your advice and comments. We have expanded our manuscript with these suggestions (see summary above and specific comments below).

#### **Page 3, Methods, Plasmid construction**

**Sentence 1 needs a reference. "...which were constructed previously."**

We've added two citations for this.

#### **Page 3, Methods, Phenotypes and Morphological Analysis**

**I didn't realize this was a split drive design until the 3rd sentence of this section. Since it is a split drive, I think the background needs to include some discussion on split drive designs, not just the classic CRISPR design. It also needs to be clarified that RIDD is a split drive design, as that's neither how the original CRISPR designs worked nor how fsRIDL is designed.**

We added a paragraph to introduce split drive and complete drive in the introduction, and have clarified that our system is split drive system. However, much of our modeling and our cage experiment is designed to approximate a complete drive system, which we have also clarified.

**The Akbari and Bier groups have done a lot with split drives, anything specifically relating to dominant effect drives or female specific drives that should be noted here? These would be the proper comparison for RIDD. (I can't think of anything that is female specific, but dominant effectors are not a new idea.)**

Akbari and Bier teams designed split gene drive systems targeting essential genes with rescue for population modification. RIDD looks like split drive due to the split construction, but it bears no more resemblance to these drives than any other of the many split drive systems constructed by various research groups. Here are some specific differences:

1. RIDD is confined even in complete drive configuration (Sup Fig.6).
2. In these split drive systems, males and females are fertile, but only RIDD males can reproduce.
3. Dominant sterility alleles is formed by end-joining in the germline stage in RIDD, but there is no dominant effect in these split drive systems.
4. RIDD is self-limiting for population suppression (Sup Fig.6). These split drives are for self-limiting population modification.

We do reference some earlier works on split drives that introduce the concept, plus the recent Scott lab *D. sukuzii* paper, which is much more relevant to this work.

### **Page 3, Methods, Phenotypes and Morphological Analysis**

**Fix capitalization in this header.**

Fixed.

**“Most lines were tested in drive heterozygotes...” - why? Any untested hypotheses on how this impacted transmission?**

Homozygotes will have more Cas9, which could slightly increase cut rates (though this effect is small when Cas9 expression is already high, as in most of our lines). We added a set of drive conversion experiments by using drive heterozygotes carrying heterozygous *nanos*-Cas9-*nanos*. Now, the drive conversion data and masculine phenotype data of all five Cas9 lines are collected from heterozygous drive - heterozygous Cas9 individuals (*dsxD/+*, *Cas9/+*). We still have the extra set of tests for the *nanos*-Cas9-*nanos* line, using heterozygous drive-homozygous Cas9 individuals (*dsxD/+*, *Cas9/Cas9*), which is representative of the situation in the cage experiments. There was no significant difference between heterozygous Cas9 and homozygous Cas9 for the *nanos*-Cas9 line (though it could certainly make a difference in other situations where Cas9 expression is lower with only one allele).

### **Page 4, Methods, Male mating competitiveness assay**

**Can we specify that the males were also *nanos*-?**

We now specify that we use *nanos*-Cas9 males and females.

**Why were competitiveness assays not performed for any other promoters, or using heterozygotes?**

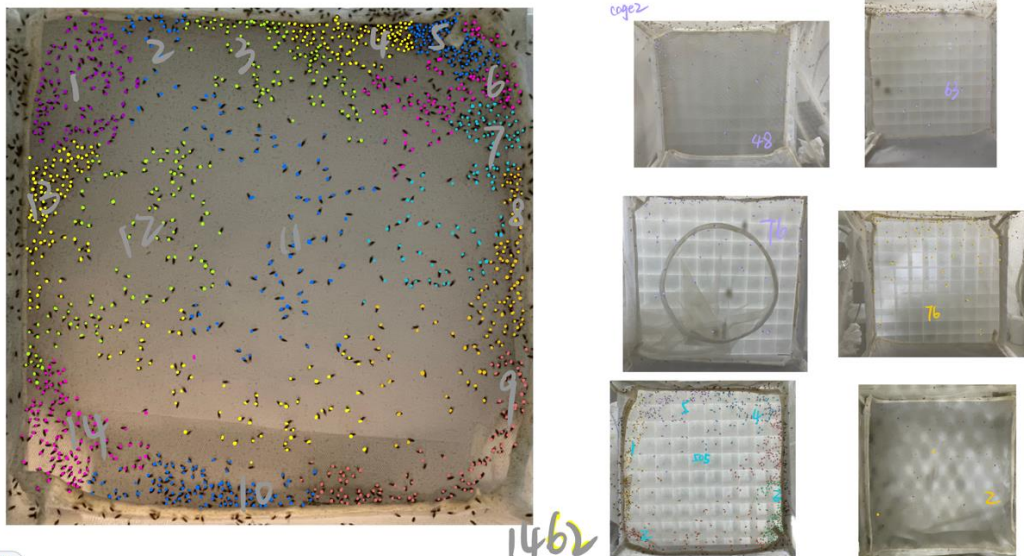
This experiment is to back up our cage study. We want to prove that there is no mating attractiveness difference between the initial cage males and released males. The genotype of released male is homozygous for *nanos-Cas9* (*dsxD/+*, *nanos-Cas9/nanos-Cas9*), the genotype of initial cage males is also homozygous for *nanos-Cas9* but without drive (*dsx+/+*, *nanos-Cas9/nanos-Cas9*). Therefore, the competitive assay was designed to compare the flies of these two genotypes.

#### Page 4, Methods, Cage study

**It sounds like the population sizes for the cages were estimates, but I don't see that noted in the data supplement. Can I also assume that the population size estimate was not used during the statistical analyses? Did you take multiple pictures to perform the population size estimates or just a single one – and if so, what's the standard error on the population size estimates? Was there any validation done on the accuracy of this method for population estimates? Maybe a reference demonstrating design (computer vision or segmentation approach?) and accuracy.**

To perform this experiment, we removed food bottles from cages, cleaned out the dead flies at the bottom, kept the cage still (so the flies would settle down and not move during the interval of picture taking), and then took photos for 6 sides as quickly as possible with white paperboard as background. Flies were then counted by looking at the pictures. The rules used for counting were: all the flies inside the edges will be counted; the flies on each edge will only be counted once; all the pictures are counted by the lead author. Thus, the error rate for all the pictures should be the same, and technically, it should be 0 because each fly should be counted. In preliminary tests, it was fully reproducible. We now specify this procedure in the Methods. We expect that any error is negligible.

The female population and drive carrier frequency were estimated phenotyping limited samples, which is less accurate. We now add the error bars for each datapoint in Fig5c and d.



**In the data supplement, Data Set S1, what is it? I can't find a description/reference anywhere. I'm guessing it's the drive-inheritance/conversion metrics, and maybe includes the data for the male mating competitiveness?**

This is a supplemental Excel file containing all the raw data for experiments in this study. Each experiment is labeled within the rows of the Excel sheets themselves. We've expanded these parts a bit, and we also refer to specific data sheets for experiments.

