DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY
Office for Scholarly Communication

Evolutionary dynamics of CRISPR gene drives

Permanent link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:39947206

Terms of Use
This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story .

Accessibility

Evolutionary dynamics of CRISPR gene drives

A DISSERTATION PRESENTED
BY
CHARLESTON NOBLE
TO
THE COMMITTEE ON HIGHER DEGREES IN SYSTEMS BIOLOGY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF

SYSTEMS BIOLOGY

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS

JULY 2018

©2018 – Charleston Noble
all rights reserved.

Dissertation Advisors: Martin Nowak & George Church Charleston Noble

Evolutionary dynamics of CRISPR gene drives

Abstract

The alteration of wild populations has been discussed as a solution to a number of humanity's most pressing ecological and public health problems. Enabled by the recent revolution in genome editing, CRISPR gene drive systems—selfish genetic elements that can be engineered to spread through populations even if they confer no advantage to their host organism—are rapidly emerging as a promising approach. However, before real-world applications are considered, it is imperative to develop a clear understanding of the potential outcomes of drive release in nature. Toward this aim, in this dissertation, I mathematically study the evolutionary dynamics of CRISPR gene drive systems. In the first chapter, I demonstrate that the emergence of drive-resistant alleles could present a major challenge to existing proof-of-principle constructs, and I show that an alternative design that selects against resistant alleles could potentially improve evolutionary stability. In the second chapter, I address the question of how likely it might be for a small accidental or unauthorized release of existing CRISPR gene drive organisms to result in significant spread through a wild population—despite the problem of resistance. The mathematical results in this chapter suggest that significant spread is highly likely following even small releases, and this has important implications for laboratory containment protocols and future design of field trials. Finally, in the third chapter, I study the dynamics of a new CRISPR-based gene drive system called "daisy-chain gene drive," which aims to address the issue of accidental spread discussed in the previous chapter. The results suggest that daisy-chain gene drive constructs could act as "self-limiting" drive systems, with the potential to spread to high frequency in a local population with a comparatively low risk of spreading indefinitely through many linked populations.

Contents

For Elina, with love and gratitude.

Acknowledgments

Many amazing people have helped make this dissertation possible, and I am deeply grateful for their kindness and generosity.

First, I am indebted to my advisors, Martin Nowak and George Church, who provided unwavering support and an unparalleled environment to grow and learn. Martin taught me the beauty of simplicity and the art of asking a good question, and he has been a wonderfully kind and inspirational mentor. It has been such an honor and a pleasure to work with him. Moreover, the Program for Evolutionary Dynamics that Martin leads has been a truly fantastic place to do a PhD, largely due to the example he sets. George renewed my sense of wonder at the world and expanded my view of what is possible, and working with him and his group—amid the environment of fearless imagination that he has fostered—has been an amazing experience.

I am also indebted to Kevin Esvelt, who has provided remarkable support as a mentor, as well as an abundance of fascinating and meaningful questions as a collaborator. I am excited to see where his efforts on gene drive lead, and I am grateful that I could be part of the journey.

I have also been fortunate to have a wonderful dissertation advisory committee, consisting of Andrew Murray, Jeremy Gunawardena and Michael Desai, whose excellent advice helped keep my immediate work focused in the right direction and inspired me to think broadly about the future

and what questions I was most excited to explore.

In addition to my fantastic advisors, I have been very fortunate to have an amazing group of collaborators, who have served as continual sources of support and inspiration throughout my work on this dissertation. Jason Olejarz provided excellent help and advice on deterministic models of evolution in infinite populations, and his kindness cannot be overstated. Ben Adlam generously lent his expertise in stochastic models of evolution in finite populations, and I have greatly enjoyed our collaboration. James DiCarlo and Alejandro Chavez taught me how to engineer yeast, and I am grateful for their patience and generosity. Finally, John Min has time and again contributed his experimental expertise in thinking about gene drive systems, particularly daisy-chain gene drives, and I am grateful for his help. I am also honored to have advised Adam Atanas on his undergraduate thesis; I learned much from the experience, and I am excited to see where his career takes him.

Of course, there is much more to a PhD than the dissertation. For constantly making the experience fun and refreshing, I am tremendously grateful to my amazing group of labmates and friends. In addition to everyone mentioned above, I would like to thank Jeff Gerold, Carl Veller, Pavitra Muralidhar, Sam Sinai, Ski Krieger, Anjalika Nande, Alex Heyde, Oliver Hauser, Morgan Craig, and Alison Hill from the Nowak lab; Devora Najjar from the Esvelt lab; and Jun-Han Su, Adam Carte, Isaac Plant, Silvia Canas Duarte, Jocelyn Kishi, Stephanie Hays, Cameron Myhrvold, Brendan Colon, Gabriel Filsinger, Sam Wolock, David Ding, Jim Valcourt, Adam Riesselman, Alina Chan and Marika Ziesack in the Systems Biology PhD Program and elsewhere at Harvard.

For helping me navigate academic life beyond research, I am extremely grateful to Samantha Reed and Elizabeth Pomerantz of the Systems Biology PhD program and Katherine Gallagher and May

Huang of the Program for Evolutionary Dynamics. And for providing financial support throughout my PhD, I am thankful to the National Science Foundation, in particular the Graduate Research Fellowship Program.

Next, I would like to thank my wonderful family, whose boundless love and support have made this entire experience possible. My grandfather, Bill, introduced me to science and inspired me from the very beginning; my grandmother, Geraldene, taught me poise and warmth; my cousin, Kirk, and his wife, Arisara, have been and continue to be great role models; my aunts, Carrie and Beth, have always given me loving and unconditional support, and they have my deepest gratitude; and my amazing mother, Barbara, did everything possible to give me the best life I could imagine, and I am forever grateful to her.

Finally, to my fiancée, Elina: It's been quite the journey—I can't thank you enough for the endless love, support and understanding that you've shown me along the way. I am so incredibly happy and lucky to have you in my life. This is for you.

0

Introduction

The central question of this dissertation is how one might spread a genetic trait through a wild population. This question has been studied for several decades, but I was fortunate to enter the field just as it underwent a technological revolution: CRISPR, a recently developed genome engineering technology, dramatically increased the possibilities of the field and introduced a multitude of interesting scientific questions. This dissertation addresses a few of these questions using mathematical modeling.

In addition to these scientific questions, the topic gives rise to many critical—and incredibly difficult—ethical and policy questions about whether one ought to alter a particular wild population at all. Although I do not directly address these questions here, the potential impact of the technology and the likelihood of its eventual use appear sufficiently great to at least warrant its careful study, with objectivity and a clear understanding that mathematical results are more a reflection of our thinking about reality than of reality itself, and with careful consideration of the consequences that could result from our efforts. Thus, to begin, I will discuss what motivates the field broadly and the reasons for its rapid expansion in recent years.

In short, there are a variety of wide-reaching problems that might be addressed by genetically engineering wild populations which are difficult or impossible to solve using traditional methods. These can be categorized broadly into two groups: mitigation of vector-borne diseases, and control of destructive invasive species or agricultural pests. While still in their early stages, significant experimental and theoretical progress has been made on both fronts.

Among vector-borne diseases, the most prominent potential targets are those transmitted by mosquitoes, including malaria, dengue, and Zika. All three diseases present significant burdens, and malaria alone was responsible for 438,000 deaths across 95 countries in 2015[1]. Moreover, under even the most optimistic scenarios comprising existing interventions, the WHO Global Technical Strategy for Malaria[2] estimates that, although the incidence of malaria and corresponding death rate could be decreased by 90% by 2030, including complete elimination from 35 countries, the disease would continue to persist in 58 countries, presenting an indefinite burden. There appears to be broad consensus that additional interventions will be required for eradication of malaria to be achieved[3,4].

Genetically engineering wild populations could help mitigate a vector-borne disease via two basic approaches. In the first approach, a genetic construct could be spread that reduces the capacity of a vector to transmit the disease (population *alteration*). In the second approach, a construct could be spread that brings about a reduction in the size of the vector population outright (population *suppression*).

Genetic constructs that could bring about useful alteration or suppression of mosquitoes—if they were somehow spread through a population—have already been the subject of significant experimental investigation. In an alteration approach, so-called "cargo genes" have been identified that would reduce transmission of malaria[5–9] and dengue[10,11]. Although they have not yet been developed, cargo genes could potentially be constructed to reduce Zika transmission via an RNA interference approach or via an endonuclease targeting the Zika virus genome, which consists of single-stranded RNA[12]. Population suppression approaches have also been devised and tested in laboratory populations of mosquitoes, including a genetic construct that reduces female fertility in *Anopheles gambiae*, the most prominent malaria vector[13], as well as sex-ratio-distorting constructs, which are predicted to induce a population crash by reducing the relative number of females over successive generations if they were spread in the wild[14].

Another example of the potential for population genetic engineering to combat vector-borne disease is the "Mice Against Ticks" project[15], which recently began with the goal of eventually eradicating Lyme disease—the most common vector-borne disease in the United States[16]—from the islands of Nantucket and Martha's Vineyard. Briefly, white-footed mice (*Peromyscus leucopus*) serve as a natural reservoir for Lyme disease, which is transmitted between mice via ticks, which, in turn, bite humans to transmit the disease from its natural reservoir. To combat the disease, the Mice Against Ticks project would release mice carrying genes encoding anti-Lyme antibodies, which would immunize them against Lyme and could, correspondingly, reduce the burden of Lyme among humans.

Aside from control of disease vectors, another prominent potential application of population genetic engineering technology is control of destructive invasive species and agricultural pests. There are two categories of applications that are often discussed: modification or suppression of invasive species that (i) cause ecological harm, or (ii) cause economic harm, typically via destruction of agricultural crops. As a striking example of the first category, New Zealand announced in 2016 a goal of eliminating all of its rats, possums and stoats by 2050[17] because they are invasive to New Zealand and cause a tremendous amount of damage to the country's native ecosystems[18]. An example of the second category could include suppression or modification of the citrus psyllid, which, as a vector of *Candidatus* Liberibacter species, transmits citrus greening disease[12,19]. Another example of the second category—which has already been proposed and tested in the laboratory[20]—could include suppression or modification of *Drosophila suzukii*, a major pest of soft-skinned fruits (e.g., strawberries, raspberries, cherries, etc.), which is estimated to cause a total of $511 million in annual revenue losses across California, Oregon and Washington[21].

Given the tremendous potential upside if these applications could be realized, a great deal of research has been conducted to develop strategies for actually spreading genetic constructs through wild populations, dating back at least to the 1960s[22,23]. The strategies vary widely in mechanism but are unified by a common idea: genomes of individuals are engineered to encode both a desired trait and also some mechanism—broadly referred to as a *gene drive* mechanism—that induces evolution to favor the engineered construct (often called a *drive element*) as the individuals in the population reproduce over time, even if the construct is deleterious. In principle, only a comparatively small number of individuals would need to be engineered and released in order to alter an entire population.

A variety of mechanisms have been utilized to create these gene drive systems, including underdominance, maternal effect dominant embryonic arrest (*Medea*), and endonuclease-based approaches.

Underdominance-based mechanisms utilize bistability brought about by heterozygotes exhibiting lower fitness than homozygotes: when a population is mostly wild-type, the drive element goes to extinction, but when the drive element is released at high frequency, it goes to fixation. This creates a "threshold effect," whereby a large release of engineered organisms (above the threshold frequency, typically about 0.5) leads to spread of the drive element, whereas a smaller release leads to extinction of the drive element. Two drive systems of this type have been engineered in *Drosophila*[24,25], each using a toxin-antidote mechanism. In Ref. 24, two maternally expressed, unlinked, zygotic toxins are each linked with a zygotic antidote that rescues the lethality of the opposite toxin. Hence, individuals must inherit either both or neither to be viable. In Ref. 25, a single-locus construct is engineered, which includes both a gene that targets RNAi to a haploinsufficient gene, and an RNAi-insensitive rescue gene. The idea is that both wild-type and engineered homozygotes have two functional copies of the haploinsufficient gene, resulting in near-wild-type fitness, whereas heterozygotes have only one functional copy, resulting in lower fitness. The threshold effect has both an upside and a downside: it could help contain drive systems in populations with limited migration elsewhere, but it could also preclude use in large populations due to logistical difficulties. Moreover, underdominance-based approaches can only be utilized for population alteration, not suppression.

Medea systems also use a toxin-antidote approach but are predicted to exhibit much lower release thresholds. However, they can also only be utilized for population alteration. An engineered *Medea* element consists of two components: first, it encodes a toxin (typically a microRNA) that is expressed during oogenesis in females and disrupts an embryonic essential gene in every embryo, regardless of whether it inherits the *Medea* element or not. Second, it encodes a tightly linked antidote that is expressed only in zygotes that inherit the *Medea* element. The result of this mechanism is that wild-type/*Medea* heterozygotes preferentially pass on the *Medea* construct to offspring since only *Medea*-carrying zygotes are rescued from the effects of the maternally-expressed toxin. Modeling suggests that the threshold frequency for *Medea* spread approaches zero as the fitness cost of the construct (independent from the maternal-effect lethality) approaches zero[47].

To date, three proof-of-concept *Medea* elements have been engineered in *Drosophila*[20,26,27]. A downside of *Medea*-based systems is that they are difficult to construct in diverse, non-model organisms[12], although this difficulty might be overcome by novel designs that utilize alternative silencing approaches.

Endonuclease-based systems, in contrast to the other two approaches, exhibit no threshold behavior—and can, therefore, see application even in large populations—and can be used for both alteration and suppression across a diverse range of species. These systems, first proposed by Austin Burt in a seminal 2003 paper[28], use endonucleases to increase their chance of inheritance from heterozygous parents. This inheritance bias can be achieved in one of two ways: the endonuclease can copy itself onto a homologous chromosome, guaranteeing inheritance (because one of the two chromosomes must be inherited), or it can cleave the opposite allele in such a way that it is lethal if inherited (i.e., half of the offspring are nonviable, but all viable offspring inherit the endonuclease system).

Mechanistically, the first approach (often called *homing*) proceeds via a two-step process: (i) the endonuclease cuts the opposite chromosome at a sequence that is homologous to the region where the endonuclease is encoded, and (ii) template-based DNA repair via homologous recombination copies the engineered construct—including the endonuclease and any adjacent cargo genes—into the cut site, repairing the break by inserting the engineered construct. This approach could be used for alteration or suppression applications. In contrast, the second approach (*shredding*), which is almost exclusively considered for suppression applications, proceeds by cutting the opposite chromosome in many locations near the centromere, ensuring that an incomplete copy of the chromosome is passed on during meiosis, resulting in a nonviable offspring. This approach is typically discussed in the context of sex-ratio distorting systems, wherein the construct is encoded on the Y chromosome, and the X chromosome is "shredded," guaranteeing that all viable offspring inherit the Y chromosome and are, therefore, male. This effect is predicted to serve as an extremely effective population suppression strategy, eventually causing a population crash due to the increasingly biased sex ratio.

The construction of endonuclease-based gene drive systems was long hindered by a lack of easily programmable sequence-specific endonucleases. Proof-of-concept systems were originally constructed using homing endonuclease genes (HEGs)[29–32] that targeted artificial recognition sites, but as the recognition sites of HEGs are prescribed by protein structure, producing a drive element to target an arbitrary endogenous sequence in a new organism would present a difficult challenge in protein engineering.

The recent advent of CRISPR/Cas9 genome editing technology[33–36] has revolutionized gene drive engineering by allowing for the simple design and construction of endonuclease drive systems with arbitrary target sequences. Briefly, Cas9 is an endonuclease whose target is prescribed by a 20-base sequence in an independently-expressed guide RNA (gRNA). Thus, to engineer an endonuclease-based gene drive system, all that is now required is to genomically insert a DNA sequence encoding Cas9, as well as a DNA sequence encoding a gRNA with the target sequence of interest. To date, proof-of-concept CRISPR-based gene drive systems have been constructed in yeast[37], fruit flies[20,20,38], and mosquitoes[5,13], representing both population alteration and population suppression applications.

Although CRISPR gene drive systems are now being constructed at a rapid pace across a diverse range of species, there are still significant gaps in our theoretical understanding of their evolutionary dynamics. Essentially, what would happen if these constructs were released into the wild? Would they spread? If so, how far? What challenges would they face—i.e., what are the most likely failure modes? In this dissertation, I study a few of these questions for alteration-type CRISPR gene drive systems using mathematical modeling.

In Chapter 1, I study the evolution of alleles that are resistant to CRISPR gene drives and analyze a strategy that might help mitigate the effect of resistance. In this context, a *resistant allele* is any allele at the same locus as the CRISPR gene drive construct that is immune to its effects. These are typically variants of the wild-type allele with mutations at the target sequence of the CRISPR nuclease, and they can arise spontaneously or due to misrepair following CRISPR-mediated cutting—in addition, they are expected to exist in most populations simply due to standing genetic variation. There are, of course, many possible known and unknown mechanisms that could result in resistance to CRISPR gene drives in the wild, but in this chapter, I present a design that could potentially mitigate the effects of at least this particular form of resistance. This chapter was published in *Science Advances* (Ref. 39), and I was fortunate to be able to carry out the project with fantastic collaborators, including Jason Olejarz, with whom I worked closely on all of the mathematical models and calculations, as well as Kevin Esvelt, George Church and Martin Nowak, who provided excellent advising on all aspects of the project.

In Chapter 2, I address the question of how likely it would be for an existing, proof-of-principle CRISPR gene drive system to spread in a wild population following a very small release, accounting for resistance. The basic question I seek to address in this chapter is how *invasive* CRISPR gene drives might be—that is, how difficult it might be to contain an intervention to a population of interest or to protect nearby wild populations from laboratory escapes. The results in this chapter suggest that many existing CRISPR gene drive elements (even without optimization to mitigate resistance, as discussed in Chapter 1) could potentially spread to high frequencies in wild populations following very small releases. To provide empirical grounding for this work, I include as Appendix A a review of all CRISPR gene drive experiments reported in the literature to date, including a table of reported drive efficiencies (i.e., how often the cut/copy drive mechanism succeeds, a measure of how efficiently the construct biases its inheritance). This chapter was published in *eLife* (Ref. 40). For this work, I enjoyed a very fruitful collaboration with Ben Adlam on the mathematical models, as well as extremely insightful advising from George Church, Kevin Esvelt, and Martin Nowak.

Finally, in Chapter 3, I turn to the question of how a CRISPR gene drive system might be designed so that it is easier to contain in a particular population. While some prominent CRISPR gene drive applications have ambitions of altering or suppressing species across entire continents—e.g., malaria, dengue, Zika, schistosomiasis—many potential applications are much more localized in nature, either because of ecological or policy considerations. Appendix B contains supplementary figures related to this work. This chapter is currently in review, and I have greatly enjoyed working together with a variety of experimental and mathematical modeling collaborators in a highly collaborative and interdisciplinary project. I worked closely with Jason Olejarz on the mathematical models; John Min, Joanna Buchthal and Alejandro Chavez designed and performed experiments to assemble a collection of CRISPR guide RNAs that were required for the proposed approach to be feasible in reality; Erika DeBenedictis wrote a helpful web-based user interface for visualizing the results of the model; Andrea Smidler helped build a preliminary discrete generation precursor to our model; Kevin Esvelt conceived the project, and he, George Church, and Martin Nowak continued providing extremely helpful advice and unwavering support.

1

Evolutionary dynamics of drive resistance

1.1 Foreword

In this chapter, I explore the evolutionary dynamics of CRISPR gene drive systems in the face of a particular form of resistance—that which is genetically encoded at the target site of the drive construct and blocks recognition by CRISPR guide RNA(s). I study two different designs: one that is typical of existing proof-of-principle gene drive constructs, and a second that was previously proposed as a means of combating resistance and enhancing the long-term stability of the drive in a population.

I performed this work together with Jason Olejarz, who contributed great help with developing and analyzing the mathematical models presented here. We benefited greatly from insight, advising and support from Kevin Esvelt, George Church and Martin Nowak.

This chapter was first published in Ref. 39:

Charleston Noble*, Jason Olejarz*, Kevin M. Esvelt, George M. Church and Martin A. Nowak. Evolutionary dynamics of CRISPR gene drives. *Science Advances* 3, e1601964 (2017). (*equal contribution)

1.2 Introduction

Gene drive systems are selfish genetic elements which bias their own inheritance and spread through populations in a super-Mendelian fashion (Fig. 1.1A). Such elements have been discussed as a means of contributing to the eradication of insect-borne diseases, such as malaria, reversing herbicide and pesticide resistance in agriculture, and controlling destructive invasive species[5,13,24,28,29,37,38,41–45]. Various examples of gene drive can be found in nature, including transposons[46], Medea elements[26,47], and segregation distorters[48–51], but for ecological engineering purposes, endonuclease gene drive systems received the most significant attention in the literature[5,13,28–30,37,38,41–44,52,53]. In general, these elements function by converting drive heterozygotes into drive homozygotes through a two-step process: (i) the drive construct, encoding a sequence-specific endonuclease, induces a double-strand break (DSB) at its own position on a homologous chromosome, and (ii) subsequent DSB repair by homologous recombination (HR) copies the drive into the break site. Any sequence adjacent to the endonuclease will be copied as well; if a gene is present, we refer to it as "cargo", as it is "driven" by the endonuclease through the population.