**Data Set S2 – only 2 cages? I don't see that noted in the methods, and I'm a little curious where there isn't at least 3? 5 would be amazing, but that's also a ton of work.**

Three repeats will be more convincing, but two cages are still enough to show the conclusion, especially considering that they performed similarly. We can observe obvious population decreasing and drive frequency increasing in both cages when we increased the drop size to ~560. Also, according to the modeling, ~560 drop size for ~2400 initial population (~1600 male population) per week exceeds the predicted suppressed threshold at drive conversion 0.69. These results together show that the suppression of cages was convincing.

Because this was repeated release-overlapping generation cage study, the workload of continuing 3 cages (2 experiment, 1 control) was substantial. We realize that only two cages may be less than the reviewer is used to seeing for this kind of study. However, please keep in mind that the population sizes of our cages tend to be substantially higher than those used by most other laboratories. We thus believe that two of these larger cages has the same experimental power of several smaller cages. While each cage in some ways could be considered a separate experiment, total experimental power may be perhaps more closely (if not completely) represented by the total population size of all cages combined.

**Data Set S3 – were the “Weekly Release size” estimates as well, or exact counts? Did I miss a note explaining the “NA” in week 2? The control cage was stopped at week 10 – why?**

Weekly released flies were exact counts.

For “NA” in week1, we initially planned to collect data every two weeks initially, but then decided to more closely monitor the dynamics of the cage population (this was our first time conducting an overlapping generation cage study). We thus switched to collecting data weekly.

We considered the population in the control cage to have reached equilibrium for approximately 9 weeks, so we did not continue. At this point, fluctuations in the cage may have still been possible due to variation in food quality or other similar factors, but by allowing this cage to continue for nine weeks, well past the lifespan for most flies in the study, we showed that the population is robust in the absence of the RIDD suppression intervention.

## **Page 5, Methods, Population models**

**The population model is discrete generation, but I didn't get the impression from the “Cage study” that it was discrete generation. Was it discrete generation, and I missed, or does your model not match your experiment?**

In the previous version manuscript, we built a discrete-generation model for general modeling while using an overlapping generation model for the cages. To better illustrate the RIDD system and match the cage study, all the models we present now are overlapping generation (Fig5,6, supplementary Fig6).

**I'm really confused what your first model is good for – it's not a split drive, it's a population frequency model, it's discrete generation. It neither matches your experiments nor reality (which I think it is used to describe later). I think we could claim it matches the experiments b/c they were performed in a homozygous Cas9 background – that should be stated with the “unlike our experiments ..” as a simplifying assumption. The model is assumed to be “...a complete drive...” which is unsupported by anything in the methods, and not supported by the “Drive performance ...” section in the results (looks like max 85%?). If the purpose of this model is to fit data from the experiments, I think it needs to reflect the experiments. If it's to perform predictive simulations, I think it needs to reflect reality. At the moment, it's does neither.**

The idea behind this model was to compare theoretical a new measure of theoretical efficiency for different genetic strategies. While not designed to match reality, we believe that this new method (“constant population genetic load) is the best way to compare different strategies in a way that does not involve any ecological parameters. It is somewhat analogous to “genetic load” measurements for gene drive, but not quite as good. This is because, by necessity, a continuous release strategy is more dependent on ecological and species lifecycle characteristics than a single-release self-sustaining gene drive strategy (where the power can be measured only as a function of the drive performance in many cases). Yet, it still allows easy comparison of suppressive power, which we believe is interesting and relevant to a wide range of scenarios. It certainly shows the substantial advantages of RIDD, and this correlates well with our more realistic simulations, showcasing the utility of this new method. In the new version of the manuscript, we expand the comparison including RIDD with different drive conversion, fsRIDL-drive with different drive conversion, as well as fsRIDL and SIT (Fig.6a).

However, we certainly agree that we could better support our claims with additional modeling that is a closer match to reality than our previous use of discrete-generation models for the second part of the figure. We now introduce an overlapping generation model that is more similar to the one we used for the cage, including RIDD-complete drive, RIDD-split drive, fsRIDL, and SIT. Besides being more realistic, it can show the efficiency and release rates required for population elimination (Fig.6b-i heatmaps, the 2<sup>nd</sup> model in Methods Population modeling). While good for comparisons, keep in mind that actual results will heavily depend on a species' ecology, plus any fitness effects on released individuals.

We still use an overlapping generation panmictic model to fit the cages, using fly-specific life cycle/experiment drop size/ecological parameters/experiment fly genotypes, which can reflect the experiment. (Fig.5c, Methods-Modeling of the cage study)

As a side note, the deterministic model is for complete drive, but our design is split drive, so we explain that it as “unlike our experiment” in modeling part. However, our cage is with homozygous Cas9 as background, so the population dynamics are actually closer to complete drive. We have deleted the sentence, “unlike our experiments, which utilized a split drive” to avoid confusion, and we have increased the clarity of our description. Keep in mind that while we build a split drive, our aim is very much to assess the full range of RIDD strategies, both complete drive and now split drive, which is included in our expanded modeling.

**SLiM – “...single panmictic population of diploids...” – so not split-drive?  
“Remaining wild-type alleles are all converted to resistance alleles.” – So the SLiM simulations also assume “a complete drive”?**

Before, we didn’t model split drive. Now, we have built SLiM panmictic models for both split drive and complete drive in the updated manuscript (Fig.6). For the cages, there is a Cas9 background, so population dynamics are equivalent to complete drive.

**Are there 2 sets of simulations using SLiM? It seems like there was one set, with discrete generations (“Stochastic simulations were performed in SLiM 4.0, similar to previous ...”), which I can’t tell where/how this was used, and then another set of simulations (“A SLiM framework was also used to simulate the dynamics...”) using weekly steps and some ecological models. Would it be possible to clarify this and specify what each model was used for? Why did the first set of simulations not use density-dependent death on the eggs? Why the different distributions on eggs (Binomial vs Poisson? At  $E[x]=2$ , they are significantly different)?**

Yes. Two sets of simulations using SLiM (general RIDD-full model and cage model) were in last version of manuscript, with some differences based on a different “lineage” for these models. In the updated version, there are seven SLiM models for RIDD system. Here is the summary.

- ◆ Four general population model variants: overlapping generation panmictic models for RIDD-complete drive, RIDD-split drive, fsRIDL and SIT. These models are more realistic and can predict the release ratio at different conversion rate. (Fig.6b-i, the 2<sup>nd</sup> paragraph of Methods-Population modeling).
- ◆ One model for cage: overlapping generation panmictic model fit for *Drosophila melanogaster* cages, adapted from general RIDD split model. In this model, we consider fly life cycle/experiment drop size/ecological parameters/experiment fly genotypes, which can reflect the experiment. (Fig.5c, Methods-Modeling of the cage study)
- ◆ One model for confinement: model with two demes connected by unidirectional migration (adapted from general RIDD-complete drive) .
- ◆ One model variant for self-limiting drive: similar to RIDD in the first two variants, but cease release and then measure how fast the population recovers, aiming to prove the self-limiting property of complete and split RIDD systems.

All above models are overlapping generations, and the distribution of egg number is now Poisson for everything. We consider density-dependent in larva competition for all models.

**“100 drive males were released into ...” – Why? The cages trials didn’t do this, correct? Why the increase after those 10 weeks?**

We redid the cage simulations. Initial population is 2400, weekly release size is now exactly the same as the size in experiments.

During the course of the experiment, we found that the initial release amount was not enough to suppress the population, we increased it.

**“The low-density growth rate was 10, roughly estimated ...” – how was this estimate performed? A short note (<1 page) in the supplement of how that was done would be plenty.**

It is 8 in updated manuscript (We update the code). This is the value that best matches the cage experiment tested in simulation. We adjust the ecological model parameters to make the simulated data more consistent with the experiment data, and then report those adjusted ecological parameters. We cannot test all internal parameters of our cage in these experiments (indeed, they may vary over time with different food quality, density conditions, and male rearing conditions), but these adjusted parameters used in the simulation can reflect the actual situation of the cages. We have clarified some cage aspects in the methods, results, and discussion to reflect these uncertainties.

**“Because the effective population size...” – Interesting, I’ve never seen this issue noted in experimental settings before. Can you explain how your experimental setup matches the Wright-Fisher model used to demonstrate this effect? If this result applies here, how come that model wasn’t used for parameter estimation and simulations here?**

The effective population size ( $N_e$ ) refers to the effective number of individuals in a population that contribute offspring to the next generation, based on the actual genetic diversity in the next generation. Generally, the effective population size is often considerably smaller than the actual population size due to various genetic, inbreeding, and demographic factors. For example, there might be 500 males, but if only 100 actually mate, the effective population size is far lower. Effective population size can partially explain the fluctuation and stochasticity of the cage. Thus, we also consider  $N_e$  in our cage simulations (Fig.5). We did not use maximum likelihood to estimate  $N_e$  population, since that estimate method is only available for discrete generations and could not be easily adapted. Thus, we set it to 5% of the total population (based on previous studies, but this should only be considered a rough estimate) and use this assumption in our cage model. This does not have a large effect on the ultimate outcome. It only affects the average level of fluctuations from a hypothetical “deterministic” result.

However, we don’t directly use the  $N_e$  concept in other modeling (Fig.6, and supplementary Fig.6), since these models are general models for all diploid organisms.  $N_e$  is largely based on species life cycle. One could envision these models as already being based around the true  $N_e$  of a certain situation.

There is no requirement to match  $N_e$  with a Wright-Fisher model. That just makes it easier. Our model is non-Wright Fisher model to allow for changes in population sizes based on how it is actually regulated - a consequence of the balance between individual reproduction and individual mortality. This is just like in natural populations, rather than being dictated by a predetermined population size as in Wright-Fisher models.

We added some notes for  $N_e$  considerations in the Methods-cage study in the updated version manuscript.

#### **Page 6, Results, Drive performance and ...**

**I don't see any note in here that some lines were homozygous Cas9 and some heterozygous. That's an important caveat that needs to be repeated with the results, possibly with some discussion on how it impacted said results.**

We did miss an important experiment before (drive test combined with heterozygous *nanos*-Cas9). We added this experiment result now and annotated with "Homo., Heter." below each data group. See Fig1c and supplementary Fig.3. There is no significant difference between hom. *nanos*-Cas9 and het. *nanos*-Cas9 in this experiment.

#### **Page 7, Results, Figure 1**

**Following on from the Methods, Phenotypes ... section – how come some lines were done as Cas9 heterozygotes, some as homozygotes, and the cage trials as homozygotes? It seems like an inconsistency, which might be part of the discontinuities with the models.**

We now clarify which test correspond to het./hom.Cas9. in Methods, Results and Fig.1 now. For homozygous *nanos*-Cas9 that represent the situation in cage experiments, it is also the same as the simulation setting (Fig5c). Homozygous Cas9 was used in all cage individuals (with modeling based on homozygous crosses) because this would mimic the performance of a complete RIDD system in natural population. If we released such flies into a wild-type cage population, then it would have been analogous to using a split RIDD system.

**Looking at 1a, this drive uses multiplexing? That should be noted in the intro, with appropriate references.**

Yes, we use multiplexed gRNAs. We added this background information for multiplexing gRNAs in the introduction with references, and also noted that our design uses 3 gRNAs in the last paragraph of the introduction.

#### **Page 8, Results, Figure 2**

**Panel "2a" uses "drive", while panel "2c" uses "D", and the description uses "D". All changed to "D" now.**

**I appreciate the note on homozygous Cas9 line used here.**

**Section "2b" needs to have "dsx" italicized – "female-specific dsx transcripts."**  
Italicized now.

**Suggestion – It would be nice to have panel “b” on top, relabeled as panel “a”, with panels “a” and “c” moved to the bottom and labels updated. That would make it consistent with figure 1 and would fix the strange “a”/”c” panel combination on top, so that it flows from top-to-bottom, then left-to-right.**

Sounds good. Panel has been changed.

#### **Page 9, Results, Resistance allele formation and classification**

**“...and targeted by three gRNAs.” – This notes that it is a multiplexed drive, which hasn’t been stated up to this point. That literature needs included in the intro, as part of the ethos behind the construct design, and what the expected impact is (why complicate something, unless you get a benefit?). Were any tests done on single or double gRNA constructs, to determine that 3 gRNAs was necessary for the best conversion rate?**

We added some literature for multiple gRNAs in introduction. Also, we added some discussion on why we use multiple gRNAs in Results-Homing drive design targeting *dsx*. Multiplexed gRNA targets ensure that frameshift mutations or disruption of sexual alternative splicing of the female-specific exon would firmly disrupt the *dsx* function in females. Even though the target site did not show functional resistance in *Anopheles gambiae* small cage experiments with one gRNA, we wanted to ensure that we avoided this. Modeling in one reference also indicated that drive conversion would be somewhat increased, but this was a secondary consideration. Multiplexed gRNAs are very easy to construct in *Drosophila*, so this complexity did not create any risk in our design.

We did not make 1-gRNA or 2-gRNA constructs. Though these could have been interested as controls, we were confident that even if the 3-gRNA design did not provide benefits with the moderate population sizes we used, then it would at least not have any substantial drawbacks. Due to the high workload involved with assessing multiple drive variants, we thus attempted to proceed directly to what we thought the best drive design would be.

#### **Page 9, Results, New pest control strategy RIDD system**

**Need to move the header to page 10 (might be a note for printing later).**

Paragraph length has changed a lot in the updated version. All lengths will be rechecked in the final version (though the journal will likely rearrange everything before publication).

**“...high total germline cut rates in males...” – where was this measured? Figure 1 only shows transmission with no discussion of cleavage. Is this why your models assume “a complete drive”?**

Supplementary fig.3 shows the masculine phenotype of female non-drive offspring of male drive heterozygotes (showing drive only to compare and assess somatic expression). Masculine rate among non-drive offspring can represent the fraction of alleles that were converted to dominant resistance alleles, for those wild-type alleles that did not undergo drive conversion. These are typically over 90%, so the total cut rate (including both these and drive conversion, which happens to about 70% of wild-type alleles in the first place) is close to 100%, especially since some resistance alleles may not have a dominant phenotype.