Although originally proposed over a decade ago[28], the chief technical difficulty of this approach—inducing easily programmable cutting at arbitrary target sites—has only recently been overcome by the discovery and development of the CRISPR/Cas9 genome editing system[33–36,54]. Briefly, Cas9 is an endonuclease whose target site is prescribed by an independently expressed guide RNA (gRNA) via a 20-nucleotide protospacer sequence. Because virtually any position in a genome can be uniquely targeted by Cas9, so-called RNA-guided gene drive elements can be constructed by inserting a suitable sequence encoding both Cas9 and gRNA(s).

Recent studies have demonstrated highly functional CRISPR gene drive elements in mosquitoes[5,13], yeast[37], and fruitflies[38]. In each case, the basic construct consists of a copy of Cas9 with a single corresponding gRNA and cargo sequence (Fig. 1.1B). Despite drive inheritance of about 95%, on average, in the published studies (compared to 50% expected by Mendelian inheritance), the evolutionary stability of these constructs in large populations has been debated due to the potential emergence of drive resistance within a population[28,41,53]. A resistant allele is anticipated to arise whenever the cell repairs the drive-induced DSB using non-homologous end joining (NHEJ) instead of HR, a process that typically introduces a small insertion or deletion mutation at the target sequence. Because the reported constructs cut only at a single site, a substantial fraction of NHEJ events will create drive-

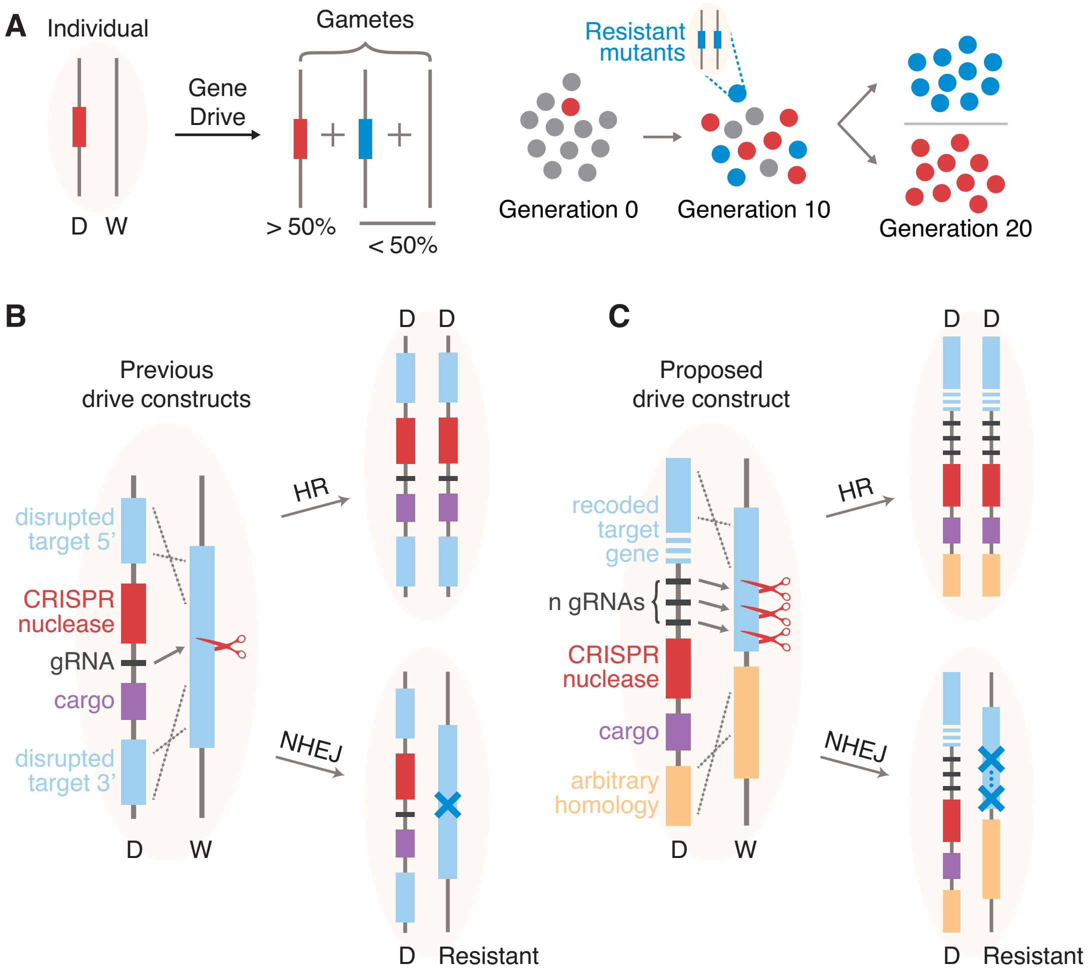


Figure 1.1: CRISPR gene drive inheritance and spread in wild populations. (A) Inheritance and spread of a gene drive construct, D, in a population of individuals homozygous for the wild-type, W. In the late germline, the drive construct induces a DSB at its own position on the homologous chromosome which is repaired either by HR, converting the individual to a DD homozygote, or by NHEJ, producing a small insertion/deletion/substitution mutation at the cut site which results in a drive-resistant allele. There is also the possibility of no modification, in which case the W allele remains unchanged. This mechanism can lead to rapid spread of the gene drive in a population or the spread of resistant alleles, depending on their relative fitness effects. (B) To achieve this mechanism, previously demonstrated drive constructs are inserted at some target sequence (blue) and carry a CRISPR nuclease (for example, Cas9) with a gRNA, as well as a "cargo gene" which can be chosen arbitrarily for the desired application. Disruption of the target sequence must be nearly neutral for the drive to spread. (C) The construct we model here, which was proposed by Esvelt *et al.*[41], reconstitutes the target gene after cutting—so an essential gene can be chosen as the target to select against resistant alleles—and employs multiple (n) gRNAs.

resistant alleles that could prevent the construct from spreading to the entire population (Fig. 1.1B).

Drive resistance was first mathematically studied in the context of single-cutting homing endonuclease-based drive elements[53]. There, it was concluded that drive is most effective when the fitness cost of the drive is low and the fitness cost of resistance is high (see Section 1.6.1 for a description of that work). Unfortunately, in the drive constructs reported thus far, these two requirements are fundamentally at odds: the fitness cost of resistance arises from disruption of the target sequence, but the drive copies itself precisely by disrupting the target sequence.

Here we study the evolutionary dynamics of an alternative drive architecture that decouples these effects by rescuing function of the target gene, but only if the drive cassette is successfully copied. This design was first proposed conceptually by Esvelt *et al.*[41] but has not yet been modeled or constructed in the laboratory; hence, we refer to it here as the "proposed" construct. It involves targeting multiple sites within the 3' end of a gene for cutting by the drive and including a completely genetically recoded[55–57] copy of this 3' target sequence in the drive construct (Fig. 1.1C). The 3' untranslated region of the gene is also replaced with an equivalent sequence in order to remove all homology between the cut sites and the drive components, which ensures that the drive cassette is copied as a single unit. If repair occurs by HR, then the target gene is restored to functionality as the drive is copied. However, if repair occurs by NHEJ, then the target gene is mutated, potentially resulting in a knockout and a corresponding loss of fitness. Using this design, drive resistance can be selected against by choosing an essential or even haploinsufficient gene as the drive target.

Because the success of this design is contingent on the ability to genetically recode the 3' end of an essential gene without imposing a large fitness cost, we now briefly discuss the plausibility of this strategy. In a study of CRISPR-based gene drive in yeast, DiCarlo *et al.*[37] showed that a drive construct targeting the essential *ABD1* gene and encoding a recoded copy of *ABD1* functioned with high efficiency without exhibiting "any obvious fitness defects as compared to wild-type strains". In the most comprehensive study of essential gene recoding to date, Ostrov *et al.*[57] showed that computationally minimizing disruption of existing RNA-binding motifs and secondary structures while preserving overall codon usage allowed the elimination of seven codons from 91% of essential genes in *Escherichia coli* with an overall fitness cost of less than 10%. Moreover, many attempted recodings were costless on the first try without requiring optimization. Wang *et al.*[58] obtained similar results. Finally, work in *Drosophila* on underdominance-based drive systems[24,27] has shown that partial recoding of haploinsufficient genes in metazoans is possible, although in both studies this involved RNA interference.

In addition to 3' target recoding, the construct uses multiple gRNAs. The use of multiple gRNAs offers two important benefits with respect to resistance: (i) all gRNA target sites must be mutated or lost before a single allele becomes drive-resistant, and (ii) if cutting occurs at two or more gRNA target sites simultaneously, then the intervening DNA sequence is lost, resulting in a large deletion and a knockout of the target gene. This is in contrast to single-cutting constructs, where a knockout can be avoided by an in-frame indel or substitution mutation.

1.3 Results

To study this construct, we formulate a deterministic model (Sections 1.5, 1.6.2 and 1.6.3) that considers the evolution of a large population of diploid organisms and focuses on a specific locus with $2n + 2$ alleles (Fig. 1.2A). First, there are the wild-type (W) allele and the gene drive allele with n gRNAs (D). There are then n distinct "cost-free" resistant alleles that are resistant to drive-induced cutting at $1, 2, \ldots, n$ target sites but are otherwise identical to the wild-type (denoted $S_1, S_2, \ldots, S_n$). These could arise via, for example, mutations that block cutting by disrupting the gRNA target sequences but do not cause a shift in the reading frame. Finally, there are n distinct "costly" resistant alleles, which have fitness effects that are distinct from those of the wild-type (denoted $R_1, R_2, \ldots, R_n$). Only the alleles S_n and R_n are fully resistant to cutting by the drive. We also refer to the wild-type allele as S_0 for notational convenience. Last, we say that individuals having genotype AB, where A and B are any of the alleles above, have fitness f_{AB} (alternatively, genotype AB is associated with a cost $1 - f_{AB}$) and produce gametes having haplotype C with probability $p_{AB,C}$. Note that these probabilities $p_{AB,C}$ abstract all individual-level drive dynamics and are agnostic to the mechanism that produces drive. We allow these parameters to be arbitrary for our analytical calculations and derive corresponding results that hold for any underlying drive mechanism—including both the previous drive constructs and the new ones considered here.

For numerical simulations, we further consider a mechanistic model that explicitly describes the mechanism of drive in individuals (Fig. 1.2B, Section 1.6.4). We assume that, in the germ line of an individual that is heterozygous for a drive construct and a susceptible allele (DS$_i$ where $0 \leq i < n$

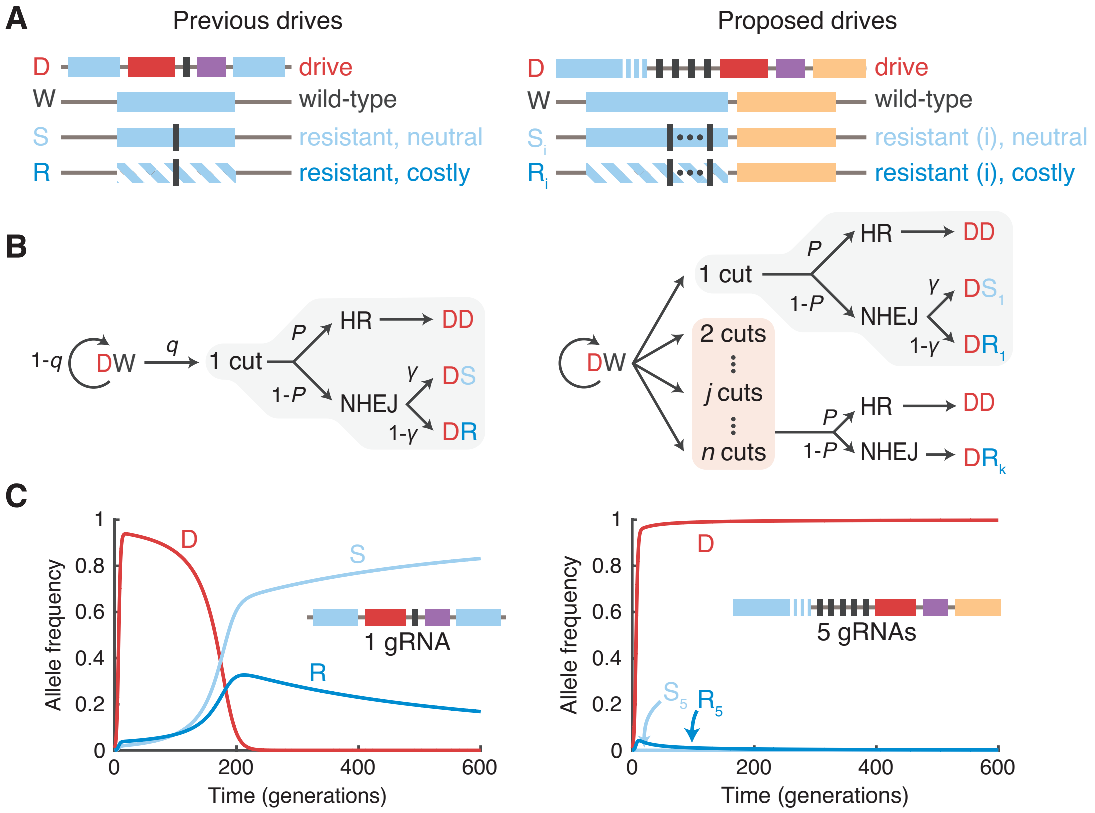


Figure 1.2: Modeling framework and representative simulations. (A) We consider $2n + 2$ alleles, where n is the number of drive target sites (prescribed by CRISPR gRNAs): the drive construct (D), the wild-type (W), n "neutral" resistant alleles (S$_i$), and n "costly" resistant alleles (R$_i$). Previous drives (left) used one target site, whereas our proposed drives use multiple target sites (right). (B) Conversion dynamics within DW germline cells during early gametogenesis. Cutting occurs at each susceptible target independently with probability q. Then, repair occurs by HR with probability P or by NHEJ with probability $1 - P$. In the case of a single cut (light gray), if there is NHEJ repair, then repair produces a functional target gene with probability γ or a non-functional target with probability $1 - \gamma$. Two or more cuts (light red) certainly produce non-functional targets after NHEJ repair. (C) Representative simulations are shown using high cutting and HR probabilities ($q = P = 0.95$), for an initial drive release of 1% in a wild-type population, with $\gamma = 1/3$. Fitness parameters are (left) $f_{SS} = f_{SR} = 1, f_{SD} = 95\%, f_{RR} = 99\%,$ $f_{DD} = f_{DR} = (99\% \times 95\%) = 94.1\%$, where S refers to neutral alleles (either S or W), and (right) $f_{SS} = f_{SR} = 1, f_{SD} = f_{DD} = f_{DR} = 95\%; f_{RR} = 1\%$, where S and R refer to alleles W, S$_1, \ldots,$ S$_5$ and R$_1, \ldots,$ R$_5$, respectively. See Section 1.6.4.2 for details regarding our assignments of the inheritance probabilities.

or DR$_i$ where $1 \leq i < n$), each susceptible target site undergoes cutting independently with probability q. If there is at least one cut, then HR occurs with probability P, whereas NHEJ occurs with probability $1 - P$. If HR occurs, then the cell is converted to a drive homozygote. However, if mutagenic NHEJ occurs, then there are a few possibilities, depending on the number of cuts.

If there is exactly one cut, then one gRNA target is lost on the susceptible allele. If the susceptible allele was initially functional (S$_i$), then with probability γ it retains function and converts to S$_{i+1}$; otherwise, it loses function and converts to R$_{i+1}$. We assume that the parameter γ is the probability that the reading frame is unaffected, so $\gamma = 1/3$. If the susceptible allele is initially nonfunctional (R$_i$) then we assume that it cannot regain function, so it converts to R$_{i+1}$.

If there are two or more cuts, then all j susceptible gRNA targets between and including the outermost damaged targets in the locus are lost ($2 \leq j \leq n - i$). The resulting allele is certainly nonfunctional and thus converts to R$_{i+j}$. The probability distribution for the number of lost targets is described in Section 1.6.4.2. It follows directly from our assumptions that cutting at each target site is independent and that sequential cutting and repair events do not occur.

Regarding initial conditions, our simulations and analytical invasion analysis assume that drive-homozygotes (genotype DD) are released into a population consisting initially of fully susceptible wild-type homozygotes (genotype S$_0$S$_0$). However, depending on the sequence targeted by the drive, standing genetic variation in real populations could result in preexisting resistance at one or more gRNA targets. For example, in a genome-wide analysis of 192 inbred strains of *Drosophila melanogaster* derived from a single natural population, MacKay et al.[59] found the genome-wide averaged polymorphism value[60] to be $\pi = 0.0056$. If we assume that polymorphism at each base pair is independent, then the number of mismatches at a gRNA target sequence in a particular individual is binomial with 20 trials and success probability π. And if each gRNA can tolerate, on average, one mismatch in its target, then single guide-resistant alleles should exist at frequencies roughly on the order of 10^{-3}. Further assuming that resistance at each gRNA is independent, two guide-resistant alleles should exist at frequencies roughly on the order of 10^{-5}, and so on. In this example and with these assumptions, using five guide RNAs would reduce the frequency of preexisting fully-resistant alleles to 10^{-12}. Of course, complications could arise, such as nonindependence of polymorphism within or between guides, so we anticipate this to serve as a low estimate of the frequency of preexisting resistance in a natural population. Therefore, before any application is considered, standing variation in the target population should be carefully measured, and the target gene as well as the number of guides should be adjusted accordingly.

Now, we address two fundamental questions: whether a CRISPR gene drive will invade a resident wild-type population and, if so, whether it will be evolutionarily stable[61]. We begin with the former. We find that a CRISPR gene drive will invade a wild population if

$$2p_{WD,D}f_{WD} > f_{WW} \tag{1.1}$$

A derivation of this result can be found in Sections 1.6.2.2 and 1.6.3.2. For the drive to spread when initially rare, the advantage from inheritance biasing ($p_{WD,D}$)—typically about 95% in published studies—must overcome the lower fitness of the drive/wild-type heterozygote (f_{WD}) compared with the wild-type (f_{WW}). Note that this condition holds in the context of drive resistance, is agnostic to individual-level drive dynamics, and thus applies both to previous drive architectures and our proposed architecture. Equation (1.1) explains the apparent success of CRISPR drive constructs reported in the literature[5,13,37,38], which easily invade wild-type laboratory populations, or would be predicted to do so after optimization of drive expression: Over short time scales, drive resistance is rare and thus does not affect the dynamics.

However, over longer time scales, NHEJ-mediated resistance will markedly affect the dynamics. We find that a resident drive population is stable against invasion by resistant alleles if and only if

$$\max_{A \in S \cup R} (2p_{DA,A} f_{DA}) < f_{DD} \tag{1.2}$$

Here, the maximization is over all nondrive alleles $\mathrm{S}_0, \ldots, \mathrm{S}_n$ and $\mathrm{R}_1, \ldots, \mathrm{R}_n$. Intuitively, the drive is stable if and only if no other allele can invade, and each of these has an invasion condition identical in form to Eq. (1.1). (A derivation of this result can be found in Sections 1.6.2.3 and 1.6.3.3).

Disconcertingly, Eq. (1.2) suggests that drive constructs are necessarily unstable in sufficiently large populations. An individual who is heterozygous for the drive and the fully resistant cost-free allele S_n has probability $p_{DS_n,S_n} = 1/2$ of producing an S_n gamete, and this individual has fitness equivalent to (or potentially greater than) the drive/wild-type heterozygote. Thus, if the drive construct has lower fitness than the wild-type, and if the fully resistant cost-free allele has a nonzero rate of production in the population, then the latter will certainly invade a resident drive population. This is especially problematic for highly deleterious population suppression drives, as in the study by Hammond *et al.*[13], which have low fitness relative to the wild-type and less costly resistant alleles.

However, population alteration drives (sometimes referred to as replacement drives) might not require long-term persistence in a population to produce their desired effect. Some applications might still be successful as long as the drive construct attains and persists at a sufficiently high frequency in the population over some length of time.

To quantify the relative effectiveness of the previous and proposed drive architectures, we consider three quantities: (i) the maximum frequency achieved by a drive construct released in a wild population, (ii) the time required for a drive construct to attain 90% of its maximum frequency, and (iii) the frequency of the drive construct after 200 generations, roughly the longest relevant timescale for a typical application. We compute these quantities numerically for drives featuring cutting and HR probabilities consistent with average drive inheritance rates observed in previous fruitfly[38] and mosquito[5,13] experiments ($q = P = 0.95$, modeling a reported drive inheritance rate of roughly 95% from DW individuals).