These considerations are separate from whether the drive is split or complete.

**“It is more powerful and efficient than fsRIDL...” – based on what? Your intro simply states that fsRIDL requires mass-rearing facilities. The current paragraph states “...which are released continuously in large number into a population...” – which is exactly what fsRIDL requires. The effect on mass-rearing needs to be clarified, as well as the expected real-world performance difference between RIDD and fsRIDL. Are smaller releases required, and if so, by how much? Is RIDD easier to rear, why? I think the actual improvement is in transmission rates from males, which is kinda alluded to in the intro “...it has been proposed to combine fsRIDL with drive”.**

See the deterministic model and general model in Fig.6 for direct comparisons. To better prove that RIDD is more powerful and efficient, we build a fsRIDL panmictic model for comparison. Both RIDD and fsRIDL required repeated release (mass rearing). According to the modeling result, the required release ratio for population elimination with fsRIDL is  $\sim 2.9$ , while for RIDD, it is  $< 1.4$  (even smaller with drive conversion, see heatmaps). The difference is because the RIDD male can produce a lot of dominant  $r_2$  alleles by end-joining, even with no drive conversion, but fsRIDL allele is inherited following a Mendelian manner, which is required for female lethality.

RIDD is not too easy to rear because RIDD females are sterile, but we discuss the future possibilities to solve this problem (see discussion the 3<sup>rd</sup> paragraph). More importantly, we model RIDD releases as heterozygotes because of this, rather than homozygous like fsRIDL.

fsRIDL-drive is another strategy that improves suppression efficiency. We also compare RIDD with fsRIDL-drive for theoretical power (Fig.6a). RIDD is better than fsRIDL-drive system, though both converge to the same efficiency when drive conversion is 100%. While a full comparison of fsRIDL drive is beyond the scope of this paper (the system has not been thoroughly investigated in other studies, and we want to avoid repeating new data on this system), we give a preprint reference for more information on this combination. Fig.6a should be sufficient for this, considering that it's results for everything strongly correlates with the modeling in the rest of this figure.

**“...allowing for highly confined population suppression...” – theoretically, I see what you're trying to say. Any proof that this is real? Are there any requirements on migration rate to make this true? Any restrictions on the size/frequency/total amount of releases to prevent spillage into nearby populations? Any caveats a local government should be aware of before attempting to apply this method?**

We added a migration test model in the updated manuscript to show this directly. We model two demes linked by unidirectional migration. The individuals in the target deme can move to the neighbor deme. Even when migration rate reaches up to 10% per time step (an enormous number considering that the generation time in these simulations was  $8/3$  time steps), the maximum drive carrier frequency is less than 18% and the minimum fertile female population is higher than 42000 (initially 50000) in the neighbor deme. Thus, the neighbor population is safe from RIDD-based suppression.

**Page 11, Figure 5**

**“...fly populations with overlapping generations” – This needs to be clarified in the methods and then which simulations does it match? At least 2 of the simulations use discrete generations, as well as the Wright-Fisher model.**

All the SLiM models in this manuscript are overlapping generation and follow a non-Wright-Fisher model. We clarify this background modeling information in the Methods now. We also added a short description for non-Wright-Fisher models, where population regulation is a consequence of the balance between individual reproduction and individual mortality, just as it is in natural populations, rather than being dictated by a predetermined population size as in Wright-Fisher models.

**“For the first 10 weeks...” – okay, so here is the 10-weeks of releases I questioned in the model earlier. Why aren’t these noted in the methods for the cage trials? Why the range of 3-7%? B/c of the population size being an estimate, and if so, why the massive variance when you can more accurately hit 20% later? Given the release sizes here (3-7%, then 20%), why the difference with model releases (100 flies, then 600 flies)? Given an actual simulated population of 2500 (right? Or is that pre-reduction? And if so, did you reduce release sizes accordingly?) (I’m assuming the 3 set of sims, the second SLiM ones), 100 flies is 4% and 600 is 24%, not exactly in-line with the experimental setup.**

All release ratios here refer to the ratio of released males to initial male population. In the cage study, ~125 was average released males among first ten weeks, and ~560 was the weekly released later on, based on the initial population of about 2400 (which was 2/3 males). For the first stage release ratio, 3-7% is a range description, and the exact average is about 5%. It varied because we collected a variable number of males during our rearing process, and we released all of them. For the second stage, the release ratio is 23%. Since we found that the initial release size was not enough to suppress the population, we increased it to ~500. In modeling, we now use the exact number of released flies.

**Panel “a” – what’s the error on the population sizes? That should be plotted.**

Because the flies were counted individually and probably completely, error was negligible. There may well be some small error in this process, but the actual error bars (if such information could be obtained) would be invisible because they’d be under 1%.

For panel c, female population size, which is estimated by the aspirator phenotype data (female fraction) multiplied counting population size, there should have error. We have added the error bar for it (Fig.5c).

**Panel “c” – “This data was in agreement with ...” – is that a measurable/statistical agreement or you looked at the plot? T0 doesn’t match, T8/9 is questionable, the drop at T11/12 is not matched but the sims eventually catch up around T20. It’d be nice to have a**

**distributional measure (or time-series measure) to quantify the similarity between sims and cages.**

It would certainly be nice to have a more rigorous analysis here, but our previously developed maximum likelihood method can only be used for discrete generations in its current implementation. We are not aware of any other method that is commonly used for genetic control cages. For example:

Li F, Yamamoto A, Belikoff EJ, Berger A, Griffith EH, Scott MJ. A conditional female lethal system for genetic suppression of the global fruit crop pest *Drosophila suzukii*. *Pest Manag Sci*, 77, 4915–4922, 2021.

This paper does not even have a model for their cages. Just simple comparisons with control cages, which are fairly self-evident. Also, note that their control cage had a different population throughout the experiment (these things can fluctuate based on environmental factors - probably accounting for the different equilibrium population between our control cages and the first part of the experimental cages).

Kaduskar B, Kushwah RBS, Auradkar A, Guichard A, Li M, Bennett JB, Julio AHF, Marshall JM, Montell C, Bier E. Reversing insecticide resistance with allelic-drive in *Drosophila melanogaster*. *Nat Commun*, 13, 1–8, 2022.

Li M, Yang T, Bui M, Gamez S, Wise T, Kandul NP, Liu J, Alcantara L, Lee H, Edula JR, Raban R, Zhan Y, Wang Y, DeBeaubien N, Chen J, Sánchez C HM, Bennett JB, Antoshechkin I, Montell C, Marshall JM, Akbari OS. Suppressing mosquito populations with precision guided sterile males. *Nat Commun*, 12, 1–10, 2021.

These two papers didn't even try to do any statistical analysis for cages, They just put a bunch of simulations next to the cage, allowing readers to roughly compare the two (even though it involved discrete generations, which would have made a statistical analysis somewhat easier).