Our results suggest that, as anticipated from Eq. (1.1), both the previous and proposed drive constructs should spread similarly in the short term, immediately following release (Fig. 1.3, A, B, and D). However, over longer time scales, the two constructs undergo markedly different dynamics. The proposed drive constructs, released at an initial frequency of 1% in a wild population, using five gRNAs and targeting an essential gene, can attain $> 99\%$ frequency in a population (Fig. 1.3, B and C) in 10 to 20 generations (Fig. 1.3, B and D) and remain above 99% for at least 200 generations (Fig. 1.3, B and E). Furthermore, this is seen over a large range of drive fitness costs, up to approximately 30% (Fig. 1.3, C to E). In contrast, the previously demonstrated constructs attain maximum frequencies between 90% and 95% over a narrower range of fitness values (Fig. 1.3, A and C) and demonstrate significantly reduced stability (Fig. 1.3E). In particular, previous constructs exceeding 8% fitness cost invariably fall below their initial release frequency in fewer than 200 generations.

1.4 Discussion

In summary, we constructed and analyzed a mathematical model of CRISPR gene drive that includes multiplex cutting via multiple guide RNAs and allows for multiple costly and cost-free resistant alleles. Our results suggest that previously demonstrated CRISPR gene drives constructed as proofs of principle should effectively invade wild populations—consistent with experimental observations—but could have limited utility due to their inherent instability, brought about by their production of resistant alleles and vulnerability to preexisting ones. We studied an alternative drive architecture, first proposed by Esvelt *et al.*[41], which contains (i) multiple CRISPR guide RNAs which target the 3' end of a gene, and (ii) a recoded copy of the target gene which is functional but resistant to cutting. We discussed the plausibility of building such a construct in light of recent experimental reports, and we concluded that this architecture could substantially improve the stability of CRISPR gene drives by minimizing the effects of NHEJ-mediated resistance.

Another alternative strategy which we have not modeled here would involve multiple independent single-guide drive constructs targeting the same locus. This is conceptually symmetric to the strategy considered here: Rather than a single drive with multiple (n) gRNAs ("multiple guides"), one might consider multiple (n) drives with one gRNA each ("multiple drives"). In this strategy, each independent drive would behave similarly to the previously demonstrated constructs. The

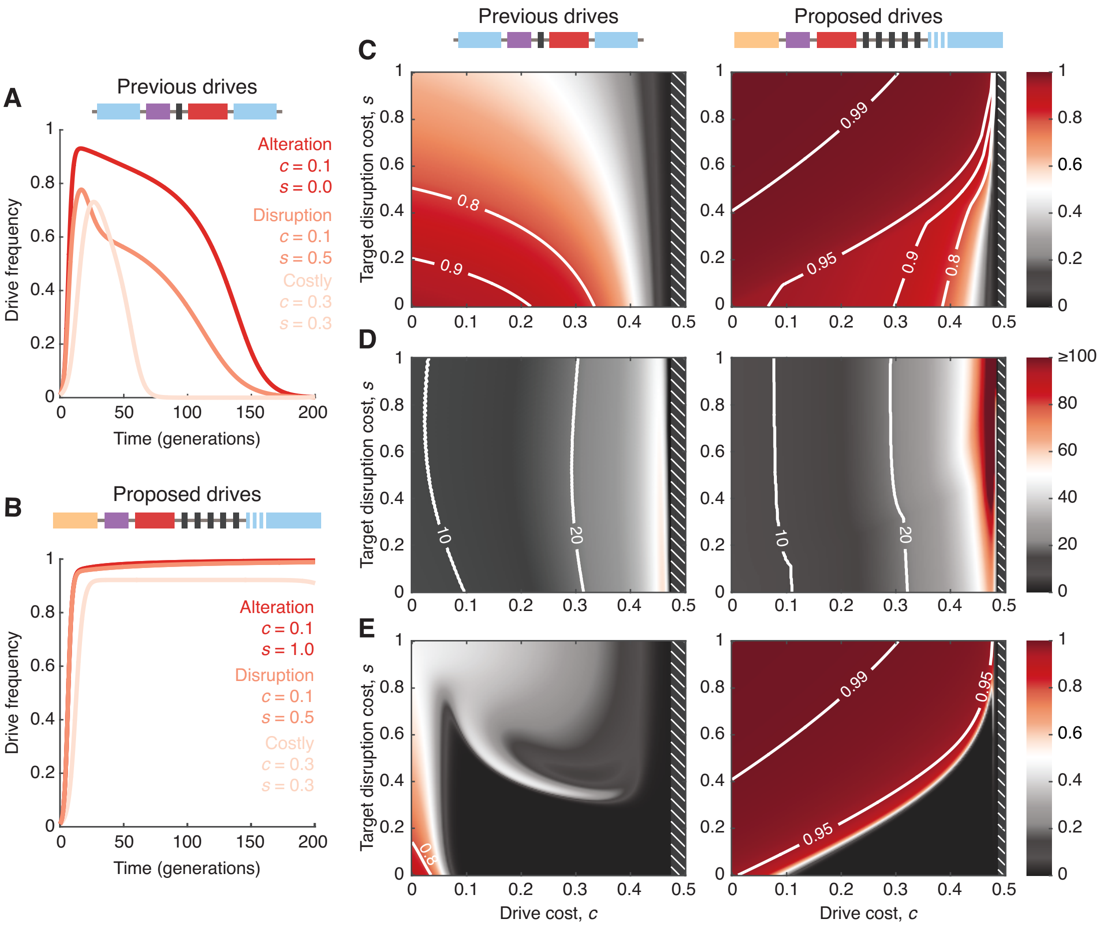


Figure 1.3: Quantitative comparison of previously demonstrated and recently proposed drive constructs. (A and B) Drive frequency over time for three particular scenarios: a low-cost alteration drive carrying a cargo gene and targeting a neutral site (previous drives) or an essential gene (proposed drives) (red), a low-cost drive whose aim is to disrupt an important target gene (orange), and a high-cost drive (tan). (C) The maximum drive allele frequency (heat) observed in simulations across 200 generations, following an initial release of drive-homozygous organisms comprising 1% of the total population. In white hatched regions, Eq. (1.1) is not satisfied, so no invasion occurs. (D) Generations to 90% of the maximum frequency. (E) Frequency of the drive constructs after 200 generations, a measure of stability in the population. Parameters used are as follows: (throughout) $q = P = 0.95, \gamma = 1/3$; (previous drives) $n = 1$, $f_{SS} = f_{SR} = 1, f_{SD} = 1 - c, f_{DD} = f_{DR} = (1 - c)(1 - s), f_{RR} = 1 - s$; (proposed drives) $n = 5$, $f_{SS} = f_{SR} = 1, f_{SD} = f_{DD} = f_{DR} = 1 - c, f_{RR} = 1 - s$, where S and R refer to any alleles S$_0, \ldots,$ S$_n$ and R$_1, \ldots,$ R$_n$, respectively. Inheritance probabilities are assigned as illustrated in Fig. 1.2B and described in Section 1.6.4.2.

multiple-drive strategy would likely outperform the previous strategy, but we anticipate that it would not outperform the multiple-guide strategy. This is because, in the multiple-drive strategy, each gRNA target can independently undergo NHEJ-mediated mutation, providing stepping-stones to fully resistant alleles. Furthermore, the multiple-drive strategy lacks the benefit of large NHEJ knockouts from multiple simultaneous cuts, which help combat cost-free resistance (Fig. 1.2B, red box), although it would be capable of editing regions unimportant to fitness. And, regardless, each single-guide drive construct could itself be built in the way we have described here, by using multiple gRNAs.

An important caveat of our work is that we specifically studied resistance that is genetically encoded at the drive locus and is generated by the action of the drive. Many other mechanisms of resistance are certainly possible. For example, standing genetic variation and de novo mutation might be important considerations, particularly if the target locus is not highly conserved. However, in recent work [62], Unckless *et al.* showed that NHEJ-mediated resistance should be more impactful for realistic NHEJ rates (specifically, greater than the inverse of the population size). Aside from these mechanisms of within-locus resistance, resistance could also arise in trans, for example as heightened ribonuclease activity or as the evolution of small RNAs which would lead to knockdowns via RNA interference. In addition, even beyond direct molecular effects, resistance could arise via higher-level effects, for example as selection for inbreeding behavior in hermaphrodites in response to extremely costly population suppression drives, as recently studied by Bull[63]. The large variety of potential resistance mechanisms underscores the need for further theoretical and experimental work on this topic.

Although our work has focused on how to maximize the invasibility and stability of gene drive systems, "global" CRISPR gene drives, such as those considered here, should only be actively developed for severe problems that (i) cause a great deal of suffering, and (ii) have few other potentially viable solutions. Examples include malaria and schistosomiasis. Other applications—such as precision alterations to local populations—will require robust methods to ensure limited spatial and/or temporal spread. Toward this aim, there are several existing approaches, including non-drive strategies such as multi-locus assortment[64] and threshold-dependent drives ()like toxin-based underdominance systems)[24,26]. Moreover, we, among others, recently proposed an alternative theoretical approach termed "daisy drive"[65].

In conclusion, our results suggest three concrete design principles for future CRISPR gene drive systems. Constructs will minimize the impact of misrepair and thus maximize evolutionary stability if (i) multiple gRNAs with minimal off-target effects are used, (ii) disruption of the target locus is highly deleterious, and (iii) any cargo genes are as close to neutral as possible.

1.5 Brief model description

Here, we briefly state the model used for the numerical simulations presented above. The remainder of this Chaper is largely used to develop and explain this model (beginning in Section 1.6.2 and later extended to include neutral resistance in Section 1.6.3).

Throughout this work, we study a genetics-based evolutionary dynamics model. We consider the evolution of diploid individuals, x_{IJ} where $I, J = \mathrm{W}, \mathrm{D}, \mathrm{R}_1, \mathrm{R}_2, \ldots, \mathrm{R}_n, \mathrm{S}_1, \mathrm{S}_2, \ldots, \mathrm{S}_n$. Here

D corresponds to the drive with n gRNAs; $\mathrm{R}_1, \mathrm{R}_2, \ldots, \mathrm{R}_n$ correspond to alleles that are resistant to cutting at $1, 2, \ldots, n$ target sites, respectively, and $\mathrm{S}_1, \ldots, \mathrm{S}_n$ are resistant alleles with no fitness cost, and W corresponds to the wild-type (which we also denote by S_0 for notational convenience). In Section 1.6.2 (extended to neutral resistance in Section 1.6.3), we present a continuous-time model for the evolutionary dynamics of this population, as well as derivations for the invasion and stability conditions discussed above. Here, we briefly describe this model. First, it makes the following assumptions: (i) an infinitely large population; (ii) random mating; (iii) standard segregation of allele pairs at meiosis, unless an individual has genotype DA (where A is one of $\mathrm{S}_0, \ldots, \mathrm{S}_{n-1}$ or $\mathrm{R}_1, \ldots, \mathrm{R}_{n-1}$) , in which case gametes receive a D allele with probability $p_{DA,D}$ or an A allele with probability $p_{DA,A}$; and (iv) viability selection where each genotype IJ has fitness f_{IJ}.

Using these rules, we can formally express the rates at which each of the $2n + 2$ types of gametes is produced in terms of the frequencies of individuals in the population. We denote by $F_D(t)$ the rate (at time t) at which drive gametes (D) are produced by individuals in the population. We denote by $F_{S_i}(t)$ the rate (at time t) at which wild-type gametes ($i = 0$) or gametes with varying levels of cost-free resistance ($1 \leq i \leq n$) are produced by individuals in the population. Last, we denote by $F_{R_i}(t)$ the rate (at time t) at which gametes with varying levels of costly resistance ($1 \leq i \leq n$) are produced by individuals in the population. We have

$$
\begin{aligned}
F_D(t) &= f_{DD}x_{DD}(t) + \sum_{k=1}^{n} p_{R_kD,D} f_{R_kD} x_{R_kD}(t) + \sum_{k=0}^{n} p_{S_kD,D} f_{S_kD} x_{S_kD}(t) \\
F_{S_i}(t) &= \sum_{k=0}^{n} \frac{1+\delta_{ki}}{2} f_{S_kS_i} x_{S_kS_i}(t) + \frac{1}{2}\sum_{k=1}^{n} f_{R_kS_i} x_{R_kS_i}(t) + \sum_{k=0}^{i} p_{S_kD,S_i} f_{S_kD} x_{S_kD}(t) \\
F_{R_i}(t) &= \sum_{k=1}^{n} \frac{1+\delta_{ki}}{2} f_{R_kR_i} x_{R_kR_i}(t) + \frac{1}{2}\sum_{k=0}^{n} f_{R_iS_k} x_{R_iS_k}(t) \\
&\quad + \sum_{k=1}^{i} p_{R_kD,R_i} f_{R_kD} x_{R_kD}(t) + \sum_{k=0}^{i-1} p_{S_kD,R_i} f_{S_kD} x_{S_kD}(t)
\end{aligned}
$$

where δ_{ki} is the Kronecker δ. $x_{IJ}(t)$ denotes the frequency of individuals (at time t) with genotype IJ, where $I, J = \mathrm{D}, \mathrm{S}_0, \mathrm{S}_1, \ldots, \mathrm{S}_n, \mathrm{R}_1, \ldots, \mathrm{R}_n$. Similarly, f_{IJ} is the fitness of IJ individuals, and $p_{IJ,K}$ denotes the probability of an individual with genotype IJ producing a K gamete. From conservation of probability, we have the following identities:

$$
\begin{aligned}
p_{R_kD,D} + \sum_{i=k}^{n} p_{R_kD,R_i} &= 1 \\
p_{S_kD,D} + \sum_{i=k}^{n} p_{S_kD,S_i} + \sum_{i=k+1}^{n} p_{S_kD,R_i} &= 1
\end{aligned}
$$

Notice that type R$_n$D and type S$_n$D individuals are fully resistant to being manipulated by the drive construct; such a fully resistant individual shows standard Mendelian segregation in its production of gametes. Thus, we have $p_{R_nD,R_n} = p_{S_nD,S_n} = 1/2$.

The selection dynamics are modeled by the following system of equations

$$
\begin{aligned}
\dot{x}_{DD}(t) &= F_D^2(t) - \psi^2(t)x_{DD}(t) \\
\dot{x}_{R_iD}(t) &= 2F_{R_i}(t)F_D(t) - \psi^2(t)x_{R_iD}(t) \\
\dot{x}_{S_iD}(t) &= 2F_{S_i}(t)F_D(t) - \psi^2(t)x_{S_iD}(t) \\
\dot{x}_{R_iS_j}(t) &= 2F_{R_i}(t)F_{S_j}(t) - \psi^2(t)x_{R_iS_j}(t) \\
\dot{x}_{R_iR_j}(t) &= (2-\delta_{ij})F_{R_i}(t)F_{R_j}(t) - \psi^2(t)x_{R_iR_j}(t) \\
\dot{x}_{S_iS_j}(t) &= (2-\delta_{ij})F_{S_i}(t)F_{S_j}(t) - \psi^2(t)x_{S_iS_j}(t).
\end{aligned}
$$

The quantity $\psi^2(t)$ represents a density-dependent death rate for the individuals in the population. At any given time, t, we require that the total number of individuals sums to one

$$x_{DD}(t)+\sum_{i=1}^{n} x_{R_iD}(t)+\sum_{i=0}^{n} x_{S_iD}(t)+\sum_{i=1}^{n}\sum_{j=0}^{n} x_{R_iS_j}(t)+\sum_{i=1}^{n}\sum_{j=1}^{i} x_{R_iR_j}(t)+\sum_{i=0}^{n}\sum_{j=0}^{i} x_{S_iS_j}(t) = 1$$

To enforce this density constraint, we set

$$\psi(t) = F_D(t) + \sum_{i=1}^{n} F_{R_i}(t) + \sum_{i=0}^{n} F_{S_i}(t)$$

For further details of the model, as well as derivations of our invasion and stability conditions, please see Sections 1.6.2 and 1.6.3.

1.6 Supplementary model details and derivations

In the remainder of this Chapter, we briefly review a closely-related previous study, develop the mathematical model described in Section 1.5 and present derivations of Equations (1.1) and (1.2). We begin with our discussion of a previous study of homing endonuclease-based gene drive systems in Section 1.6.1. In Section 1.6.2, we propose a simple model of population genetics of CRISPR-based gene drive systems with multiple guide RNAs, and we analyze the selection pressure acting on an engineered drive construct. In Section 1.6.2.2, we derive a condition for an engineered drive allele to invade a natural population. In Section 1.6.2.3, we derive a condition for a population in which the drive has fixed to resist invasion by either wild-type or drive-resistant alleles. In Section 1.6.2.4, we derive equations for interior equilibria permitted by our system. In Section 1.6.2.5, we present numerical examples of the system's dynamics. Lastly, in Section 1.6.3, we extend the model from Section 1.6.2 to include the effects of "neutral resistance", leading to the model presented in Section 1.5 and used in numerical simulations throughout Section 1.3.

1.6.1 Previous work on homing endonuclease gene drives

At the time this Chapter was written, the most closely-related existing theoretical study of nuclease-based gene drive with resistance was presented by Deredec *et al.*[53]. In this Section, we briefly review that study in order to highlight the parallels and points of difference between our theoretical approaches and, more importantly, the underlying biological systems.

In the study by Deredec *et al.*, the authors mathematically investigate gene drive systems that utilize homing endonuclease genes (HEGs). Essentially, HEGs encode proteins that have both nuclease and DNA-targeting activity. Thus, an HEG-based gene drive can be thought of conceptually as an example of the "previous" constructs described in Fig. 1.1, if the CRISPR nuclease and gRNA were fused into one contiguous unit. This leads to an important difference between HEG and CRISPR-based gene drive systems: There is no analogue to "multiple guides" for HEG-based systems—each drive system has exactly one target site.

The authors begin their analysis of HEG-based gene drive systems with a two-allele model precluding resistance, consisting of a wild-type allele and a gene drive allele (pp. 2014–2016 of Deredec *et al.*[53]). As described previously, the model can be thought of as implicitly considering a single guide RNA because it was motivated by HEGs. In their notation, p is the frequency of the wild-type allele, and q is the frequency of the drive allele. The authors assume Hardy-Weinberg proportions at all times, and they write a recurrence for q:

$$q' = \frac{(1-s)q^2 + (1-sh)pq(1+e)}{1 - sq^2 - 2shpq}$$

Here, s is the fitness cost associated with a drive homozygote, sh is the fitness cost associated with a drive/wild-type heterozygote, and e is the probability that the HEG copies itself onto the homologous chromosome ("homes").

The authors identify that there are three possible fixed points:

$$q^* = 0$$

$$q^* = 1$$

$$q^* = \frac{e - (1+e)hs}{s(1-2h)}$$

The authors obtain the following invasion condition for the drive allele:

$$s < \frac{e}{h(1+e)}$$

Intuitively, the fitness cost, sh, of a drive/wild-type heterozygote must be less than a monotonically increasing function of the homing rate, e, for the homing endonuclease gene to spread when rare. Low fitness costs of the drive and high homing rates facilitate the invasion of the drive. More specifically, the authors show that, if the drive/wild-type heterozygote has fitness close to the wild-type (i.e., h close to zero), then the drive invades and fixes (if s is small relative to e), coexists with the wild-type allele (if s is comparable in magnitude to e), or does not invade and is unstable (if s is large relative to e). The authors also show that, if the drive/wild-type heterozygote has fitness close to the drive homozygote (i.e., h close to one), then the drive invades and fixes (if s is small relative to e), is bistable with the wild-type allele (if s is comparable in magnitude to e), or does not invade and is unstable (if s is large relative to e). These are important insights into the evolutionary dynamics of HEG-based gene drive systems.

Deredec *et al.* then extend their model to consider also a single resistant allele (pp. 2018–2019 of

Deredec et al.[53]). In their notation, p is the frequency of the wild-type allele, q_H is the frequency of the drive allele, and q_M is the frequency of the misrepaired (resistant) allele. The authors assume Hardy-Weinberg proportions at all times, and they write recurrences for q_H and q_M:

$$\begin{aligned}
q'_H &= \frac{q_H^2(1-s_H) + pq_H(1+e(1-\gamma))(1-h_Hs_H) + q_Mq_H(1-s_I)}{\overline{W}} \\
q'_M &= \frac{q_M^2(1-s_M) + pq_M(1-h_Ms_M) + pq_H(1-h_Hs_H)e\gamma + q_Mq_H(1-s_I)}{\overline{W}}
\end{aligned}$$

Here, $\overline{W}$ is the mean fitness of the population, and γ is the probability of misrepair.

The authors then consider a variety of special cases and make observations about each. A general theme is that low misrepair rates, high fitness of the drive, and low fitness of resistance alleles all act to improve drive spread. These are crucial points for understanding the evolutionary dynamics of HEG-based gene drive systems.