Simoni A, Hammond AM, Beaghton AK, Galizi R, Taxiarchi C, Kyrou K, Meacci D, Gribble M, Morselli G, Burt A, Nolan T, Crisanti A. A male-biased sex-distorter gene drive for the human malaria vector *Anopheles gambiae*. *Nat Biotechnol*, 38, 1054–1060, 2020.

As above, but from a different modeling lead. They used several simulations attempting to match the cage, with model parameters based on their mosquitoes and the drive performance individual crosses. No attempt was made to statistically analyze how well the model matched to the cage.

Thus, while it's not ideal, our study still tends to match “most common practice” for many cage studies, and certainly for studies involving overlapping generations. We believe that our model is a rough match to the data, or at least about as well as the model matched in the above listed studies. We make no claim that the model is perfect, only that the cages were roughly within expectations.

Specifically, we don't see any drive frequency mismatch as week 0 (population was not measured here). At week 8/9, one of the cage frequencies and one of the cage population sizes

was a good match to most of the model lines. The population in one cage was a little high, but this was clearly a temporary fluctuation. One of the cages has a somewhat low drive frequency, but it was in line with the most extreme simulations in the model. At week 11/12, it seems to be a pretty good match to us, though around week 10, the drive frequency in the cage is a bit low before catching up quickly and again matching model. At week 20, both cages had an unexpected downward fluctuation in drive frequency, but after this, their performance was still close to the model, just lagging a bit behind (a short-term change can make the cage and model lines diverge even if the genetic construct is performing as modeling in most time points). Even with worse drive parameters, such a downward fluctuation was likely a stochastic effect, or perhaps the released males were just reared in slightly too crowded conditions for a brief period, hurting their performance. At worst, we could potentially say that we overestimated the effective population size in the model, which would result in more fluctuations in the simulation. This was only a rough estimate in the first place, though, so we are overall reasonably pleased by our model's match. A rigorous statistical analysis would likely indicate that the cages are statistically different from the model, but it is unclear how much value this would produce considering the complicated situation. A rough match is probably the best that can be hoped for in most population cages, especially for suppression. Again, these differences between experiment and model are typical of genetic control cage studies.

## **Page 11, Results, Assessment of the RIDD in cage populations**

**“...all flies were homozygous for the split...” – was it ever clarified why you picked this line to do cage trials? Were any other lines tested in the mating competitiveness assay?**

*nanos*-Cas9 line performs better than others overall:

1. high drive conversion rate. (Fig.1c)
2. highest germline resistance rate. (Supplementary Fig.3) Masculine rate of non-drive offspring is the highest, which represents high cleavage rate and high dominant resistance formation rate. The core idea of RIDD is dominant resistance allele, so we wanted a line with high resistance formation rate.

We did not test other Cas9 lines in mating competitiveness assay, since this experiment aimed to prove there is no attractiveness difference between the initial cage flies and released flies.

It's possible that another Cas9 would have been slightly better with one of our tested promoters or another in our lab (though these differences would be small). However, *nanos* has been our *D. melanogaster* “standard” throughout many studies, and we actually carried out some of the later experiments before the promoter experiments were even completed.

**“...was substantially increased to approximately 500 per cage...” –why don't your models reflect this? “We observed the sex ratio biased ...” – none of the models mention reflecting this.**

Rather than an approximation, the modeling now uses the exact release numbers from the experiment. We describe this in Methods - Modeling of the cage, and in Fig5c.

The sex ratio bias also happens in modeling, caused by males repeatedly released to the cages, which was obvious. Considering the female population of cages is more interesting, we only present that data. For the sex bias in the wild-type population, we include this by giving females an increased death rate at high density to match our experimental observations prior to the drive release.

**“..which was approximately 1200 flies.” – why was the control cage so much smaller? It’s nearly ½ the size of the test cages. Somewhere else I already asked why it was only kept for ~1/3 the time of the cage trials.**

It’s a little unexpected that it didn’t reach the same size. However, see above references and responses. The control was done at a different time, and like other studies, may therefore have had a different carrying capacity at different times. Maybe it would have if we gave it longer. However, it was done at a different time, with different food (which we know is variable). Thus, the exact time is of less importance in the first place. The more important thing is that this cage demonstrated that the fly population is robust in our overlapping generations model. The population would not decline to low levels in the absence of the RIDD releases.

**“...using estimates for several parameters related to the ...” – why only simulate 1 resistant allele, when you have population measures for 3 of them?**

In experiment, we found three types of resistance alleles, dominant with moderate intersex, dominant with stronger intersex, and recessive with mild intersex. The first two are the same for modeling (they are both dominant sterile, just different in appearance). It will be more realistic if we considered recessive resistance allele (as well as perhaps a small number of remaining wild-type alleles), but such recessive resistance alleles are only 5.3% of all resistance alleles (possibly less). We thus built only a simple model, assuming as an approximation that all the resistance alleles were dominant sterile. This very slightly increased drive performance in the model, possibly explaining in part why the model performed, on average, a little better than the actual cages.

**This section also assumes 100% cleavage – need the data shown/noted somewhere.**

See above explanations. Supplementary fig.3 shows the masculine phenotype of female offspring of male drive heterozygote. Masculine rate among non-drive offspring can represent the ratio of germline resistance formation rate. For homozygous *nanos*-Cas9, total masculine rate of non-drive female is 92%, and drive conversion rate is 69%. Thus, the cleavage rate is >97.5% (some additional alleles may be recessive sterile). For heterozygous *nanos*-Cas9 test, the apparent total cleavage rate is even higher at >99.6%. Thus, to keep the model simple, we assumed 100% cleavage.

**“The trajectories ... closely matched ...” – “closely” needs quantified. KL-divergence or earth-movers distance for distributions, or a time-series approach to quantify this. There may be better methods out there. A binomial test at each datapoint is inappropriate for this – it fails to account for the correlated nature of each datapoint.**

Please see above comment. While it might be useful to have a better statistical analysis, our manuscript's methods follow most common practice in the field.

## Page 12, Results, Modelling of RIDD system

**“For RIDD and similar technologies such as SIT and RIDL...” – RIDD is a split-gene drive, SIT and RIDL are not the most related comparisons, you should compare to other split-drive designs (Akbari has some, Hay has some).**

RIDD is actually a half split-gene drive, since only male heterozygote can help spread drive allele. We compare it with SIT and fsRIDL, because of repeat release strategy as well as the self-limiting property. In the updated version manuscript, we also provide a comparison between RIDD and the fsRIDL-drive, which is the most similar genetic control strategy to RIDD.

**“Here, we introduce a possible method of comparison based on genetic load...” – isn't that published here - <https://doi.org/10.1111/eva.12583>? I believe there is another paper that introduces this too, written by several Asian authors who used Mathematica to create a unified framework for testing and measuring several different gene drives (I'm really sorry I can't remember anything more useful about this paper).**

While showing a unified framework for (mostly self-sustaining) comparing gene drives in some ways, that manuscript discusses different concepts, and can still only handle a limited selection of the total possibilities. As far as we are aware, “constant population genetic load” of repeated release genetic control methods is presented here for the first time, and nothing in that manuscript or the one indicated in the doi above is similar to this.