For a classic homing endonuclease gene drive, the latter two properties—high fitness of the drive and low fitness of resistance alleles—are naturally difficult to reconcile with each other, as we describe in Section 1.3. Since cost-free resistance to a drive construct can certainly arise, alternative drive designs are necessary for effective population modification. The CRISPR-based gene drive systems studied in this Chapter facilitate targeting arbitrary (and many) locations in a genome, which greatly expands the creative potential for manipulating wild populations. However, while CRISPR-based constructs offer enhanced opportunities for constructing gene drive systems, they also inevitably exhibit more complex dynamics that must be firmly understood.

1.6.2 Model with only costly resistance

In this Section, we present and analyze our model for a CRISPR-based gene drive system featuring n gRNAs and n costly resistant alleles ($\mathrm{R}_1, \ldots, \mathrm{R}_n$, as described in Fig. 1.2). We later extend this model to also include the neutral resistant alleles $\mathrm{S}_1, \ldots, \mathrm{S}_n$ (Section 1.6.3), but for simplicity we begin with only the former class of resistant alleles.

1.6.2.1 Model description

To describe the evolutionary dynamics of such a system, we consider a population of diploid organisms featuring a drive allele, D, a wild-type allele, 0, and n resistance alleles, i (with $1 \leq i \leq n$). (In Section 1.3, we use the notation "W" for a wild-type allele rather than "0". The notation "0" is more natural for doing calculations.) There are $(n+2)(n+3)/2$ possible genotypes in the population: ij (with $0 \leq i \leq n$ and $0 \leq j \leq n$), iD (with $0 \leq i \leq n$), and DD. The drive mechanism works as follows.

Consider a type $0D$ individual; one allele is wild-type, and the other allele is the drive. There are n guide RNAs and therefore n targets for the drive to cut. At meiosis, the drive can cut any number of targets between 0 and n. If the drive cuts no targets, then the individual remains with genotype $0D$. If the drive cuts k targets (with $1 \leq k \leq n$), then one of several things can happen: One possibility is that homologous recombination copies the drive allele onto the damaged chromosome, so that the individual's genotype becomes DD. This is how the drive construct effects its spread through a population. Another possibility is that non-homologous end joining repairs the damaged chromosome without restoring the lost targets, so that the individual's genotype becomes iD (with $1 \leq i \leq n$). This is how resistance to the drive construct emerges. Yet another possibility is that non-homologous end joining perfectly repairs the damaged chromosome, so that the individual's genotype remains $0D$.

The drive allele can effect its spread as long as there is at least one remaining target. In an individual with genotype iD, either the drive cuts at no targets, with the individual's genotype remaining iD, or the drive cuts at some number, k, of the $n-i$ remaining targets (so that $1 \leq k \leq n-i$). After cutting, the individual can become homozygous in the drive allele (DD), the individual can lose additional targets by acquiring genotype jD (with $i+1 \leq j \leq n$), or the individual can remain with genotype iD.

Using these rules, we can formally express the rates at which each of the $n+2$ types of gametes are produced in terms of the frequencies of individuals in the population. We denote by $F_D(t)$ the rate (at time t) at which drive gametes (D) are produced by individuals in the population. We denote by $F_i(t)$ the rate (at time t) at which wild-type gametes ($i = 0$) or gametes with varying levels of resistance ($1 \leq i \leq n$) are produced by individuals in the population. We have

$$\begin{aligned} F_D(t) &= f_{DD} x_{DD}(t) + \sum_{k=0}^{n} p_{kD,D} f_{kD} x_{kD}(t) \\ F_i(t) &= \sum_{k=0}^{i} p_{kD,i} f_{kD} x_{kD}(t) + \sum_{k=0}^{n} \frac{1+\delta_{ki}}{2} f_{ki} x_{ki}(t). \end{aligned} \tag{1.3}$$

Here, δ_{ki} is the Kronecker delta. We use the following notation: $x_{ki}(t)$ denotes the frequency of individuals (at time t) with only wild-type or resistance alleles, $x_{kD}(t)$ denotes the frequency of individuals (at time t) with one wild-type or resistance allele and one drive allele, and $x_{DD}(t)$ denotes the frequency of individuals (at time t) that are homozygous in the drive allele. (We define $x_{ki}(t)$ for $k \neq i$ and $x_{kD}(t)$ such that the ordering of the indices does not matter, i.e., $x_{ki}(t) = x_{ik}(t)$ is the frequency of individuals with one copy of the k allele and one copy of the i allele, and $x_{kD}(t) = x_{Dk}(t)$ is the frequency of individuals with one copy of the k allele and one copy of the drive allele.) f_{ki} denotes the fitness of individuals with only wild-type or resistance alleles, f_{kD} denotes the fitness of individuals with one wild-type or resistance allele and one drive allele, and f_{DD} denotes the fitness of individuals that are homozygous in the drive allele. $p_{kD,D}$ denotes the probability that an individual of genotype kD produces a D gamete. $p_{kD,i}$ denotes the probability that an individual of genotype kD produces an i gamete. From conservation of probability, we have the following identity:

$$p_{kD,D} + \sum_{i=k}^{n} p_{kD,i} = 1.$$

Notice that a type nD individual is fully resistant to being manipulated by the drive construct; such a fully resistant individual shows standard Mendelian segregation in its production of gametes. Thus, we have

$$p_{nD,n} = \frac{1}{2}.$$

We understand Equations (1.3) as follows: Type DD individuals only produce type D gametes, hence the term $f_{DD}x_{DD}(t)$ in the equation for $F_D(t)$. Type kD individuals produce type D gametes with probability $p_{kD,D}$, hence the terms $p_{kD,D}f_{kD}x_{kD}(t)$ in the equation for $F_D(t)$. Type kD individuals produce type i gametes with probability $p_{kD,i}$, hence the terms $p_{kD,i}f_{kD}x_{kD}(t)$ in the equation for $F_i(t)$. Type ki individuals produce type i gametes with probability 1 if $k = i$ or with probability $1/2$ if $k \neq i$, hence the terms $[(1+\delta_{ki})/2] f_{ki} x_{ki}(t)$ in the equation for $F_i(t)$.

The selection dynamics are modeled by the following system of equations:

$$\begin{aligned} \dot{x}_{ij}(t) &= (2-\delta_{ij})\, F_i(t) F_j(t) - \psi^2(t) x_{ij}(t) \\ \dot{x}_{iD}(t) &= 2F_i(t) F_D(t) - \psi^2(t) x_{iD}(t) \\ \dot{x}_{DD}(t) &= F_D^2(t) - \psi^2(t) x_{DD}(t). \end{aligned} \tag{1.4}$$

Here, an overdot denotes the time derivative, d/dt. In formulating the population dynamics, we assume random mating; i.e., two random gametes meet to form a new individual. Notice that the products $(2-\delta_{ij}) F_i(t) F_j(t)$, $2F_i(t) F_D(t)$, and $F_D^2(t)$ in Equations (1.4) represent the pairings of the different types of gametes to make new offspring. The quantity $\psi^2(t)$ represents a density-dependent death rate for the individuals in the population.

At any given time, t, we require that the total number of individuals sums to one:

$$x_{DD}(t) + \sum_{i=0}^{n} x_{iD}(t) + \sum_{i=0}^{n} \sum_{j=0}^{i} x_{ij}(t) = 1. \tag{1.5}$$

To enforce this density constraint, we set

$$\psi(t) = F_D(t) + \sum_{i=0}^{n} F_i(t). \tag{1.6}$$

Throughout this Chapter, we choose to work in the framework of continuous time (Equations

(1.4)), since we feel that this approach simplifies the mathematical analysis. In much of the remainder of this Chapter, we omit explicitly writing the time dependence on dynamical quantities for notational convenience.

1.6.2.2 Invasion of the drive construct

Consider a wild-type population in which all individuals have genotype 00. We perturb the wild-type population by introducing a small amount of the drive allele, D. What happens? Does the drive allele catalyze its own spread in the population, or is it eliminated?

For a perturbation to a wild-type population, we write the frequencies of the genotypes as

$$\begin{aligned}
x_{00} &= 1 \quad -\epsilon\delta_{00}^{(1)} - \epsilon^2\delta_{00}^{(2)} \quad -\mathcal{O}(\epsilon^3)\\
x_{0D} &= \quad +\epsilon\delta_{0D}^{(1)} + \epsilon^2\delta_{0D}^{(2)} \quad +\mathcal{O}(\epsilon^3)\\
x_{0i} &= \quad +\epsilon\delta_{0i}^{(1)} + \epsilon^2\delta_{0i}^{(2)} \quad +\mathcal{O}(\epsilon^3)\\
x_{ij} &= \quad + \epsilon^2\delta_{ij}^{(2)} \quad +\mathcal{O}(\epsilon^3)\\
x_{iD} &= \quad + \epsilon^2\delta_{iD}^{(2)} \quad +\mathcal{O}(\epsilon^3)\\
x_{DD} &= \quad + \epsilon^2\delta_{DD}^{(2)} \quad +\mathcal{O}(\epsilon^3)
\end{aligned} \tag{1.7}$$

In Equations (1.7), it is implied that $1 \leq i \leq n$ and $1 \leq j \leq n$. The expansions (1.7) are understood as follows. The frequency of the wild-type allele is approximately one, since we only introduce a small amount of the drive allele. The frequency of the drive allele is of order $\epsilon \ll 1$. The small number of $0D$ individuals in the population also produce resistance alleles, and the frequency of these resistance alleles shortly after the perturbation is also small (i.e., of order $\epsilon \ll 1$). Notice that:

- New type 00 individuals are produced by pairing two wild-type gametes (each at a frequency $\mathcal{O}(1)$), so new type 00 individuals are generated at a rate $\mathcal{O}(1)$.
- New type $0D$ individuals are produced by pairing a wild-type gamete (at a frequency $\mathcal{O}(1)$) and a drive gamete (at a frequency $\mathcal{O}(\epsilon)$), so new type $0D$ individuals are generated at a rate $\mathcal{O}(\epsilon)$.
- New type $0i$ individuals (for $1 \leq i \leq n$) are produced by pairing a wild-type gamete (at a frequency $\mathcal{O}(1)$) and a resistant gamete (at a frequency $\mathcal{O}(\epsilon)$), so new type $0i$ individuals are generated at a rate $\mathcal{O}(\epsilon)$.
- New type ij individuals (for $1 \leq i \leq n$ and $1 \leq j \leq n$) are produced by pairing two resistant gametes (each at a frequency $\mathcal{O}(\epsilon)$), so new type ij individuals are generated at a rate $\mathcal{O}(\epsilon^2)$.
- New type iD individuals (for $1 \leq i \leq n$) are produced by pairing a resistant gamete (at a frequency $\mathcal{O}(\epsilon)$) and a drive gamete (at a frequency $\mathcal{O}(\epsilon)$), so new type iD individuals are generated at a rate $\mathcal{O}(\epsilon^2)$.
- New type DD individuals are produced by pairing two drive gametes (each at a frequency $\mathcal{O}(\epsilon)$), so new type DD individuals are generated at a rate $\mathcal{O}(\epsilon^2)$.

Also, notice that a nonzero amount of the drive allele and the resistance alleles are produced at order ϵ^2 by type ij, iD, and DD individuals, so there also exist terms of order ϵ^2 in the expansions for x_{0D} and x_{0i}. Hence, we arrive at the expansions (1.7).

Note that (1.7) and (1.5) impose a constraint on the $\mathcal{O}(\epsilon)$ terms in the genotype frequencies:

$$\delta_{00}^{(1)} = \delta_{0D}^{(1)} + \sum_{i=1}^{n} \delta_{0i}^{(1)}. \tag{1.8}$$

Also, note that (1.7) and (1.5) impose a constraint on the $\mathcal{O}(\epsilon^2)$ terms in the genotype frequencies:

$$\delta_{00}^{(2)} = \delta_{0D}^{(2)} + \delta_{DD}^{(2)} + \sum_{i=1}^{n} \delta_{0i}^{(2)} + \sum_{i=1}^{n} \delta_{iD}^{(2)} + \sum_{i=1}^{n} \sum_{j=1}^{i} \delta_{ij}^{(2)}.$$

Substituting (1.6), (1.3), (1.7), and (1.8) into the equation for $\dot{x}_{0D}$ in (1.4), we obtain

$$\dot{\delta}_{0D}^{(1)} = f_{00} \left(2p_{0D,D} f_{0D} - f_{00}\right) \delta_{0D}^{(1)}.$$

The drive allele invades a wild-type population if $\dot{\delta}_{0D}^{(1)} > 0$, i.e., if

$$2p_{0D,D} f_{0D} > f_{00}. \tag{1.9}$$

1.6.2.3 Stability of the drive construct

Consider a population in which the drive construct has fixed, so that all individuals have genotype DD. We perturb the DD population by introducing a small amount of the wild-type allele, 0. What happens? Is the DD population stable to perturbations, or does the wild-type allele or one of the resistance alleles invade the population?

For a perturbation to a population in which the drive construct has fixed, we write the frequencies of the genotypes as

$$\begin{aligned}
x_{DD} &= 1 \quad -\epsilon\delta_{DD}^{(1)} - \epsilon^2\delta_{DD}^{(2)} \quad -\mathcal{O}(\epsilon^3) \\
x_{iD} &= \qquad +\epsilon\delta_{iD}^{(1)} + \epsilon^2\delta_{iD}^{(2)} \quad +\mathcal{O}(\epsilon^3) \\
x_{ij} &= \qquad\qquad + \epsilon^2\delta_{ij}^{(2)} \quad +\mathcal{O}(\epsilon^3)
\end{aligned} \tag{1.10}$$

In Equations (1.10), it is implied that $0 \le i \le n$ and $0 \le j \le n$. The expansions (1.10) are understood as follows. The frequency of the drive allele is approximately one, since we only introduce a small amount of the wild-type allele. The frequency of the wild-type allele is of order $\epsilon \ll 1$. The small number of $0D$ individuals in the population also produce resistance alleles, and the frequency of these resistance alleles shortly after the perturbation is also small (i.e., of order $\epsilon \ll 1$). Notice that:

- New type DD individuals are produced by pairing two drive gametes (each at a frequency $\mathcal{O}(1)$), so new type DD individuals are generated at a rate $\mathcal{O}(1)$.
- New type iD individuals (for $0 \le i \le n$) are produced by pairing a non-drive gamete (at a frequency $\mathcal{O}(\epsilon)$) and a drive gamete (at a frequency $\mathcal{O}(1)$), so new type iD individuals are generated at a rate $\mathcal{O}(\epsilon)$.
- New type ij individuals (for $0 \le i \le n$ and $0 \le j \le n$) are produced by pairing two non-drive gametes (each at a frequency $\mathcal{O}(\epsilon)$), so new type ij individuals are generated at a rate $\mathcal{O}(\epsilon^2)$.

Also, notice that a nonzero amount of the non-drive alleles are produced at order ϵ^2 by type ij individuals, so there also exist terms of order ϵ^2 in the expansions for x_{iD}. Hence, we arrive at the expansions (1.10).

Note that (1.10) and (1.5) impose a constraint on the $\mathcal{O}(\epsilon)$ terms in the genotype frequencies:

$$\delta_{DD}^{(1)} = \sum_{i=0}^{n} \delta_{iD}^{(1)}. \tag{1.11}$$

Also, note that (1.10) and (1.5) impose a constraint on the $\mathcal{O}(\epsilon^2)$ terms in the genotype frequencies:

$$\delta_{DD}^{(2)} = \sum_{i=0}^{n} \delta_{iD}^{(2)} + \sum_{i=0}^{n} \sum_{j=0}^{i} \delta_{ij}^{(2)}. \tag{1.12}$$

Substituting (1.6), (1.3), (1.10), and (1.11) into the equations for $\dot{x}_{iD}$ in (1.4), we obtain

$$\dot{\delta}_{iD}^{(1)} = B_i \delta_{iD}^{(1)} + \sum_{k=0}^{i-1} A_{k,i} \delta_{kD}^{(1)}. \tag{1.13}$$

Here, we use the shorthand notation

$$A_{k,i} = 2p_{kD,i} f_{kD} f_{DD}$$

$$B_i = A_{i,i} - f_{DD}^2.$$

To solve (1.13), we take its Laplace transform. Using the notation $\Delta_{iD}^{(1)}(s) = \mathcal{L}\{\delta_{iD}^{(1)}(t)\}(s) = \int_0^\infty e^{-st} \delta_{iD}^{(1)}(t) dt$, we have

$$\Delta_{iD}^{(1)}(s) = \frac{1}{s - B_i} \delta_{iD}^{(1)}(0) + \frac{1}{s - B_i} \sum_{k=0}^{i-1} A_{k,i} \Delta_{kD}^{(1)}(s) \tag{1.14}$$

Here, we use $\delta_{iD}^{(1)}(0)$ to denote $\delta_{iD}^{(1)}(t)$ evaluated at time $t = 0$. Equation (1.14) specifies $\Delta_{iD}^{(1)}(s)$ in terms of each $\Delta_{kD}^{(1)}(s)$ for which $0 \leq k < i$. Simplifying, we have

$$\begin{aligned}\Delta_{iD}^{(1)}(s) = \frac{\delta_{iD}^{(1)}(0)}{s - B_i} + \sum_{k=0}^{i-1} \frac{\delta_{kD}^{(1)}(0)}{s - B_k} \Bigg[& \frac{A_{k,i}}{s - B_i} \\ & + \sum_{u=k+1}^{i-1} \frac{A_{k,u} A_{u,i}}{(s - B_u)(s - B_i)} \\ & + \sum_{u=k+1}^{i-2} \sum_{v=u+1}^{i-1} \frac{A_{k,u} A_{u,v} A_{v,i}}{(s - B_u)(s - B_v)(s - B_i)} \\ & + \sum_{u=k+1}^{i-3} \sum_{v=u+1}^{i-2} \sum_{w=v+1}^{i-1} \frac{A_{k,u} A_{u,v} A_{v,w} A_{w,i}}{(s - B_u)(s - B_v)(s - B_w)(s - B_i)} \\ & + \cdots \Bigg] \end{aligned} \tag{1.15}$$

We are interested in the time dependence of $\delta_{iD}^{(1)}(t)$. From Equation (1.15), notice that when the Laplace transform is inverted, the time dependence of each term in the resulting equation for $\delta_{iD}^{(1)}(t)$ has the form $t^{\alpha} \exp(B_j t)$, where $\alpha \geq 0$.

To demonstrate this, consider a set of real numbers $\{\beta_j\}$ and a set of positive integers $\{\nu_j\}$, and define $\mathcal{F}_k(s)$ for $k \geq 0$:

$$\mathcal{F}_k(s) = \prod_{j=0}^{k} \frac{1}{(s - \beta_j)^{\nu_j}}$$

If the inverse Laplace transform of $\mathcal{F}_k(s)$, denoted by $\mathcal{L}^{-1}\{\mathcal{F}_k(s)\}(t)$, is equal to a sum of factors of the form $\mathcal{L}^{-1}\{1/(s - \beta_j)^{\xi}\}(t)$, where ξ is a positive integer, then each term in the solution for $\delta_{iD}^{(1)}(t)$ has the form $t^{\alpha} \exp(B_j t)$, where $\alpha \geq 0$.

To prove that $\mathcal{L}^{-1}\{\mathcal{F}_k(s)\}(t)$ is equal to a sum of factors of the form $\mathcal{L}^{-1}\{1/(s - \beta_j)^{\xi}\}(t)$, we use induction. Define

$$\mathcal{G}_{k+1}(t) = \mathcal{L}^{-1}\left\{\mathcal{F}_{k+1}(s)\right\}(t) = \mathcal{L}^{-1}\left\{\mathcal{F}_k(s)\frac{1}{(s-\beta_{k+1})^{\nu_{k+1}}}\right\}(t) \tag{1.16}$$

The inverse Laplace transform in (1.16) is calculated as follows:

$$\mathcal{G}_{k+1}(t) = \int_0^t d\tau \ \left[\mathcal{L}^{-1}\left\{\mathcal{F}_k(s)\right\}(\tau)\right]\left[\mathcal{L}^{-1}\left\{\frac{1}{(s-\beta_{k+1})^{\nu_{k+1}}}\right\}(t-\tau)\right] \tag{1.17}$$

First, for the base case, consider Equation (1.16) for $k = 0$. We have

$$\mathcal{G}_1(t) = \mathcal{L}^{-1}\left\{\frac{1}{(s-\beta_0)^{\nu_0}}\frac{1}{(s-\beta_1)^{\nu_1}}\right\}(t) \tag{1.18}$$

From (1.17), this becomes

$$\mathcal{G}_1(t) = \int_0^t d\tau \ \left[\mathcal{L}^{-1}\left\{\frac{1}{(s-\beta_0)^{\nu_0}}\right\}(\tau)\right]\left[\mathcal{L}^{-1}\left\{\frac{1}{(s-\beta_1)^{\nu_1}}\right\}(t-\tau)\right]$$

Substituting the expressions for $\mathcal{L}^{-1}\{1/(s-\beta_0)^{\nu_0}\}(\tau)$ and $\mathcal{L}^{-1}\{1/(s-\beta_1)^{\nu_1}\}(t-\tau)$, the equation for $\mathcal{G}_1(t)$ becomes

$$\mathcal{G}_1(t) = \int_0^t d\tau \ \left[\frac{\tau^{\nu_0-1}e^{\beta_0\tau}}{(\nu_0-1)!}\right]\left[\frac{(t-\tau)^{\nu_1-1}e^{\beta_1(t-\tau)}}{(\nu_1-1)!}\right]$$

Performing the integration over τ, we have

$$
\begin{aligned}
\mathcal{G}_1(t) = {} & \frac{(-1)^{\nu_0}}{(\nu_0-1)!(\nu_1-1)!(\beta_0-\beta_1)^{\nu_0}} \sum_{j=0}^{\nu_1-1} \binom{\nu_1-1}{j} \frac{(j+\nu_0-1)!}{(\beta_0-\beta_1)^j} \\
& \times \Bigg[(\nu_1-j-1)! \mathcal{L}^{-1}\left\{ \frac{1}{(s-\beta_1)^{\nu_1-j}} \right\}(t) \\
& - \sum_{k=0}^{j+\nu_0-1} (-1)^k \frac{(\nu_1-j+k-1)!}{k!} (\beta_0-\beta_1)^k \mathcal{L}^{-1}\left\{ \frac{1}{(s-\beta_0)^{\nu_1-j+k}} \right\}(t) \Bigg]
\end{aligned}
$$

Manipulating the indices and simplifying, we obtain

$$
\begin{aligned}
& \mathcal{G}_1(t) = \frac{(-1)^{\nu_0}}{(\nu_0-1)!(\nu_1-1)!(\beta_0-\beta_1)^{\nu_0+\nu_1}} \\
& \times \Bigg[\sum_{j=1}^{\nu_1} \mathcal{L}^{-1}\left\{ \frac{1}{(s-\beta_1)^j} \right\}(t) \binom{\nu_1-1}{\nu_1-j} (\nu_0+\nu_1-j-1)!(j-1)!(\beta_0-\beta_1)^j \\
& - \sum_{j=1}^{\nu_1} \mathcal{L}^{-1}\left\{ \frac{1}{(s-\beta_0)^j} \right\}(t) \sum_{k=0}^{j-1} (-1)^k \binom{\nu_1-1}{\nu_1-j+k} \frac{(\nu_0+\nu_1-j+k-1)!(j-1)!}{k!} (\beta_0-\beta_1)^j \\
& - \sum_{j=\nu_1+1}^{\nu_0+\nu_1-1} \mathcal{L}^{-1}\left\{ \frac{1}{(s-\beta_0)^j} \right\}(t) \sum_{k=0}^{\nu_1-1} (-1)^{\nu_1-j-k} \binom{\nu_1-1}{k} \frac{(\nu_0+k-1)!(j-1)!}{(j+k-\nu_1)!} (\beta_0-\beta_1)^j \Bigg]
\end{aligned}
\tag{1.19}
$$

We see that $\mathcal{G}_1(t)$ is equal to a sum of factors of the form $\mathcal{L}^{-1}\{1/(s-\beta_j)^\xi\}(t)$.

Next, consider Equation (1.16) for $k > 0$. From (1.17), we have

$$
\mathcal{G}_{k+2}(t) = \int_0^t d\tau \; \left[\mathcal{L}^{-1}\left\{ \mathcal{F}_{k+1}(s) \right\}(\tau) \right] \left[\mathcal{L}^{-1}\left\{ \frac{1}{(s-\beta_{k+2})^{\nu_{k+2}}} \right\}(t-\tau) \right]
$$

This is equal to

$$\mathcal{G}_{k+2}(t) = \int_0^t d\tau \ \left[\mathcal{G}_{k+1}(\tau)\right] \left[\mathcal{L}^{-1}\left\{\frac{1}{(s-\beta_{k+2})^{\nu_{k+2}}}\right\}(t-\tau)\right] \tag{1.20}$$

For the inductive step, suppose that $\mathcal{G}_{k+1}(t)$ reduces to a sum of factors of the form $\mathcal{L}^{-1}\{1/(s - \beta_j)^{\xi}\}(t)$:

$$\mathcal{G}_{k+1}(t) = \sum_j \sum_i \mathcal{L}^{-1}\left\{\frac{1}{(s-\beta_j)^{\xi_i}}\right\}(t) \tag{1.21}$$

Substituting (1.21) into (1.20), we have

$$\mathcal{G}_{k+2}(t) = \sum_j \sum_i \int_0^t d\tau \ \left[\mathcal{L}^{-1}\left\{\frac{1}{(s-\beta_j)^{\xi_i}}\right\}(\tau)\right] \left[\mathcal{L}^{-1}\left\{\frac{1}{(s-\beta_{k+2})^{\nu_{k+2}}}\right\}(t-\tau)\right]$$

This is equal to

$$\mathcal{G}_{k+2}(t) = \sum_j \sum_i \mathcal{L}^{-1}\left\{\frac{1}{(s-\beta_j)^{\xi_i}} \frac{1}{(s-\beta_{k+2})^{\nu_{k+2}}}\right\}(t)$$

Then from Equations (1.18) and (1.19), we see that $\mathcal{G}_{k+2}(t)$ also necessarily reduces to a sum of factors of the form $\mathcal{L}^{-1}\{1/(s-\beta_j)^{\xi}\}(t)$, thus completing the proof.

Since $\delta_{iD}^{(1)}$ is equal to a sum of factors of the form $t^{\alpha}\exp(B_j t)$, where $\alpha \geq 0$, we see that if all $B_j < 0$, then all $\delta_{iD}^{(1)}$ approach zero in the long-time limit, and, from (1.11), we have that $\delta_{DD}^{(1)}$ approaches zero in the long-time limit. Therefore, if $B_j < 0$ for all values of $0 \leq j \leq n$, then the drive construct is evolutionarily stable.

If, instead, $B_j > 0$ for at least one value of j, then $\delta_{iD}^{(1)}$ has a term whose magnitude grows exponentially in time. The leading-order (in ϵ) terms in the expansions for x_{iD} in (1.10) are necessarily positive. Therefore, if the condition $B_j > 0$ is satisfied for at least one value of j, then $\delta_{iD}^{(1)}$ is positive and grows exponentially in time; i.e., the DD population is unstable to perturbations.

The resulting condition is that the DD population is stable to perturbations with a wild-type allele if

$$2 \max \left(p_{kD,k} f_{kD}\right) < f_{DD}. \tag{1.22}$$

Completely recessive fitness cost for a resistance mutation Now, we consider a special case in which the fitness cost associated with having resistance to the drive is completely recessive. If the fitness of each heterozygote with a resistance allele, f_{kD}, exactly equals f_{DD} for all k, then is the DD population stable to perturbations? We expect that $p_{kD,k} < 1/2$ for all $0 \leq k < n$. Therefore, if $f_{kD} = f_{DD}$ for all k, then the inequality (1.22) is satisfied for all $k < n$ and becomes an equality for $k = n$.

All resistance alleles with at least one target ($0 \leq k < n$) are removed from the population by selective forces. We must focus on the fully resistant allele, n. To probe the stability of the DD population, we substitute (1.6), (1.3), (1.10), (1.11), and (1.12) into (1.4), and we keep terms that are $\mathcal{O}(\epsilon^2)$. We have

$$-\dot{\delta}_{DD}^{(2)} = f_{DD}\left(f_{DD} - 2 f_{nn}\right) \delta_{nn}^{(2)} + \frac{1}{4} f_{DD}^2 \left[\delta_{DD}^{(1)}\right]^2. \tag{1.23}$$

We also have

$$\dot{\delta}_{nn}^{(2)} = -f_{DD}^2 \delta_{nn}^{(2)} + \frac{1}{4} f_{DD}^2 \left[\delta_{DD}^{(1)}\right]^2 . \tag{1.24}$$

We can integrate (1.24). We get

$$\delta_{nn}^{(2)} = \frac{1}{4} \left[\delta_{DD}^{(1)}\right]^2 \left[1 - \exp\left(-f_{DD}^2 t\right)\right] . \tag{1.25}$$

We are interested in the regime $1 \ll t \ll \epsilon^{-1}$. We must consider the sign of $\dot{\delta}_{DD}^{(2)}$ at large times $t \gg 1$ but before the terms in (1.10) become similar in magnitude. Our condition for stability of the DD population is therefore

$$\lim_{\substack{\epsilon t \to 0 \\ t \to \infty}} \dot{\delta}_{DD}^{(2)} < 0.$$

Shortly after the perturbation, the exponential in the solution for $\delta_{nn}^{(2)}$ will approach zero. Substituting (1.25) into (1.23) and simplifying, we see that the DD population is stable to perturbations if

$$f_{nn} < f_{DD}. \tag{1.26}$$

1.6.2.4 Interior equilibria

A drive construct increases in frequency when rare if Equation (1.9) is satisfied. A drive construct that has already fixed is stable to perturbations if Equation (1.22) is satisfied (or if Equation (1.26) is satisfied for the case of a completely recessive fitness cost for resistance). But if a small amount of the drive construct is introduced into a wild-type population, then does the drive spread completely to fixation?

To answer this question, it is helpful to know if the model for the drive dynamics, Equations (1.4), admits an interior equilibrium. Notice that, if all time derivatives are zero, then Equations (1.4) simplify to

$$\begin{aligned}
x_{ij} &= \frac{(2-\delta_{ij})\,F_iF_j}{\psi^2}\\
x_{iD} &= \frac{2F_iF_D}{\psi^2}\\
x_{DD} &= \frac{F_D^2}{\psi^2}.
\end{aligned}$$

Next, we define x_i to equal the frequency of allele i in the population. Thus, x_0 is the frequency of the wild-type allele, and x_i for $1 \leq i \leq n$ is the frequency of a resistance allele with i damaged targets. Also, x_D is the frequency of the drive allele. These allele frequencies can be calculated from the frequencies of individuals of the various genotypes:

$$\begin{aligned}
x_i &= \frac{1}{2}x_{iD} + \sum_{j=0}^{n} \frac{1+\delta_{ij}}{2} x_{ij}\\
x_D &= x_{DD} + \frac{1}{2}\sum_{i=0}^{n} x_{iD}.
\end{aligned}$$

Similar to Equation (1.5), the sum of all allele frequencies equals one at all times:

$$x_D + \sum_{i=0}^{n} x_i = 1.$$

We directly compute the following results:

$$\begin{aligned}
x_i^2 &= \left(\frac{1}{2}x_{iD} + \sum_{j=0}^{n} \frac{1+\delta_{ij}}{2} x_{ij}\right)^2 = \frac{F_i^2}{\psi^4}\left(F_D + \sum_{j=0}^{n} F_j\right)^2 = \frac{F_i^2}{\psi^2} = x_{ii} \\
x_D^2 &= \left(x_{DD} + \frac{1}{2}\sum_{j=0}^{n} x_{jD}\right)^2 = \frac{F_D^2}{\psi^4}\left(F_D + \sum_{j=0}^{n} F_j\right)^2 = \frac{F_D^2}{\psi^2} = x_{DD} \\
2x_i x_D &= \frac{2F_iF_D}{\psi^2} = x_{iD} \\
(2-\delta_{ij})\, x_i x_j &= \frac{(2-\delta_{ij})\, F_iF_j}{\psi^2} = x_{ij}.
\end{aligned} \tag{1.27}$$

In summary, we obtain

$$\begin{aligned}
x_{ij} &= (2-\delta_{ij})\, x_i x_j \\
x_{iD} &= 2x_i x_D \\
x_{DD} &= x_D^2.
\end{aligned} \tag{1.28}$$

From (1.27), we have that

$$\begin{aligned}
\psi x_i &= F_i \\
\psi x_D &= F_D.
\end{aligned} \tag{1.29}$$

By substituting Equation (1.6) for ψ and Equations (1.3) for F_i and F_D into (1.29), and substituting (1.28), we obtain

$$\left(f_{DD}x_D^2 + 2\sum_{k=0}^{n} f_{kD}x_kx_D + \sum_{j=0}^{n}\sum_{k=0}^{j} (2-\delta_{jk})\, f_{jk}x_jx_k\right) x_i = 2\sum_{k=0}^{i} p_{kD,i} f_{kD} x_k x_D + \sum_{k=0}^{n} f_{ki}x_kx_i. \tag{1.30}$$

We also obtain

$$\left(f_{DD}x_D^2 + 2\sum_{k=0}^{n} f_{kD}x_k x_D + \sum_{j=0}^{n}\sum_{k=0}^{j}\left(2-\delta_{jk}\right) f_{jk}x_j x_k\right) x_D = f_{DD}x_D^2 + 2\sum_{k=0}^{n} p_{kD,D} f_{kD} x_k x_D. \tag{1.31}$$

Equations (1.30) and (1.31) must be simultaneously satisfied for $0 \leq x_D \leq 1$ and $0 \leq x_i \leq 1$ for each i at each interior fixed point. If Equations (1.30) and (1.31) cannot be simultaneously solved for a given set of parameter values, then no interior fixed point exists.

1.6.2.5 Numerical examples

Numerical simulations of Equations (1.4) are helpful for understanding the evolutionary dynamics of a drive construct. For simplicity, we consider a single guide ($n = 1$), and we choose the following parameter values:

$$\begin{aligned} f_{00} = f_{10} &= 1 \\ f_{0D} = f_{1D} = f_{DD} &= 1 - c \\ f_{11} &= 1 - s \\ p_{0D,0} &= 0. \end{aligned} \tag{1.32}$$

We make the following assumptions: The fitness cost of the drive, c, is dominant. The fitness cost of the resistant allele, s, is recessive. Also, the drive construct in a $0D$ heterozygote always cuts at the target, and either the drive allele is copied by homologous recombination or resistance emerges. Thus, we have $p_{0D,0} = 0$.

In Fig. 1.4 (a and b), numerical simulations demonstrate evolutionary invasion of the drive construct. For these simulations, the initial condition is $x_{AA} = 1 - 10^{-4}$ and $x_{DD} = 10^{-4}$. The relevant condition for determining evolutionary invasion is Equation (1.9).

- In Fig. 1.4 (a), we set $p_{0D,D} = 0.75$ and $s = 0.4$. From Equation (1.9), the critical value of c for invasion is $1/3$. If $c = 0.34$ (green curve), then the drive construct does not invade. If $c = 0.33$ (blue curve), then the drive construct invades.
- In Fig. 1.4 (b), we set $p_{0D,D} = 0.65$ and $s = 0.3$. From Equation (1.9), the critical value of c for invasion is approximately 0.23. If $c = 0.235$ (green curve), then the drive construct does not invade. If $c = 0.225$ (blue curve), then the drive construct invades.

In Fig. 1.4 (c and d), numerical simulations demonstrate evolutionary stability of the drive construct. For these simulations, the initial condition is $x_{DD} = 1 - 10^{-2}$ and $x_{AA} = 10^{-2}$. From (1.32), notice that the condition (1.22) becomes an equality. Therefore, the relevant condition for determining evolutionary stability is Equation (1.26).

- In Fig. 1.4 (c), we set $p_{0D,D} = 0.75$ and $c = 0.32$. From Equation (1.26), the critical value of s for stability is 0.32. If $s = 0.315$ (green curve), then the drive construct is unstable. If $s = 0.325$ (blue curve), then the drive construct is stable.
- In Fig. 1.4 (d), we set $p_{0D,D} = 0.65$ and $c = 0.2$. From Equation (1.26), the critical value of s for stability is 0.2. If $s = 0.195$ (green curve), then the drive construct is unstable. If $s = 0.205$ (blue curve), then the drive construct is stable.

In Fig. 1.4 (e and f), numerical simulations demonstrate the behavior of the drive construct at intermediate frequencies. For these simulations, the initial condition is $x_{AA} = 1 - 10^{-4}$ and $x_{DD} = 10^{-4}$. If Equations (1.30) and (1.31) cannot simultaneously be solved numerically, then there is no interior equilibrium.

- In Fig. 1.4 (e), we set $p_{0D,D} = 0.75$ and $c = 0.32$. From numerical analysis of Equations (1.30) and (1.31), values of s that are slightly below approximately 0.815 permit an interior equilibrium, while values of s that are slightly above approximately 0.815 do not. If $s = 0.81$ (green curve), then the drive construct reaches an equilibrium frequency that is strictly between 0 and 1. If $s = 0.82$ (blue curve), then the drive construct spreads to fixation.

- In Fig. 1.4 (f), we set $p_{0D,D} = 0.65$ and $c = 0.2$. From numerical analysis of Equations (1.30) and (1.31), values of s that are slightly below approximately 0.285 permit an interior equilibrium, while values of s that are slightly above approximately 0.285 do not. If $s = 0.28$ (green curve), then the drive construct reaches an equilibrium frequency that is strictly between 0 and 1. If $s = 0.29$ (blue curve), then the drive construct spreads to fixation.

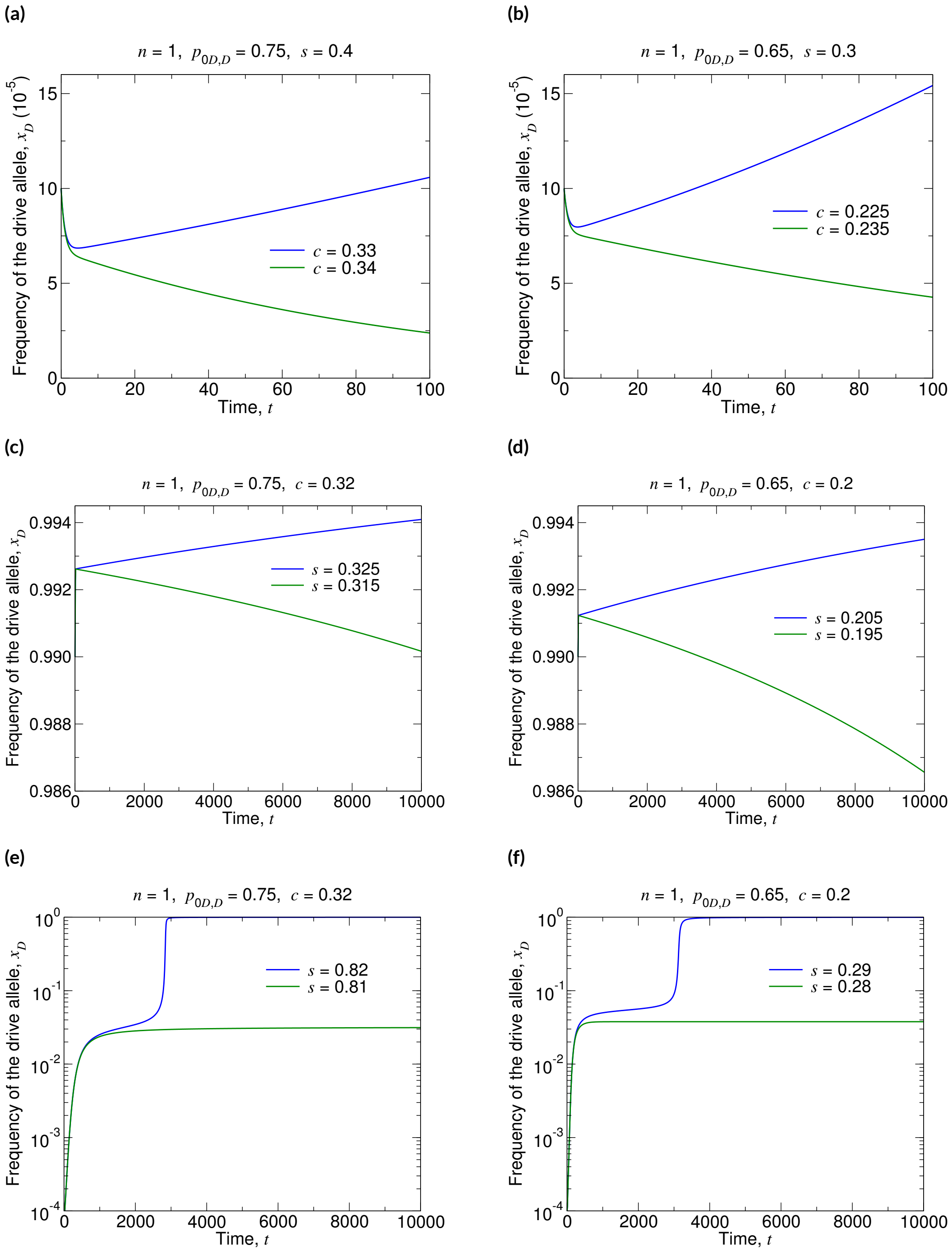


Figure 1.4: Numerical simulations of the evolutionary dynamics.

1.6.3 Model with neutral resistance

In this Section, we present an extension of the model that accounts for the phenomenon of "neutral resistance", concluding with the model presented in Section 1.5. Neutral resistance can occur if non-homologous end joining results in repair at a cut site that disrupts the recognition sequence of a guide RNA while nonetheless leaving the function of the target gene intact. This can occur, for example, via an in-frame insertion or deletion or a synonymous mutation. The resulting allele is similar (with respect to the drive mechanism) to the resistant alleles discussed in previous sections: the repaired target is immune to cutting by its corresponding guide RNA. However, the mutation conferring this resistance is not deleterious.