**“We developed a deterministic, discrete-generation ...the wild-type population does not change...” – Nothing in the remainder of this section matters. These simulations don't reflect reality and they don't reflect the drive design presented in this paper. This section needs to be removed or rewritten to be correct and applicable, not a misleading statement on the possible efficacy of some other design in an impossible population.**

Please see above for a detailed discussion of this method, which we believe is useful for assessing the power of a drive system, even if it does not try to represent the actual reality of population suppression. We have also added more realistic simulations to show most comparisons for both complete and split RIDD, together with fsRIDL and SIT.

## Page 13, Discussion

**“...and homing suppression drives that are dominant-sterile due to somatic expression.” – this is old news, and has been ameliorated through better target choices and the use of germline-specific promoters (such as the nanos promoter implemented here).**

Keep in mind that while the somatic expression problem has been solved for *Anopheles* mosquitoes, it may not be easily solvable in other species. In these cases, alternate pest control methods, even if less efficient, might still represent an improvement over existing pest control

methods. Further, there may be times when a self-eliminating pest control system is required. We thus believe that this is a worthy comparison for our discussion.

**“Our overlapping-generation cage study...” – I don’t think I said it earlier, but I really appreciate this cage trial design. Very good choice, I hope it propagates through the field.**

Thanks for these comments. While discrete-generation cages have their uses, this is our first foray into repeated-release genetic control methods, and we thought that overlapping generations would potentially be even more important for this situation. We are glad that our cages seemed to be a good model system.

**“...increasing to approximately 50~60% ...” – this is a huge percentage of the population. Is this even feasible in a real population? It seemed like ~20% was the minimum effective release size to crash the population, how feasible is that?**

Typical SIT and fsRIDL program release numbers of males equal to many times the wild population. Needless to say, this only provides localized suppression, but continental scale elimination was eventually achieved with screwworm SIT, despite lower efficiency requiring huge releases. It might seem like a lot, but it could still be less expensive than alternatives.

**“RIDD is substantially more efficient...” – how much more efficient in a population w/o Cas9? Then it’s no longer a Cas9 homozygous background, so the sims won’t match.**

We now simulate both RIDD split drive and RIDD complete drive in our updated manuscript (Fig.6). For split drive, performance is similar until around 50% drive conversion efficiency, at which point the performance of split drive does not improve (though it’s still quite efficient compared to existing measures). Keep in mind that we are not proposing to use our specific design in *D. melanogaster* to suppress wild populations. Instead, this manuscript shows methods that can potentially be generally used in other species, where scientists will have the option to design either complete or split systems. We thus wanted to show the performance of each type.

**“If releases are ceased at any time ...” – can you prove this? This is a male-only drive, so it won’t propagate well, but how many generations before the population crashes do we need to stop releases?**

We now directly show this with modeling (see above for details).

**How long do we expect it to hang around in the population? Is this impacted by the split-drive nature of your design?**

We now model this explicitly, see Fig 6fghi.

**“...highly confined population suppression...” – that is not demonstrated here.**

See above. The data is in new Supplementary Fig.6c.

## Reviewers' Comments:

### Reviewer #3:

#### Remarks to the Author:

I have been asked to comment on this manuscript following the first review, and in particular if the first two reviewers concerns have been addressed. Obviously I cannot speak for those reviewers but here is my take based on the revised ms.

I agree with both the reviewers that this is an interesting and impressive study and I think it is suitable for the journal. The first reviewers comments were relatively minor and I believe they have been adequately addressed. The second review is very detailed and I believe the authors have adequately addressed the numerous minor comments. The major concern of this reviewer relates to the modelling, finding it to be confusing and ill-matched to the empirical studies. I can see the authors have made a number of revisions in response to these claims (even if I have not seen the original), though even in this version I share some of the reviewers concerns. My impression is the various modelling components of the manuscript could be streamlined and better explained. I will elaborate below.

The authors use a variety of models to address different questions, adding considerable complexity to an already complex study. I would like to see these modelling aims spelled out clearly in the MS (preferably Introduction), but what I am reading is: 1. Modelling to demonstrate understanding of the cage dynamics; 2. Modelling to give a general appraisal/comparison of their RIDD system with various types of similar genetic control systems, that are independent of species-specific ecology; 3. Modelling to give a more specific appraisal/comparison of their RIDD system against the other systems assuming the focal species; 4. Modelling to demonstrate self-limiting nature of RIDD. Please correct/ revise this summary if I am missing something.

#### 1. Modelling to demonstrate understanding of the cage dynamics (stochastic model)

I am assuming the purpose of modelling the cage populations is to demonstrate this understanding, but this is not clearly stated (unless I missed it), so it would be helpful if this was clarified. I am puzzled as to why the authors don't try to do more here – for instance to re-estimate the parameters using an ABC method. I did not see discussion of the (admittedly not large) discrepancy between simulation and cage dynamics (Fig. 5d), which ABC parameter estimation may shed some light on.

#### 2. Modelling to give a general appraisal/comparison of their RIDD system with various types of similar genetic control systems, that are independent of species-specific ecology (deterministic model).

I understand the authors intentions here, but I am unconvinced that the pay-off from having the additional model is worth its cost. “Constant population genetic load” is not an intuitive enough metric, and I don’t find this analysis is giving any great additional insight over that of the stochastic model. I suggest dropping this part.

3. Modelling to give a more specific appraisal/comparison of their RIDD system against the other systems assuming the focal species

I think this is the most important of the modelling components in the MS, and the authors approach is generally sensible. The presentation of results in Fig. 6b-i could be a bit clearer. My suggestion from above would lead to fig. 6A being dropped – would it then be possible to replace this with a similar figure produced from the stochastic model? The rest of fig 6 is generally nice but the colour scheme (especially the red) is a bit hard to read. In 6d (SIT) why is it lighter near the bottom (indicating more suppression from a lower release ratio??), and how should I interpret the grey at the top??

4. Modelling to demonstrate self-limiting nature of RIDD.

I understand this is a new part of the manuscript and I find it sensible. I find it a shame that the figure is a SI fig – I understand the competition for space but it is demonstrating quite an important point. Is there any way it could be squeezed into the main manuscript?

Overall comment on modelling.

Aside from my specific comments on the modelling components above, I feel the modelling methods section could be improved. The assumptions of the stochastic model variants (cage model, panmictic and two-population models), could be clearer. For example, on reading the sentences

“In any time step, the population is defined by the numbers of male and female adults of each genotype. To determine the number of offspring produced, each female first selects a random male in the population to mate with.”

I was left wondering: a. what is the time step. b. Does this mean we assume each female re-mates in each time-step (and so potentially mates numerous times through her life)?

Moreover, I suspect a better job could be made of relating how the stochastic models are varieties of the same over-arching model. (Not sure whether it would be better to

start with the cage and then the panmictic or vice-verca??). I think this streamlining is a matter of tweaking the text, but would simplify and clarify the modelling methods.

In the following, editor and reviewer comments are in bold, and our responses are in plain text.

## **REVIEWER COMMENTS**

### **Reviewer #3 (Remarks to the Author):**

**I have been asked to comment on this manuscript following the first review, and in particular if the first two reviewers concerns have been addressed. Obviously I cannot speak for those reviewers but here is my take based on the revised ms.**

**I agree with both the reviewers that this is an interesting and impressive study and I think it is suitable for the journal. The first reviewers comments were relatively minor and I believe they have been adequately addressed. The second review is very detailed and I believe the authors have adequately addressed the numerous minor comments. The major concern of this reviewer relates to the modelling, finding it to be confusing and ill-matched to the empirical studies. I can see the authors have made a number of revisions in response to these claims (even if I have not seen the original), though even in this version I share some of the reviewers concerns. My impression is the various modelling components of the manuscript could be streamlined and better explained. I will elaborate below.**

Revisions summary:

1. We estimated several parameters for the cage models using RMSE to identify the optimal set and provide a clearer illustration of the cage dynamics.
2. We reorganized the modeling results, moving the constant-population genetic load figure to the supplementary material and relocating the self-limiting figure to the main body.
3. We made minor adjustments to enhance the manuscript's readability, including a distinctive color scheme for heatmaps and more specific legend descriptions.
4. We rearranged the order of the modeling method descriptions and streamlined the wording for clarity.

**The authors use a variety of models to address different questions, adding considerable complexity to an already complex study. I would like to see these modelling aims spelled out clearly in the MS (preferably Introduction), but what I am reading is: 1. Modelling to demonstrate understanding of the cage dynamics; 2. Modelling to give a general appraisal/comparison of their RIDD system with various types of similar genetic control systems, that are independent of species-specific ecology; 3. Modelling to give a more specific appraisal/comparison of their RIDD system against the other systems assuming the focal species; 4. Modelling to demonstrate self-limiting nature of RIDD. Please correct/revise this summary if I am missing something.**

We added a short summary of modeling at the end of introduction. We really just have three basic types of modeling.

1. Stochastic SLiM-based modeling to directly support the RIDD cage experiment, using specific ecologies appropriate for this.
2. General SLiM-based modeling to assess RIDD and other systems, which we used in multiple different ways (Figures 6-7, Supplementary Figure 7) to show basic performance, confinement, and the self-limiting nature of the drive (there were slight variations in these, of course, but they used the same framework).
3. A deterministic model to find constant-population genetic load. This has been de-emphasized and moved to the supplementary material.

### **1. Modelling to demonstrate understanding of the cage dynamics (stochastic model)**

**I am assuming the purpose of modelling the cage populations is to demonstrate this understanding, but this is not clearly stated (unless I missed it), so it would be helpful if this was clarified. I am puzzled as to why the authors don't try to do more here – for instance to re-estimate the parameters using an ABC method. I did not see discussion of the (admittedly not large) discrepancy between simulation and cage dynamics (Fig. 5d), which ABC parameter estimation may shed some light on.**

Thanks for this suggestion. To better understand the cage dynamics, we now estimate some parameters by Root Mean Square Error. Fixed parameters that were accurately measured in individual experiments are not varied, but we now assess drive male mating fitness, generation duration (lifespan), and low-density growth rate. We co-vary each of these to identify the optimal parameter set that most closely matches the cage experiments, considering both drive frequency and female population size.

Though this is the most common way of assessing this sort of thing, please keep in mind that it is not necessarily highly accurate. A fluctuation early in a cage from expectation can then result in bad performance of the whole model. For example, released males may have had variable fitness in different time intervals due to slightly different rearing conditions. It is also a little hard to find a parameter set that can both perfectly match the female population and drive carrier frequency. Overall, though, our optimal parameter set was reasonable. No fitness costs were detected, and the other two parameters we close to our rough guesses from earlier. The cage simulations shown in the manuscript are now matched to these optimal parameters.

### **2. Modelling to give a general appraisal/comparison of their RIDD system with various types of similar genetic control systems, that are independent of species-specific ecology (deterministic model).**

**I understand the authors intentions here, but I am unconvinced that the pay-off from having the additional model is worth its cost. “Constant population genetic load” is not an intuitive enough metric, and I don't find this analysis is giving any great additional insight over that of the stochastic model. I suggest dropping this part.**

For all genetic population suppression strategies, success depends on not just the genetic system performance and release parameters, but also heavily on the ecology of the species, in particular the density growth curve. It is thus common to assess self-sustaining gene drives using “genetic load”, allowing a direct assessment of the drive’s power, regardless of ecology. Genetic load focuses on the fraction of sterile females (or whatever other mechanism the system uses to reduce population size), avoiding any input from density-dependent growth. Thus, genetic load estimation can be used to generally compare the suppressive power of repeated release strategies. Ideally, we would like to do something similar for repeated release strategies.

However, for repeated release strategies, it’s not just the low-density growth rate that matters, but also the shape of the density growth curve. This is because repeated releases will give an increased release ratio of transgenic to wild-type males as the wild-type population shrinks. Thus, the release size does not need to overcome the low-density growth rate when the wild-type population is at equilibrium. This complexity can make it difficult to compare different transgenic constructs in the general case. We thus decided to go a step beyond usual genetic load assessment (which is only relevant to self-limiting systems) with “constant-population genetic load”. We believe that this is probably the best way to directly compare RIDD, SIT, fsRIDL, and fsRIDL-drive in terms of suppressive power.

However, considering the opinion of all the reviewers, perhaps we did not give this concept the attention we think it really deserves. Rather than expanding this peripheral point in this manuscript, we’ll save it for a future study. We are loath to completely remove it from the manuscript, but we have moved this figure panel to the supplementary material, while also de-emphasizing it in the main text.

### **3. Modelling to give a more specific appraisal/comparison of their RIDD system against the other systems assuming the focal species**

**I think this is the most important of the modelling components in the MS, and the authors approach is generally sensible. The presentation of results in Fig. 6b-i could be a bit clearer. My suggestion from above would lead to fig. 6A being dropped – would it then be possible to replace this with a similar figure produced from the stochastic model? The rest of fig 6 is generally nice but the colour scheme (especially the red) is a bit hard to read. In 6d (SIT) why is it lighter near the bottom (indicating more suppression from a lower release ratio??), and how should I interpret the grey at the top??**

For old Figure 6A, the results are applicable for any model (this is the idea of the new concept... broadly applicable to any ecological situation, which could be thought of in terms of model type). We could use the stochastic model to derive a nearly identical figure, but all that would do is introduce some “noise” that we would be trying to reduce with high sample size. The graph would still be the same - perhaps more confusing than would be ideal for this manuscript that generally has a different focus.

We reset the color scheme for both blue and red heatmaps. The color differences between the grids is hopefully easier to distinguish now.

For 6d-SIT (6b-SIT in new version), it's a good question, and something that has been seen in previous models. Basically, this is called "overcompensation", and can occur when larval competition is fierce and the density growth curve is linear (a common modeling assumption). Having fewer larva (due to fewer females laying viable eggs) can actually result in more larval survival. Thus, a little bit of suppressive effort actually increases the population, until enough effort is applied that the population is finally eliminated.