1.6.3.1 Model description

We represent this scenario of neutral resistance by an extension of our original model developed in Section 1.6.2. We consider a drive allele, D, n "costly" resistant alleles, R_i (with $1 \leq i \leq n$), n "neutral" resistant alleles, S_i (with $1 \leq i \leq n$), and the wild-type allele, S_0. The drive mechanism works as follows (see Fig.1.2 for an illustration).

Consider a type S_0D individual; one allele is wild-type, and the other allele is the drive. There are n guide RNAs and therefore n targets for the drive to cut. At meiosis, the drive can cut any number of targets between 0 and n. If the drive cuts no targets, then the individual remains with genotype S_0D. If the drive cuts k targets (with $1 \leq k \leq n$), then one of several things can happen: One possibility is that homologous recombination copies the drive allele onto the damaged chromosome, so that the individual's genotype becomes DD. Another possibility is that non-homologous end joining repairs the damaged chromosome without restoring the lost targets, and the resulting resistant allele is either costly, in which case the individual's genotype becomes R_iD (with $1 \leq i \leq n$), or cost-free, in which case the individual's genotype becomes S_iD (with $1 \leq i \leq n$). Yet another possibility is that non-homologous end joining perfectly repairs the damaged chromosome, so that the individual's genotype remains S_0D.

The drive allele can effect its spread as long as there is at least one remaining target. In an individual with genotype R_iD or S_iD, either the drive cuts at no targets, with the individual's genotype remaining R_iD or S_iD, or the drive cuts at some number, k, of the $n - i$ remaining targets (so that $1 \leq k \leq n - i$). After cutting, the individual can become homozygous in the drive allele (DD), the individual can lose additional targets by acquiring genotype R_jD or S_jD (with $i + 1 \leq j \leq n$), or the individual can remain with genotype R_iD or S_iD. We assume that costly resistant alleles R_i cannot convert to cost-free resistant alleles S_j, but cost-free resistant alleles S_i can convert to costly resistant alleles R_j.

Using these rules, we can formally express the rates at which each of the $2n + 2$ types of gametes are produced in terms of the frequencies of individuals in the population. We denote by $F_D(t)$ the rate (at time t) at which drive gametes (D) are produced by individuals in the population. We denote by $F_{S_i}(t)$ the rate (at time t) at which wild-type gametes ($i = 0$) or gametes with varying levels of cost-free resistance ($1 \leq i \leq n$) are produced by individuals in the population. And we denote by $F_{R_i}(t)$ the rate (at time t) at which gametes with varying levels of costly resistance ($1 \leq i \leq n$)

are produced by individuals in the population. We have

$$\begin{aligned}
F_D(t) &= f_{DD}x_{DD}(t) + \sum_{k=1}^{n} p_{R_kD,D} f_{R_kD} x_{R_kD}(t) + \sum_{k=0}^{n} p_{S_kD,D} f_{S_kD} x_{S_kD}(t) \\
F_{S_i}(t) &= \sum_{k=0}^{n} \frac{1+\delta_{ki}}{2} f_{S_kS_i} x_{S_kS_i}(t) + \frac{1}{2}\sum_{k=1}^{n} f_{R_kS_i} x_{R_kS_i}(t) + \sum_{k=0}^{i} p_{S_kD,S_i} f_{S_kD} x_{S_kD}(t) \\
F_{R_i}(t) &= \sum_{k=1}^{n} \frac{1+\delta_{ki}}{2} f_{R_kR_i} x_{R_kR_i}(t) + \frac{1}{2}\sum_{k=0}^{n} f_{R_iS_k} x_{R_iS_k}(t) \\
&\quad + \sum_{k=1}^{i} p_{R_kD,R_i} f_{R_kD} x_{R_kD}(t) + \sum_{k=0}^{i-1} p_{S_kD,R_i} f_{S_kD} x_{S_kD}(t).
\end{aligned}$$

Here, δ_{ki} is the Kronecker delta. $x_{IJ}(t)$ denotes the frequency of individuals (at time t) with genotype IJ, where $I, J = D, S_0, S_1, \ldots, S_n, R_1, \ldots, R_n$. Similarly, f_{IJ} is the fitness of IJ individuals, and $p_{IJ,K}$ denotes the probability of an individual with genotype IJ producing a K gamete. From conservation of probability, we have the following identities:

$$\begin{aligned}
p_{R_kD,D} + \sum_{i=k}^{n} p_{R_kD,R_i} &= 1 \\
p_{S_kD,D} + \sum_{i=k}^{n} p_{S_kD,S_i} + \sum_{i=k+1}^{n} p_{S_kD,R_i} &= 1
\end{aligned}$$

Notice that type R_nD and type S_nD individuals are fully resistant to being manipulated by the drive construct; such a fully resistant individual shows standard Mendelian segregation in its production of gametes. Thus, we have

$$p_{R_nD,R_n} = p_{S_nD,S_n} = \frac{1}{2}.$$

The selection dynamics are modeled by the following system of equations:

$$\dot{x}_{DD}(t) = F_D^2(t) - \psi^2(t)x_{DD}(t)$$

$$\dot{x}_{R_iD}(t) = 2F_{R_i}(t)F_D(t) - \psi^2(t)x_{R_iD}(t)$$

$$\dot{x}_{S_iD}(t) = 2F_{S_i}(t)F_D(t) - \psi^2(t)x_{S_iD}(t)$$

$$\dot{x}_{R_iS_j}(t) = 2F_{R_i}(t)F_{S_j}(t) - \psi^2(t)x_{R_iS_j}(t)$$

$$\dot{x}_{R_iR_j}(t) = (2-\delta_{ij})F_{R_i}(t)F_{R_j}(t) - \psi^2(t)x_{R_iR_j}(t)$$

$$\dot{x}_{S_iS_j}(t) = (2-\delta_{ij})F_{S_i}(t)F_{S_j}(t) - \psi^2(t)x_{S_iS_j}(t).$$

The quantity $\psi^2(t)$ represents a density-dependent death rate for the individuals in the population.

At any given time, t, we require that the total number of individuals sums to one:

$$x_{DD}(t)+\sum_{i=1}^{n} x_{R_iD}(t)+\sum_{i=0}^{n} x_{S_iD}(t)+\sum_{i=1}^{n}\sum_{j=0}^{n} x_{R_iS_j}(t)+\sum_{i=1}^{n}\sum_{j=1}^{i} x_{R_iR_j}(t)+\sum_{i=0}^{n}\sum_{j=0}^{i} x_{S_iS_j}(t) = 1$$

To enforce this density constraint, we set

$$\psi(t) = F_D(t) + \sum_{i=1}^{n} F_{R_i}(t) + \sum_{i=0}^{n} F_{S_i}(t).$$

1.6.3.2 Invasion of the drive construct

The steps for determining if the drive construct invades when there is neutral resistance are the same as in Section 1.6.2.2. The drive allele invades a wild-type population if

$$2p_{S_0D,D}f_{S_0D} > f_{S_0S_0}.$$

1.6.3.3 Stability of the drive construct

The steps for determining if the drive construct is stable when there is neutral resistance are the same as in Section 1.6.2.3. The DD population is stable to perturbations with a wild-type allele if

$$2\max_{A\in S\cup R}(p_{AD,A}f_{AD}) < f_{DD}.$$

1.6.4 Explicit cellular models of CRISPR gene drive

We now specify values of the inheritance probabilities, $p_{AB,C}$, and fitness values, f_{AB}, which explicitly describe possible scenarios by which a CRISPR gene drive acts within individuals. First, we specify a parameter set that corresponds with the behavior of CRISPR gene drives as described in prior literature. Then, we specify a parameter set that corresponds with our newly proposed CRISPR gene drive construct. These specified parameter sets for the previous and newly proposed drive constructs are used for the simulations of the previous and newly proposed drive constructs, respectively, in the numerical simulations presented earlier in Section 1.3.

1.6.4.1 Previous drives

For CRISPR gene drives as described in prior literature, $n = 1$. Reasonable choices for the fitness values and inheritance probabilities are as follows.

The wild-type has the maximum fitness of $f_{S_0 S_0} = 1$, and the cost-free resistant allele, S_1, is identical to the wild-type allele, S_0, with respect to fitness. Disruption of the target gene produces a recessive fitness cost, s, and the gene drive construct produces a dominant fitness cost, c. However, since the previously demonstrated drive constructs copied themselves by inserting at (and thus disrupting) the target sequence, the drive allele contains a disrupted copy of the target gene. Thus, DD and RD individuals incur both the cost of the drive construct, c, and the cost of resistance, s. These two costs can be assumed to be independent so that the corresponding fitness effects are multiplicative, i.e., $(1 - c)(1 - s)$. Therefore, we have the following fitness values: $f_{DD} = f_{RD} = (1 - c)(1 - s)$, $f_{SD} = 1 - c$, $f_{RR} = 1 - s$, and $f_{RS} = f_{SS} = 1$.

We then compute the drive-heterozygote gamete production probabilities as follows:

- $R_1 D$ individuals produce R_1 gametes and D gametes equiprobably because the single target site is resistant to cutting, so we have

$$p_{R_1 D, R_1} = p_{R_1 D, D} = \frac{1}{2}.$$

- $S_1 D$ individuals produce S_1 gametes and D gametes equiprobably because the single target site is resistant to cutting, so we have

$$p_{S_1 D, S_1} = p_{S_1 D, D} = \frac{1}{2}.$$

- $S_0 D$ individuals produce S_0 gametes precisely when no cutting occurs. Since cutting occurs with probability q, we have

$$p_{S_0D,S_0} = \frac{1-q}{2}.$$

- S_0D individuals produce D gametes by inheriting the existing D allele, or by cutting at the single target site with probability q and undergoing HR repair with probability P. We have

$$p_{S_0D,D} = \frac{1}{2} + \frac{qP}{2}.$$

- S_0D individuals produce S_1 gametes by cutting at the single target site with probability q, undergoing NHEJ repair with probability $1 - P$, and repairing the cut perfectly with probability γ. We have

$$p_{S_0D,S_1} = \frac{q(1-P)\gamma}{2}.$$

- S_0D individuals produce R_1 gametes by cutting at the single target site with probability q, undergoing NHEJ repair with probability $1 - P$, and repairing the cut imperfectly with probability $1 - \gamma$. We have

$$p_{S_0D,R_1} = \frac{q(1-P)(1-\gamma)}{2}.$$

1.6.4.2 Newly proposed drives

For our newly proposed CRISPR gene drive construct, any $n \geq 1$ is valid. Reasonable choices for the fitness values and inheritance probabilities are as follows.

The wild-type has the maximum fitness of $f_{S_0S_0} = 1$, and cost-free resistant alleles, S_i, are identical to the wild-type allele, S_0, with respect to fitness. The cost, c, conferred by the drive is dominant, while the cost, s, conferred by costly resistant alleles—which are disrupted copies of the target gene—is recessive. Furthermore, we assume that the drive allele contains a functional copy of the target gene, so DD and RD individuals do not incur the recessive fitness cost for target disruption. Thus, we have $f_{DD} = f_{RD} = f_{SD} = 1 - c$, $f_{RR} = 1 - s$, and $f_{RS} = f_{SS} = 1$.

We then assign values to the drive-heterozygote gamete production probabilities according to the biological description outlined in the main text and illustrated in Fig. 1.2B. We first define a probability density, $P_K(k \mid n, i, q)$, which describes the probability that k target sites undergo cutting, given that there are n total targets, of which i are currently resistant to cutting, and where each of the $n - i$ susceptible targets are cut independently with probability q. This distribution is binomial, specifically:

$$P_K(k \mid n, i, q) = \binom{n-i}{k} q^k (1-q)^{n-i-k}.$$

This distribution is defined for $0 \leq k \leq n - i$.

In the case that two or more cuts occur, we assume that all target sites between the two outermost cuts are lost due to loss of the intervening DNA sequence. To account for this effect, we further define a probability density, $P_L(l \mid k, n, i)$, which describes the probability that l targets are lost given k cuts, n total target sites, and i currently resistant sites. This distribution can be straightforwardly computed:

$$P_L(l \mid k, n, i) = (n - i - l + 1)\binom{l-2}{k-2} \bigg/ \binom{n-i}{k}.$$

This distribution is defined for $2 \leq k \leq l \leq n - i$.

We then compute the drive-heterozygote gamete production probabilities as follows:

- $R_i D$ individuals produce D gametes by inheriting the existing D allele, or by cutting at one or more sites on the R_i chromosome (each with probability q) and undergoing HR repair (with probability P). We have

$$p_{R_iD,D} = \frac{1}{2} + \frac{P}{2}(1 - (1 - q)^{n-i}).$$

- R_iD individuals produce R_i gametes precisely when no cutting occurs. Each of the $n - i$ sites is susceptible to cutting, and cutting occurs independently at each with probability q, so we have
$$p_{R_iD,R_i} = \frac{1}{2}(1 - q)^{n-i}.$$

- R_iD individuals produce R_{i+1} gametes (with $i < n$) by cutting at exactly one target site (where each is cut independently with probability q) and undergoing NHEJ repair (with probability $1 - P$). Since we assume that costly resistant alleles cannot convert back to cost-free alleles, we do not consider the efficacy of repair by NHEJ. In this case, we have
$$p_{R_iD,R_{i+1}} = \frac{1 - P}{2}(n - i)q(1 - q)^{n-i-1}.$$

- R_iD individuals produce R_k gametes (with $i + 2 \leq k \leq n$) by losing $k - i$ target sites and undergoing NHEJ repair (with probability $1 - P$). Since we assume that costly resistant alleles cannot convert back to cost-free alleles, we do not consider the efficacy of repair by NHEJ. In this case, we have
$$p_{R_iD,R_k} = \frac{1 - P}{2} \sum_{j=2}^{k-i} P_L(k - i \mid j, n, i) P_K(j \mid n, i, q).$$
The sum is over the number of simultaneous cuts, j, which could possibly give rise to a loss of $k - i$ targets.

- S_iD individuals produce D gametes by inheriting the existing D allele, or by cutting at one or more sites on the S_i chromosome (each with probability q) and undergoing HR repair (with probability P). We have
$$p_{S_iD,D} = \frac{1}{2} + \frac{P}{2}(1 - (1 - q)^{n-i}).$$

- S_iD individuals produce S_i gametes precisely when no cutting occurs. Each of the $n - i$ sites is susceptible to cutting, and cutting occurs independently at each with probability q, so we have

$$p_{S_iD,S_i} = \frac{1}{2}(1-q)^{n-i}.$$

- S_iD individuals produce S_{i+1} gametes (with $i < n$) by cutting at exactly one target site (where each is cut independently with probability q), undergoing NHEJ repair (with probability $1 - P$), and repairing the cut perfectly (with probability γ). We have

$$p_{S_iD,S_{i+1}} = \frac{1-P}{2}(n-i)q(1-q)^{n-i-1}\gamma.$$

- S_iD individuals do not produce S_k gametes when $k \geq i+2$. This is because cutting at two or more target sites would lead to a large deletion in the intervening DNA sequence, resulting in loss of target gene function. Thus

$$p_{S_iD,S_{i+2}} = \cdots = p_{S_iD,S_n} = 0.$$

- S_iD individuals produce R_{i+1} gametes (with $i < n$) by cutting at exactly one target site (where each is cut independently with probability q), undergoing NHEJ repair (with probability $1 - P$), and repairing the cut imperfectly (with probability $1 - \gamma$). We have

$$p_{S_iD,R_{i+1}} = \frac{1-P}{2}(n-i)q(1-q)^{n-i-1}(1-\gamma).$$

- S_iD individuals produce R_k gametes (with $i+2 \leq k \leq n$) by losing $k-i$ target sites and undergoing NHEJ repair (with probability $1 - P$). This is because cutting at two or more target sites would lead to a large deletion in the intervening DNA sequence, resulting in loss of target gene function. Thus we have

$$p_{S_iD,R_k} = \frac{1-P}{2}\sum_{j=2}^{k-i} P_L(k-i \mid j,n,i)P_K(j \mid n,i,q).$$

  The sum is over the number of simultaneous cuts, j, which could possibly give rise to a loss of $k-i$ targets.

For the numerical simulations of both the previous and newly proposed drive constructs shown in the main text, we set $q = P = 0.95$ and $\gamma = 1/3$.

2

Invasiveness of current CRISPR gene drives

2.1 Foreword

The results in Chapter 1 paint a general picture of CRISPR gene drive dynamics in infinite wild populations: drive elements spread when initially rare, persist in the population over some typical timescale and then go extinct in the long run. This chapter asks how these dynamics play out in finite populations with small initial introductions of drive-carrying organisms. The motivating scenario is a hypothetical field trial in which a drive element is released to alter only one population among several that are connected by low rates of migration. Our analysis asks how likely the drive is to be contained to the target population under various circumstances.

I performed this work together with Ben Adlam, who contributed great help with developing and analyzing the mathematical models presented here. We benefited immensely from insight, advising and support from George Church, Kevin Esvelt and Martin Nowak.

This chapter was first published in Ref. 40:

Charleston Noble*, Ben Adlam*, George M. Church, Kevin M. Esvelt and Martin A. Nowak. Current CRISPR gene drive systems are likely to be highly invasive in wild populations. *eLife* 7, e33423 (2018). (*equal contribution)

2.2 Introduction

Following reports of successful CRISPR gene drive systems in yeast[37] and fruit flies[38], scientists emphasized the need to employ strategies beyond traditional barrier containment as a laboratory safeguard[42,66]. These precautions were judged necessary to prevent unintended ecological effects, but also because any unauthorized release affecting a wild population could severely damage trust in scientists and governance, significantly delaying or even precluding applications of gene drive and other biotechnologies.

Drive resistance can result from, among other mechanisms, mutations that block cutting by the CRISPR nuclease, which can arise via de novo mutation, standing genetic variation or—as analyzed in detail in the previous chapter—because the drive itself is not perfectly efficient. Moreover, recent experimental and theoretical studies of resistance have predicted that this phenomenon will prevent drive fixation in large wild populations unless additional mitigating strategies are employed, such as multiplex targeting of essential genes using multiple gRNAs [28,39,41,53,62,67–69] (Section 1.3).

On the other hand, recent articles highlighting the problem of resistance have suggested that it might prevent drive *invasion* in wild populations—with some even implying that resistance could serve as a safeguard [70,71]. While we agree that resistance should prevent drive *fixation* in large populations, an allele can nonetheless spread to significant frequency without fixing. To clarify this point, we sought here to quantify the likelihood and magnitude of spread in the most likely unauthorized release scenario—a small number of engineered individuals released into a wild population.

As discussed in the previous chapter, CRISPR-based gene drive systems function by converting drive-heterozygotes into homozygotes in the late germline or early embryo [41] (Fig. 2.1A). First, a CRISPR nuclease encoded in the drive construct cuts at the corresponding wild-type allele—its target prescribed by one or more independently expressed guide RNAs (gRNAs)—producing a double-strand break [33]. This break is then repaired either through homology-directed repair, producing a second copy of the gene drive construct, or through a nonhomologous repair pathway (non-homologous end joining, NHEJ, or microhomology-mediated end joining, MMEJ), which typically introduces a mutation at the target site [34,35]. Because the drive target is determined through sequence homology, such a mutation generally results in resistance to future cutting by the gene drive. Thus, the allele converts from a wild-type to resistant allele if it undergoes repair by a pathway other than homology-directed repair. Moreover, drive-resistant alleles are expected to exist in wild populations simply due to standing genetic variation[62,68].

Deterministic models, which assume an infinite, well-mixed population, predict whether an allele is favored to increase in frequency when initially rare in a wild population. Whether gene drives are predicted to invade by deterministic models depends on two key parameters: the homing efficiency (P), or the probability of undergoing homology-directed repair instead of nonhomologous repair, and fitness (f), or the relative fecundity or death rate the drive and its cargo confer on their organism compared to the wild-type. Mathematically, drives are initially favored by selection if $f(1+P) > 1$ (which arises from Eq. (1.1) with $p_{WD,D} = (1+P)/2$, $f_{WW} = 1$, and $f_{WD} = f$), i.e., if the inheritance bias of the drive exceeds its fitness penalty[39,53,72]. Given that the homing efficiencies of reported drive systems typically range from 0.37 to 0.99 (Table A.1), current drive systems can clearly invade in deterministic models. Although the fitness parameter, f, is typically not measured in proof-of-concept studies, a substantial fitness cost is tolerable by all reported CRISPR drive constructs[5,13,37,38,67] (Fig. 2.1B).

However, in finite populations, the fate of initially rare alleles is determined not only by selection but also by stochastic fluctuations[73–75]. Therefore, stochastic models are required to predict the probability that a drive will invade a population upon the introduction of a very small number of individuals, even when deterministic models predict that they are to invade. A previous, and arguably prescient, stochastic model of endonuclease drive containment found that homing-based drives, such as those subsequently developed using CRISPR, were among the likeliest to invade of the various drive alternatives[76]. To determine whether self-propagating homing drives are still able to invade in the presence of resistance, we formulated a finite population, stochastic, Moran-based model that allows us to study small releases in finite and structured populations (Section 2.5).

2.3 Results

Our model considers three distinct allelic classes: wild-type (W), gene drive (D), and resistant (R). Consistent with experiments, we assume that the drive invariably cuts the wild-type allele in the germline of a heterozygous WD individual, converting to a drive allele with probability P, or a resistant allele with probability $1 - P$. Each genotype, AB, has a relative reproductive rate, f_{AB}, corresponding to its fitness in deterministic models, normalized such that the wild-type homozygote has fitness one ($f_{WW} = 1$), the drive confers a dominant cost ($f_{DW} = f_{DD} = f_{DR} < 1$), and resistance is neutral ($f_{WR} = f_{RR} = 1$). This ordering of the parameters conservatively represents the worst-case scenario for drive spread (Section 2.5.9).

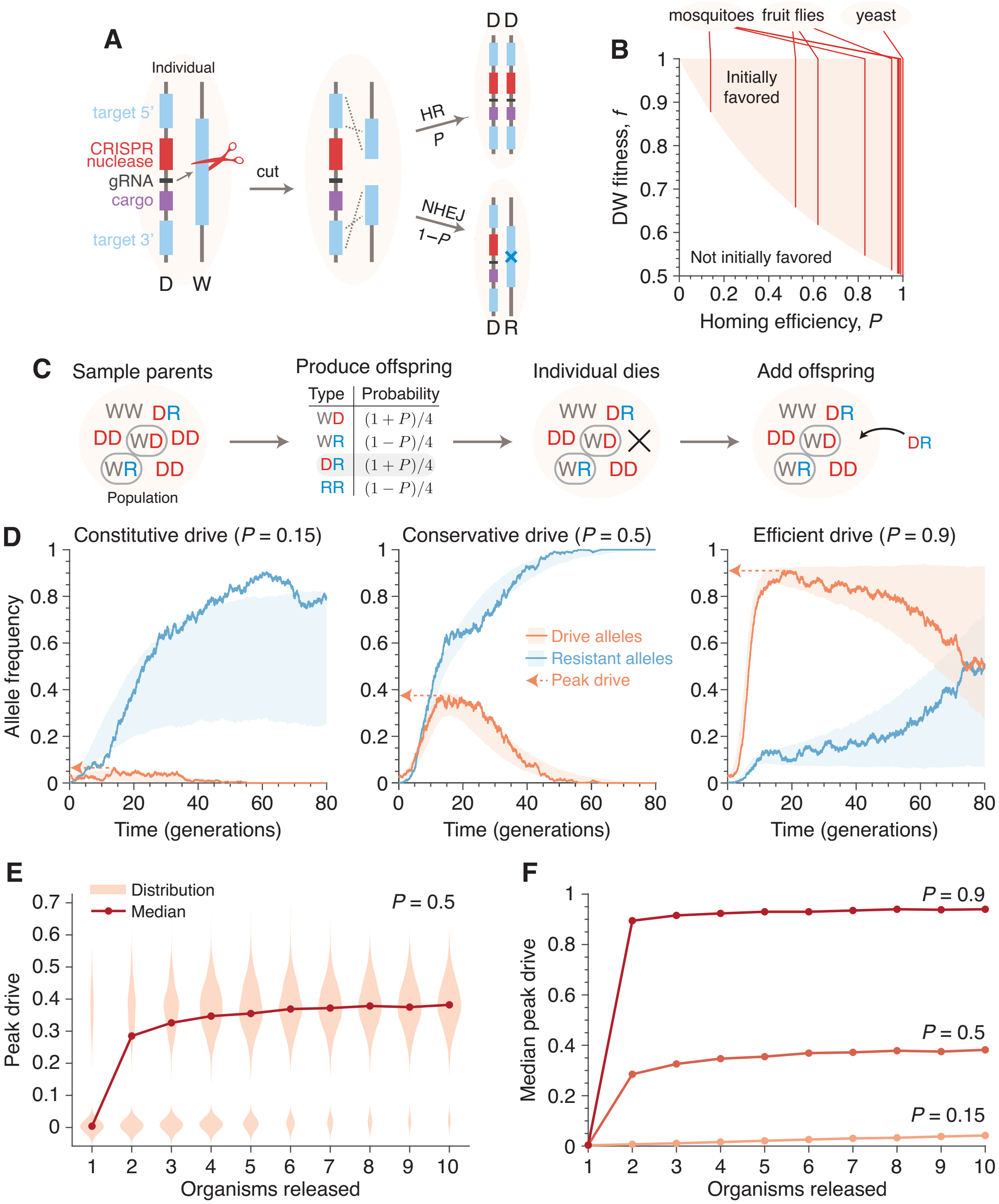
At the population level, our basic model considers N diploid individuals mating randomly. The process unfolds in discrete steps, during which parents are chosen for reproduction, an offspring is chosen according to the mechanism above, and another individual is replaced by the offspring (Fig. 2.1C and Section 2.5.1). These steps are repeated until one allele fixes. A generation is N timesteps, which corresponds to the mean lifespan of an individual.

Code to perform numerical simulations of this model and all model extensions described below (C++, Matlab), as well as data files, documentation, and code to reproduce all of the figures shown in this Chapter (Matlab) can be found at GitHub[77].

Figure 2.1D shows typical simulations for drive efficiencies of 0.15, 0.5, and 0.9, which correspond respectively to a constitutively active drive system targeting a common insertion site, and conservative and high efficiency systems (based on previous experimental studies, Table A.1, Fig. 2.1B,

Figure 2.1 ***(following page)*:** Existing alteration-type CRISPR gene drive systems should invade well-mixed wild populations. (A) Typical construction and function of alteration-type CRISPR gene drive systems. A drive construct (D), including a CRISPR nuclease, guide RNA (gRNA), and "cargo" sequence, induces cutting at a wild-type allele (W) with homology to sequences flanking the drive construct. Repair by homologous recombination (HR) results in conversion of the wild-type to a drive allele, or repair by nonhomologous end-joining (NHEJ) produces a drive-resistant allele (R). (B) Drives are predicted to invade by deterministic models when the fitness of DW heterozygotes, f, and the homing efficiency, P, are in the shaded region. Vertical lines indicate empirical efficiencies from Table A.1. (C) Diagram of a single step of the gene-drive Moran process. (D) Finite-population simulations of 15 drive individuals released into a wild population of size 500, assuming conservative ($P = 0.5$) or high ($P = 0.9$) homing efficiencies, as well as a low-efficiency, constitutively active system ($P = 0.15$). Individual sample simulations (solid lines), and 50% confidence intervals (shaded), calculated from 10^3 simulations. Drive-allele frequencies red and resistant-allele frequencies blue. Peak drive, or maximum frequency reached, is illustrated by dashed lines and arrows. (E) Peak drive distributions and medians with varying numbers of individual organisms released (P=0.5). (F) Medians of peak drive distributions for varying homing efficiencies ($P = 0.15$, bottom; $P = 0.5$, middle; $P = 0.9$, top). Throughout, we assume neutral resistance ($f_{WR} = f_{RR} = 1$) and a 10% dominant drive fitness cost ($f_{WD} = f_{DD} = f_{DR} = 0.9$).

Figure 2.1: (continued)



Appendix A). These simulations assume a dominant drive fitness cost of 10%, a population of size 500, and a release of 15 drive-homozygous individuals. (Note that the dynamics are similar for larger population sizes; see Section 2.5.4 and Fig. 2.3.) In all three cases, the drive, on average, irreversibly alters a majority of the population, either via invasion of the drive itself or via spread of drive-created resistant alleles. We call the maximum frequency of drive alleles reached during a simulation the peak drive, and we say a drive has invaded if it establishes in the population, ensuring behavior qualitatively similar to deterministic models (Section 2.5.9). Notably, for sufficiently large populations, arbitrarily low frequencies meet this standard, as it depends on the absolute number of drive alleles rather than their frequency (Section 2.5.10). Note also that each of these examples is chosen from the parameter regime in which invasion is predicted by deterministic models, since invasion is very unlikely outside of this regime.

We next calculated the distribution of peak drive while varying the number of organisms released (Figs. 2.1E and 2.1F). We find that these distributions are bimodal, with one mode centered around the initial frequency (corresponding to drift leading rapidly to extinction) and one centered roughly around the maximum values observed in the large-release scenarios in Fig. 2.1D. The former mode shrinks rapidly as more organisms are released, and for the parameters studied, a release of 10 individuals nearly guarantees invasion with substantial peak drive (Section 2.5.9, Fig. 2.10).

To understand the extent to which isolation might prevent invasion of other populations connected by gene flow, we introduced population structure. Our model consists of five subpopulations (or islands) that are equally connected by migration (Figs. 2.2A and Section 2.5.2). Typical dynamics are illustrated in Fig. 2.2C. Figures 2.2B and 2.2D show the escape probability, or the probability of the drive invading (arbitrarily defined as attaining a frequency of 0.1) at least one subpopulation other than its originating one, and Figure 2.2E shows the probability of invading a varying number of subpopulations.

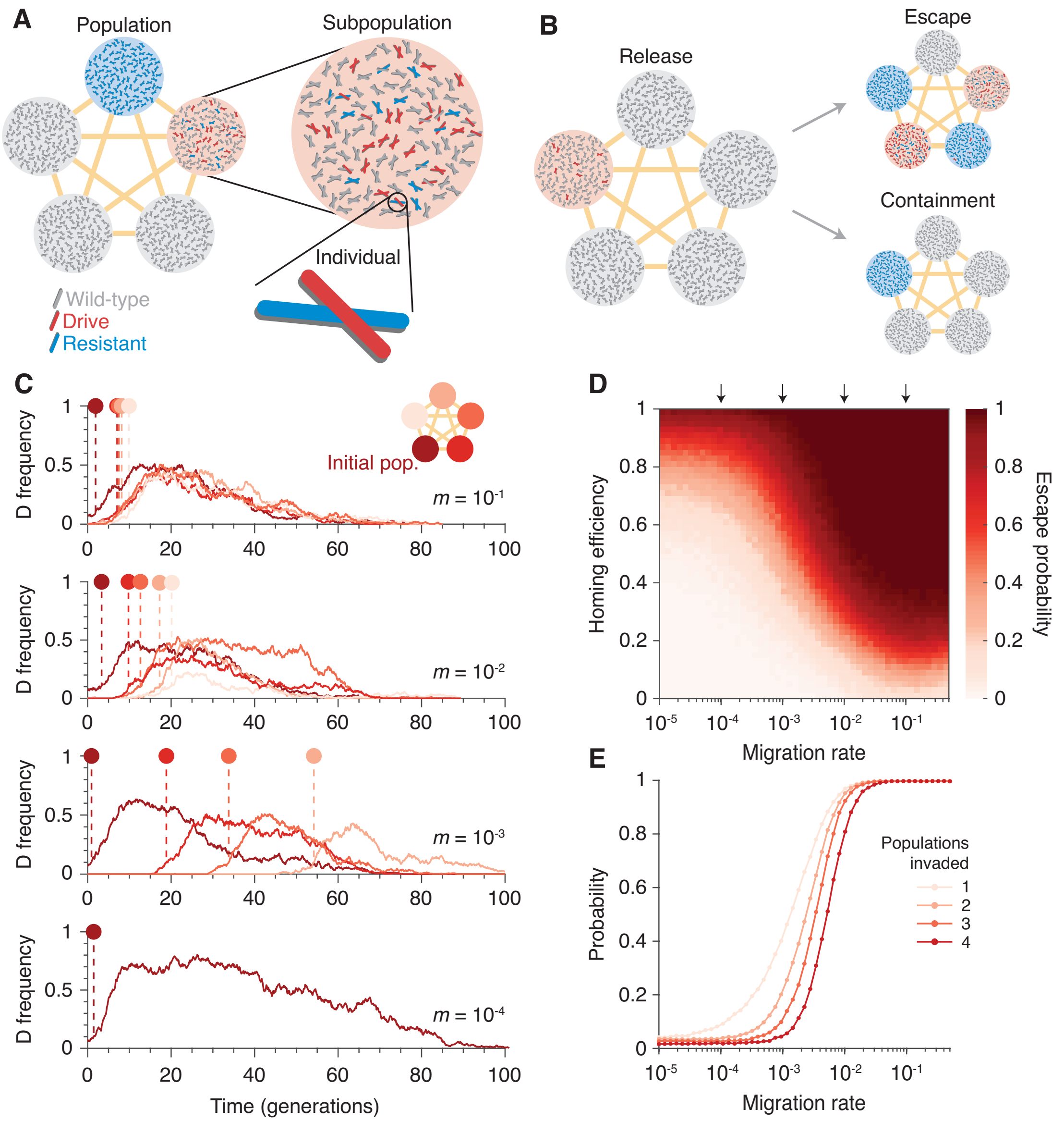
Our results in Fig. 2.2 suggest that if the migration rate is extremely low, then the drive is effectively contained in the initial subpopulation. If the migration rate is high, the drive is almost guaranteed to invade all subpopulations linked to the originating one. For intermediate migration rates—characterized roughly by migration rates on the order of the inverse of the drive extinction time—both outcomes occur. In the scenario studied in Fig. 2.2, a migration rate of 10^{-3}, which corresponds to a single migration event every 2 generations on average (Section 2.5.2), virtually guarantees escape for moderate drive efficiencies. For further details and analytical formulae allowing rapid estimation of escape probabilities, see Section 2.5.10.

Finally, we sought to understand the effects of additional mitigating factors that could potentially affect peak drive or invasion. We considered the most prominent factors that have arisen in previous papers, and we studied each by varying parameters in our basic model and developing model extensions. Our results are explored in detail in Section 2.5.

First, we considered preexisting drive resistance resulting from standing genetic variation[62,68] (Section 2.5.5). We find that increasing the proportion of the population that is initially resistant lin-

Figure 2.2 ***(following page)*****:** Existing CRISPR gene drive systems should invade linked subpopulations connected by gene flow. (A) Diagram of well-mixed subpopulations (circles) linked by gene flow (edges). Individuals represented by chromosomes with wild-type (gray), drive (red), or resistant (blue) haplotypes. (B) Few drive homozygotes are released in one subpopulation. The drive escapes if it invades another subpopulation before going extinct. Otherwise it is contained. (C) Typical simulations for varying migration rates ($m = 10^{-1}$, top, to $m = 10^{-4}$, bottom), following introduction into a single subpopulation. Lines represent drive frequencies in each subpopulation. Circles correspond to the time the drive invades a subpopulation. Population color is by invasion order, not predetermined. (D) Escape probability as a function of homing efficiency, P, and migration rate, m. Arrows indicate migration rates from B. Each pixel is calculated from 10^3 simulations. (E) Probability of invading 1, 2, 3, or 4 additional populations (aside from the originating population, which is typically invaded), assuming a homing efficiency of $P = 0.5$. Each data point is calculated from 10^4 simulations. Throughout, we consider 5 subpopulations connected in a complete graph, each consisting of 100 individuals. Initially, 15 drive homozygotes are introduced into one subpopulation. Resistance is neutral ($f_{WR} = f_{RR} = 1$) and the drive confers a dominant 10% cost ($f_{WD} = f_{DD} = f_{DR} = 0.9$).

Figure 2.2: (continued)



early decreases the mean peak drive ($R^2 = 0.996$). Using the parameters in Fig. 2.1E and considering a release of 15 individuals, more than 50% preexisting resistance is required to contain average peak drive below 10% (Fig. 2.4).

Second, we studied the effect of varying family size, which may be relevant to species such as mosquitoes with large egg batch sizes[13,78]. We extended the model so that k (adult) offspring are produced from a reproduction event, rather than one. We find that this effect scales the release and population sizes[79] by a factor of $4/(2k+6)$. For illustration, we estimated k for Anopheles gambiae to be roughly 10 (Section 2.5.6), so that a release of 7 individuals roughly corresponds to a release of 1 individual in our basic model. While this effect somewhat reduces the chance of drive invasion for small release sizes, it does not preclude it.

Third, we varied drive fitness, resistant-individual fitness and homing efficiency across their entire parameter regimes and recorded peak drive (Section 2.5.7, Fig. 2.7, Fig. 2.8). While varying drive fitness, we find that peak drive is on average greater than 30% across the majority of the regime and almost always greater than 10% (Fig. 2.7, left)—and, as a technical aside, we find that this is the case whether the fitness cost of the drive manifests itself via a reduction in birth rate or via increase in death rate (Fig. 2.7, right). Moreover, in line with previous deterministic results, we find that peak drive can be substantially increased by associating a fitness cost with resistance (Fig. 2.8), which could be expected for drive constructs intended for large-scale application, utilizing methods such as multiplex targeting of essential genes[39,41,69].

Fourth, we studied the effect of inbreeding, which has been shown in several recent theoretical studies[63,68] to impede drive spread (Section 2.5.8). We extended the model to include a probability

s of an individual selfing rather than mating with a second individual[63]. The model assumes no inbreeding depression and thus considers the worst-case scenario for drive[63]. We find that even in this scenario, high selfing probabilities are required to reduce peak drive and the probability of invasion for moderate drive costs.

There are a variety of other phenomena that could affect invasiveness, e.g., density dependence[52], environment[80], costly resistance[81], local ecology, and even mating incompatibilities between some laboratory strains and wild individuals. Such effects should be carefully studied in subsequent papers. Most importantly, the drive architecture itself should affect invasiveness; we consider here only alteration-type drive systems, while others, e.g., sex-ratio distorters and genetic load drives, would be expected to yield different dynamics. In particular, population suppression drive systems may locally self-extinguish before invading new populations. However, for alteration drives, our key qualitative finding—that peak drive is difficult to reliably contain below a socially tolerable threshold following a very small release of organisms—appears robust to a variety of mitigating factors. Fundamentally, we exercise caution by omitting application-specific phenomena that might aid containment in particular instances but not in general.

2.4 Discussion

Our results suggest that current first-generation CRISPR-based gene drive systems for population alteration are capable of far-reaching—perhaps, for species distributed worldwide, global—spread, even for very small releases. A simple, constitutively expressed CRISPR nuclease and guide RNA cassette targeting the neutral site of insertion—an arrangement that could occur accidentally—may be capable of altering many populations of the target species depending on the homing efficiency of the organism in question. More generally, resistance can be problematic for intentional applications of gene drives, but we find that it is not a major impediment to invasion of unintended populations.

These findings raise two important questions: (1) How likely are unauthorized releases of self-propagating gene drive systems in the first place? (2) How likely are serious negative consequences given the apparently high likelihood of spread to most populations of the target species? Rigorously addressing these questions is an important direction for future work, and we can offer only opinions here. The answer to the first question likely depends on a large number of factors, such as species, application, containment strategies, economic motivations, drive development stages, geography, and the caution of the investigators, so we omit speculation here. However, we consider the answer to the second question to be clearer: although most laboratory gene drive systems are unlikely to cause ecological changes—they are typically predicted to be transient and are not designed to alter traits of the host organism, least of all interactions with other species—the history of genetic engineering offers many examples suggesting that substantial social backlash could be triggered by unauthorized spread of a self-propagating gene drive[82,83]. Any such event could significantly reduce public support for interventions against diseases such as malaria that could possibly save millions of lives. We believe it would be profoundly unwise to proceed with anything less than an abundance of caution.

On a more technical note, our findings are specific to population alteration drive and cannot be directly generalized to self-propagating suppression drive, which could potentially self-extinguish before invading other populations. However, our results suggest a method for rough comparison between these scenarios: we find that the primary factor in determining drive spread between adjacent populations is the average number of migrants per generation (Section 2.5.10), which can, in principle, be compared between models. For example, an earlier model of suppression drive systems[52] predicted a total number of drive-carrying organisms over time which is remarkably similar to our example of an inefficient alteration drive system that is rapidly outcompeted by resistant alleles (Fig. 2.1D, middle). Thus, assuming comparable migration rates, it might not be surprising to see qualitatively similar levels of invasiveness. Accordingly, we urge researchers to exercise caution in developing or advocating for self-propagating suppression drives for applications other than malaria prevention—or similar projects intended to affect an entire species—until explicit models of invasiveness are available.

Additionally, our findings emphasize the importance of the containment strategy known as "ecological confinement", which was proposed previously[41,42]. Given the risk that organisms may escape through accidents or outside intervention, laboratories in regions with endemic wild populations may wish to refrain from constructing self-propagating systems capable of invading those populations and undergoing unwanted spread. Laboratories in regions with endemic wild populations can reliably prevent accidental invasion by employing intrinsic molecular confinement mechanisms such as synthetic site targeting or split drive as recommended by the National Academies' report on gene drives[66].

Perhaps most importantly, any development efforts looking ahead toward field trials, a component of the staged testing strategy outlined by the National Academies report, should be aware that there could be a high likelihood of unwanted spread across international borders, even from ostensibly isolated islands. The development of 'local', intrinsically self-exhausting gene drive systems[26,27,65,84,85], sensitive methods of monitoring population genetics, and strategies for countering self-propagating drive systems and removing all engineered genes from wild populations should be correspondingly high priorities.