Though a strange concept, overcompensation has been observed in the field populations of three *Aedes* species [1], though it doesn't always happen. Since numerous larvae must feed on a limited resource, fewer initial eggs may result in a better survival of the remaining larvae, leading to an equal or increased number of adults.

[1] Bouyer J. When less is more: accounting for overcompensation in mosquito SIT projects. *Trends Parasitol.* 2023 Apr;39(4):235-237. doi: 10.1016/j.pt.2023.02.001. Epub 2023 Feb 9. PMID: 36764849.

Grey means "not applicable" (NA - where population elimination always occurred for a, and where is never occurred for b). This is now added to the figure legend.

#### **4. Modelling to demonstrate self-limiting nature of RIDD.**

**I understand this is a new part of the manuscript and I find it sensible. I find it a shame that the figure is a SI fig – I understand the competition for space but it is demonstrating quite an important point. Is there any way it could be squeezed into the main manuscript?**

There's no figure limit, so we're happy to add this to the main text. It is now Figure 7.

#### **Overall comment on modelling.**

**Aside from my specific comments on the modelling components above, I feel the modelling methods section could be improved. The assumptions of the stochastic model variants (cage model, panmictic and two-population models), could be clearer. For example, on reading the sentences**

**"In any time step, the population is defined by the numbers of male and female adults of each genotype. To determine the number of offspring produced, each female first selects a random male in the population to mate with."**

**I was left wondering: a. what is the time step. b. Does this mean we assume each female re-mates in each time-step (and so potentially mates numerous times through her life)?**

a. Each time step represents one mating and reproduction cycle, labeled as taking place in one week.

b. Yes, females can re-mate and reproduce in each time step through their lives. Not all insects will do this, but it doesn't have a large effect on results (in some SIT systems, multiple mating can reduce the efficiency of the system, but we here assume that females can't utilize sperm from previous mating, so all systems are equal with respect to mating system).

We have clarified both these points in the main text.

**Moreover, I suspect a better job could be made of relating how the stochastic models are varieties of the same over-arching model. (Not sure whether it would be better to start with the cage and then the panmictic or vice-verca??). I think this streamlining is a matter of tweaking the text, but would simplify and clarify the modelling methods.**

Yes. We adjusted the order and the wording of the text, making it easier to understand.

1. Stochastic general model to evaluate efficiency.
2. Two-demes model. It uses the same basic model, just with a second connected population.
3. Variant model to simulate the *Drosophila* cage experiment. It has exactly the same basic framework, but a few parameters are a little different. These are now described more clearly.
4. The deterministic model that was used only for constant genetic load comparison has been removed from the main methods (to avoid confusion) and is now almost entirely in the supplement, with only a brief mention in the main results section.

## Reviewers' Comments:

### Reviewer #3:

#### Remarks to the Author:

I welcome the revisions to the ms and in particular find the modelling description and results to be clearer now. I only have some minor suggestions outstanding, and I noticed a few typos in the revised text.

#### Comments/suggestions

- L.213: “Individuals can survive at most six time steps, but their survival rates decline with age to produce a linear decline in the number of individuals in a single age classy” .
  - Is the decline in numbers with age class exactly linear? I can’t think of a standard survival function of age which produces this outcome. I suggest stating the survival function explicitly in the text.
- L267-270. “To better understand cage dynamics...”
  - I’m satisfied with this approach which is broadly equivalent to finding the best-fit line in a regression analysis. I agree with your comments in the rebuttal text: its not a perfect method but still worth doing. (You could make these points in the Discussion!). However, you don’t say anything about the extent of parameter space you explored. I suggest you state the ranges of the 3 parameters that are varied, either by their absolute values or in comparison to default parameters. 100 simulations doesn’t sound like a lot for exploring 3 parameters, but may be sufficient in this case (?).
  - Do the simulations shown in plots Fig 5c-d use the best fit parameters from this method? (Also: I know the simulation lines are deliberately faint, but they are so faint as to strain my eyes! Could they be slightly darkened?).

#### Typos and wording issues

- L104. parameterS.
- L214. Single age class (no y!).
- Fig. 6. You refer to population ‘suppression’ in the scale bar but elimination in the legend. I suggest using elimination here - its meaning is unambiguous.
- Lines 585-588. The English in these two sentences is not clear. “We then further estimated the parameters of cage modeling by RMSE, no drive fitness detected and ecological parameters are sensible. Variable fitness in different time intervals due to different rearing conditions of released drive males, perhaps accounting for the imperfect match, especially the early stage of drive carrier frequency.”

### **Reviewer #3 (Remarks to the Author):**

**I welcome the revisions to the ms and in particular find the modelling description and results to be clearer now. I only have some minor suggestions outstanding, and I noticed a few typos in the revised text.**

Thanks for your comments.

### **Comments/suggestions**

- **L.213: “Individuals can survive at most six time steps, but their survival rates decline with age to produce a linear decline in the number of individuals in a single age classy” .**

- **Is the decline in numbers with age class exactly linear? I can’t think of a standard survival function of age which produces this outcome. I suggest stating the survival function explicitly in the text.**

Here is the function we used. We add it to the methods.

$$\text{Current age survival rate} = \frac{\text{max age} - \text{current age}}{\text{sum}(1:\text{max age})}$$

It does produce a linear survival curve.

- **L267-270. “To better understand cage dynamics...”.**

- **I’m satisfied with this approach which is broadly equivalent to finding the best-fit line in a regression analysis. I agree with your comments in the rebuttal text: its not a perfect method but still worth doing. (You could make these points in the Discussion!). However, you don’t say anything about the extent of parameter space you explored. I suggest you state the ranges of the 3 parameters that are varied, either by their absolute values or in comparison to default parameters. 100 simulations doesn’t sound like a lot for exploring 3 parameters, but may be sufficient in this case (?).**

We included our previous comment for the cage simulation analysis in discussion. Also, we added the parameter space in the Methods, and we show the exact absolute RMSE value for the best match and worst match in the Results. Keep in mind that we use 100 simulations per parameter set, not for the whole analysis (we have clarified this in the text). Though there is a little stochasticity of the RMSE estimate with 100 simulations, it is sufficient for exploring our parameter range.

- **Do the simulations shown in plots Fig 5c-d use the best fit parameters from this method? (Also: I know the simulation lines are deliberately faint, but they are so faint as to strain my eyes! Could they be slightly darkened?).**

The simulation lines presented in Figures 5c and 5d were indeed generated using the optimal parameter set that we identified. We now use darkened color for simulation lines.

#### **Typos and wording issues**

- **L104. parameterS.**
- **L214. Single age class (no y!).**

Fixed.

- **Fig. 6. You refer to population ‘suppression’ in the scale bar but elimination in the legend. I suggest using elimination here - its meaning is unambiguous.**

All the annotations have now been changed to “elimination”.

- **Lines 585-588. The English in these two sentences is not clear. “We then further estimated the parameters of cage modeling by RMSE, no drive fitness detected and ecological parameters are sensible. Variable fitness in different time intervals due to different rearing conditions of released drive males, perhaps accounting for the imperfect match, especially the early stage of drive carrier frequency.”**

We rewrote this part, and also include some additional comments discussing RMSE.