2.5 Supplementary model details and extensions

2.5.1 Well-mixed finite population model

To model gene drives in finite populations, we introduce a Moran-type model with sexual reproduction (illustrated in Fig. 2.1C). We consider a population of N individuals, each of which is diploid. We focus on a locus with three allelic classes: wild-type (W), CRISPR gene drive element (D) and drive-resistant (R). There are six possible genotypes: WW, WD, WR, DD, DR, and RR. We assign to each genotype α a reproductive rate f_α.

The process proceeds in discrete time-steps, during each of which three events occur in succession (Fig. 2.1C). First, two individuals are chosen without replacement for mating with probabilities proportional to their reproductive rates, so that genotype α is selected with probability

$$\frac{f_\alpha N_\alpha}{\sum_\beta f_\beta N_\beta}. \tag{2.1}$$

Here N_α is the number of individuals having genotype α, and the sum in the denominator is over all six genotypes. Second, after selecting the two parents, the offspring genotype is chosen randomly based on the genotypes of the two parents. To proceed, we introduce notation $\alpha = AB$ to mean that genotype α consists of alleles A and B, and we index these alleles via $\alpha_1 = A$ and $\alpha_2 = B$. Note that we track only one genotype for each heterozygote, implicitly combining counts for genotypes AB and BA. Using this notation, the probability that an offspring of genotype γ is chosen given a mating between parents of genotypes α and β is given by the quantity $q^{\gamma}_{\alpha\beta}$, which is equal to

$$\frac{q^{\gamma_1}_{\alpha} q^{\gamma_2}_{\beta} + q^{\gamma_2}_{\alpha} q^{\gamma_1}_{\beta}}{1 + \delta_{\gamma_1\gamma_2}}. \tag{2.2}$$

Here q^{A}_{α} is a gamete production probability—the probability that a parent with genotype α produces a gamete with haplotype A—and δ_{AB} is the Kronecker delta, defined by $\delta_{AB} = 1$ if $A = B$ (i.e., if the offspring under consideration is a homozygote), and $\delta_{AB} = 0$ otherwise. The gamete production probabilities, q^{A}_{α}, are determined by accounting for the gene drive process described above. They are given by: $q^{W}_{WW} = q^{D}_{DD} = q^{R}_{RR} = 1$, $q^{D}_{WD} = (1 + P)/2$, $q^{R}_{WD} = (1 - P)/2$, $q^{W}_{WR} = q^{R}_{WR} = q^{D}_{DR} = q^{R}_{DR} = 1/2$. The remaining values not listed, e.g., q^{R}_{WW}, are zero. Third, an individual is chosen uniformly at random for death. Thus, the population size remains constant. The resulting counts become the starting abundances for the next iteration of the process. The process is initialized with a small number, i, of drive homozygotes (DD) and the remaining population, $N - i$, wild-type homozygotes (WW). The process continues as described above either until a specified number of time steps have elapsed or until one of the three alleles has fixed. Any of the alleles can fix, but typically either the wild-type or resistant alleles fix, due to the emergence of resistance.

2.5.2 Finite population model with population structure

To study the effects of population structure on drive containment, we extended the well-mixed model from the previous section. We now consider l well-mixed subpopulations, each consisting initially of N/l individuals. The process proceeds in discrete time steps, as before. In each time step, we either migrate an individual from one population to another, or we choose a particular subpopulation and proceed through one mating and replacement iteration, as outlined above. More specifically, one step of the process proceeds as follows (illustrated in Fig. 2.11). With probability m, we initiate a migration event. In this case, we perform three steps. First, we choose a source population with probability proportional to its size. Second, we choose an individual uniformly at random from the source population for migration. Finally, we move the chosen individual to a linked subpopulation uniformly at random. Or, with probability $1 - m$, we initiate a mating event as described in the well-mixed section. To carry this out, we first choose the population in which the event will occur. We choose this population with probability proportional to the square of its total fitness, since this counts the rate of reproduction for every possible mating pair in the population (as matings occur with rates proportional to the fitness of each parent). We then step through one iteration of the well-mixed mating process within this subpopulation. Note that in this model the migration rate has a simple interpretation. The time between migrations is geometrically distributed with parameter m, so the mean time between migrations is $1/m$ time steps. Recall that a "generation" is equal to the mean lifespan of an individual, that is, N reproduction events or $N/(1 - m)$

time steps. Then the typical time between migrations can be expressed with the units as generations:

$$\mathbb{E}[T] = \frac{1-m}{Nm}. \tag{2.3}$$

2.5.3 Deterministic model

To compare our stochastic simulations with deterministic results, we use the model from Chapter 1, described in Section 1.5 and published in Ref. 39. In particular, we use the "previous drive" model, as it was designed to agree with the existing proof-of-concept CRISPR drive constructs that we consider here. Specifically, we consider the case of 1 guide RNA ($n = 1$ in that model's notation), and zero production of costly resistant alleles ($\gamma = 1$).

2.5.4 Population size

Above, we present results from simulations which assume populations of size $N = 500$. We claim that $N = 500$ is a reasonable approximation for the dynamics in the large-population limit, which is the relevant regime for widespread invasion or for species with very large population sizes, e.g., mosquitoes. Here we briefly evaluate this claim.

Figure 2.3 recreates Figure 2.1E with additional population sizes overlaid: $N = 1000, 2500$, 5000, and 10000. The distributions narrow for larger N until plateauing at roughly $N = 5000$. However, the central tendencies show little change with increasing N.

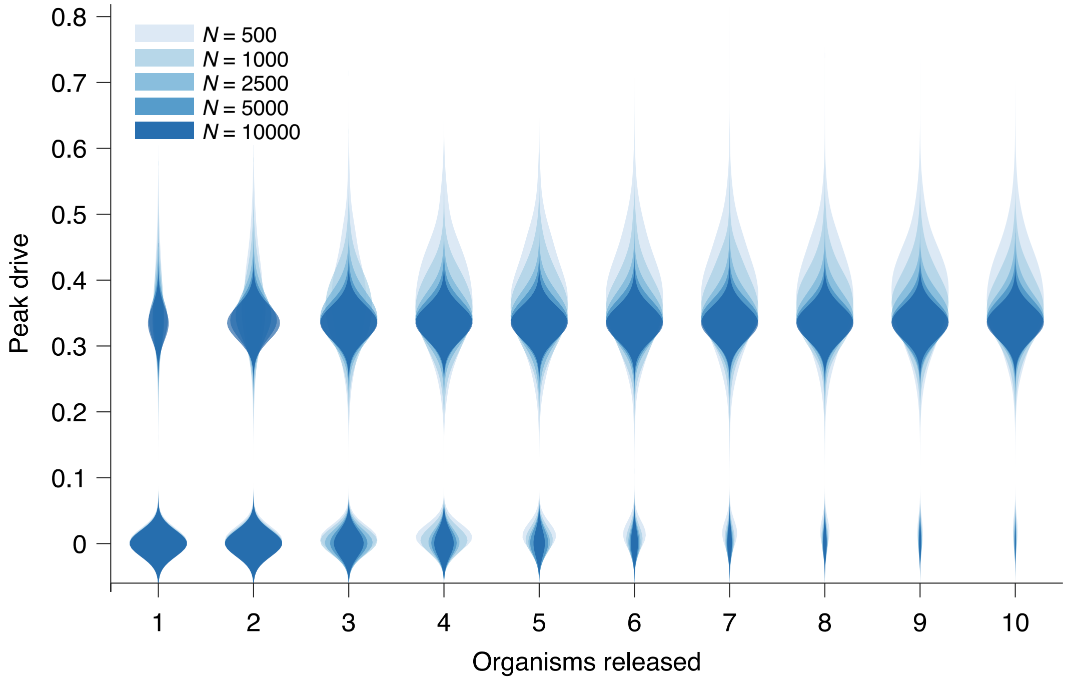


Figure 2.3: Peak drive distributions for variable release and population sizes. Parameters are chosen to correspond to Fig. 2.1E: $P = 0.5, f = 0.9$ and neutral resistance. Population sizes are, from light to dark, $N = 500, 1000, 2500, 5000, 10000$. Note that $N = 500$ corresponds exactly to Fig. 2.1E. Each distribution corresponds to 10^3 simulations.

2.5.5 Standing genetic variation

Several recent studies have explored the effect of pre-existing drive resistant alleles in a population brought about by standing genetic variation (SGV) at the target locus[62,68]. These studies developed deterministic models and showed that pre-existing resistant alleles—presumably neutral—should rapidly outcompete costly drives due to selection, resulting in rapid drive extinction. The study by Drury *et al.*[68] used sequencing to quantify this standing variation in diverse populations of flour beetles and found resistance-conferring mutations to exist at a wide range of frequencies, from 0 to

0.375, with an average of roughly 0.1.

However, these studies were primarily concerned with long-term outcomes following drive release, in which case resistance certainly outcompetes the drive. For our purposes, however, we are concerned with the intermediate time regime in which the dynamics of resistance are less clear. Moreover, these studies employed deterministic models, whereas our model is stochastic. Here, we seek to understand the effect of SGV in our model.

To incorporate SGV, we simply alter the initial conditions: rather than introducing i drive homozygotes into a population of $N - i$ wild-type homozygotes, we introduce i drive homozygotes into a population consisting of j resistant homozygotes (we choose resistant homozygotes for simplicity, since they rapidly go to Hardy-Weinberg equilibrium following release) and $N - i - j$ wild-type homozygotes. Figure 2.4 shows the effect of SGV on peak drive for pre-existing resistance frequencies up to 0.5.

We find that the effect of SGV is to linearly decrease the mean peak drive ($R^2 = 0.996$). Our intuition for this result is as follows. Because the population is well-mixed, the effect of resistance is simply to decrease the size of the population that is susceptible to the effects of the drive. This can be roughly viewed as linearly scaling the drive-frequency axis. For example, if the population has a 0.1 frequency of resistant alleles immediately prior to release, then the population that is susceptible to drive is roughly 90% of the census population size, and the drive undergoes its usual dynamics within this subpopulation. There are of course complications to this simplistic explanation, e.g., selection increasing the size of the resistant population and diploidy mixing resistant and drive alleles. Furthermore, the linear relationship only holds for sufficiently low levels of SGV. In our example here, the relationship holds to roughly 0.5 initial resistance frequency. However, this is still higher than would be anticipated for drives engineered to spread in the wild.

Overall, our results suggest that a high level of SGV would be required to protect against drive invasion. In our conservative example (Fig. 2.4) assuming 0.5 homing efficiency, 0.9 drive fitness, and neutral resistance, pre-existing resistance of greater than 0.5 frequency is required to contain peak drive to below 10% of the population, compared to 35% in the absence of SGV.

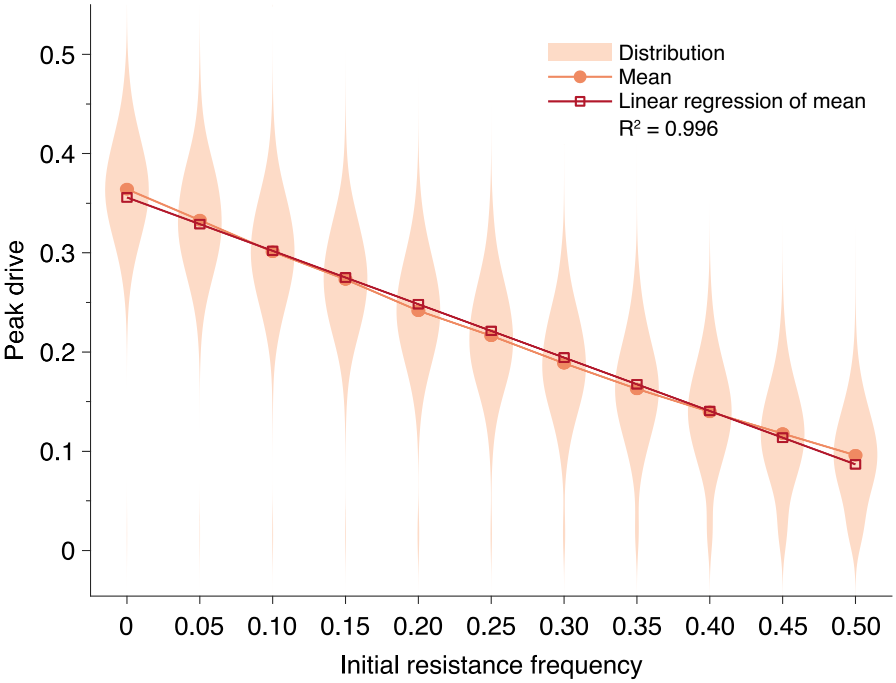


Figure 2.4: Pre-existing drive-resistant allele frequency linearly decreases peak drive. Distributions (violin plots), means (orange, circles) and linear regression of the mean values (red, squares). Parameters are chosen to correspond to Fig. 2.1E: $P = 0.5$, $f = 0.9$, neutral resistance, $N = 500$. Each distribution corresponds to 5000 simulations.

2.5.6 Offspring number distribution

In the model presented above, we assume that each mating produces one offspring. However, a variety of application-relevant species are known to produce many offspring per mating. For example, female *Anopheles gambiae* mosquitoes can lay hundreds of eggs per lifetime[13]. It is not clear, *a priori*, how varying the offspring number distribution in our model would affect the results presented above. Thus we here analyze a simple extension of the model which allows us to vary the number of offspring following a given mating event.

To begin, recall our model. We consider a population of constant size N with the following process: At each time-step, two individuals are chosen for mating; an offspring is sampled according to the parental genotypes; a third individual is chosen for removal from the population, and the parents' offspring takes its place. (We implicitly assume that these offspring are only the offspring which successfully reach adulthood, i.e., reproductive age.) We now add a new parameter, k, which determines number of (adult) offspring produced by a mating pair. The process proceeds as before, except now k offspring are independently sampled from the parental genotypes, and k individuals are chosen uniformly (without replacement) for removal from the population. Clearly the model presented in the main text is the special case $k = 1$.

Note that this parameter k is not equivalent to brood size, clutch size, egg batch size, etc.—values often considered in the ecological literature—in that k describes the number of offspring produced per mating which successfully attain reproductive age. This number can of course be much lower than these other parameters due to death during juvenile life stages. We provide an example calculation for this parameter in *An. gambiae* at the end of this section.

We now argue that increasing the number of offspring per mating, k, corresponds to decreasing the effective size of the population, N_e. We omit rigorous proof here, but we provide a formula for the effective population size in our model and present numerical simulations as support. To begin, Hill showed in 1972 that the variance effective population size in the standard Moran model is[79]

$$N_e = \frac{4N}{2 + \sigma_X^2}. \tag{2.4}$$

Here N is the census population size, and σ_X^2 is the variance in the distribution of the total number of offspring produced by an individual over the course of its lifetime (*i.e.*, its lifetime reproductive success). It was proven that this formula holds both for the Wright-Fisher model with discrete generations and for the Moran model with overlapping generations, provided that σ_X^2 is the same and that the total number of individuals entering the population in each generation is equal[79]. Our model meets both of these requirements—indeed, the only difference is that two parents are chosen to sample offspring types, rather than one, and this has no bearing on the number of offspring produced—so we conjecture that Eq. (2.4) holds for our case as well.

To proceed, we calculate σ_X^2 for our extended model and employ the variance effective population size given by Eq. (2.4). Consider one particular individual in the population, and let $t = 1, 2, \ldots$ count time-steps. As described, in each step, k individuals are uniformly sampled (without replacement) for removal. Thus, an individual has probability k/N of dying in each step. Its lifespan, T, is thus geometrically distributed, $T \sim \text{Geometric}(k/N)$.

Next, let X be a random variable describing the number of offspring an individual produces in its lifetime, so that $X|T$ is the number of such events given that the individual survives T time-steps. Because each mating event is independent, $(X|T) \sim k \cdot \text{Bin}(T, 2/N)$. The success probability derives from the fact that two individuals are chosen for mating in each time-step and that the process is neutral. Thus,

$$\mathbb{E}X = \mathbb{E}\mathbb{E}\left[X \mid T\right] = \mathbb{E}k(2/N)T = k(2/N)N/k = 2$$

and

$$\begin{aligned}\texttt{Var}(X) &= \mathbb{E}\texttt{Var}(X \mid T) + \texttt{Var}(\mathbb{E}(X \mid T)) \\ &= \mathbb{E}k^2T(2/N)(1-2/N) + \texttt{Var}(k(2/N)T) \\ &= kN(2/N)(1-2/N) + (2k/N)^2N(N-k)/k^2 \\ &= 4 + 2k(N-4)/N.\end{aligned}$$

Returning to the variance effective population size expression in Eq. (2.4), we obtain for our model:

$$N_e = \frac{4N}{2k+6}. \tag{2.5}$$

Note that in the case $k = 1$ we recover $N_e = N/2$, which is the variance effective population size for the standard Moran model.

In Fig. 2.5, we present peak drive distributions (as in Figs. 2.1E and 2.3) for varying values of k with the effective population size, N_e, and effective release size, i_e, both determined by Eq. (2.5), held constant. In this case we used $N_e = 250$ and $i_e = 8$, which correspond to $N = 500$ and an initial release of $i = 16$ in our standard model with $k = 1$. The peak drive distributions for all values of k studied are approximately identical. This suggests that the dynamics for larger k can indeed be inferred from the standard model with $k = 1$ and population/release sizes appropriately scaled via Eq. (2.5). An immediate consequence of this result is that releases of organisms which have many offspring (e.g., mosquitoes) are effectively smaller than would be expected from simply counting. For example, an organism which typically has 100 offspring that survive to adulthood would need a release size of roughly 258 to surpass the 10-individual initial release threshold we have observed. Note that the 10-individual threshold discussed throughout the text is the census release size; the effective release size is $i_e = 5$.

In Fig. 2.6, we recalculate the distributions in Fig. 2.5 holding the *actual* population and release sizes constant, rather than their effective values. Two effects are apparent. First, the decrease in effective population size, N_e, leads to greater variation in peak drive among simulations that invade, *i.e.*, the distribution centered around ≈ 0.4 widens. Second, the decrease in effective release size, i_e, leads to a greater probability of simulations immediately going extinct, *i.e.*, the relative mass of the mode centered around ≈ 0 increases. In sufficiently large populations the first effect would be less pronounced—see Fig. 2.3—while the second effect should apply for any small release.

Finally, as an example, we provide an estimate of our model's k parameter for a particularly relevant species, *An. gambiae*. To do this, we find the typical size, n, of egg batches laid by females following a particular mating event; then we estimate the total number of these which survive to adulthood using parameters from the literature.

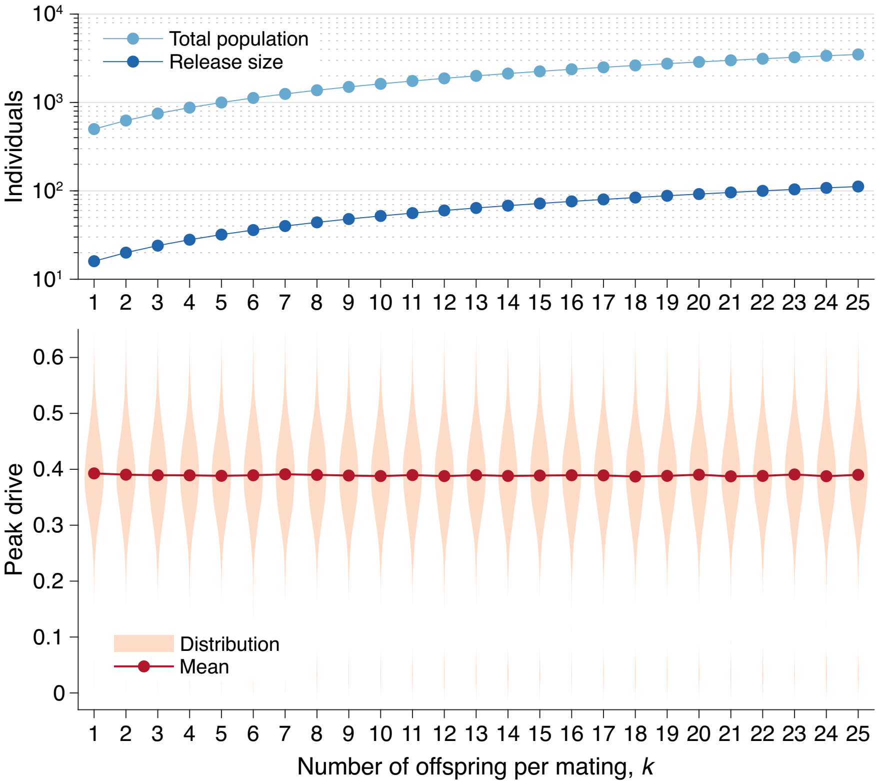


Figure 2.5: Peak drive distributions for varying numbers of offspring per mating with effective population and release sizes held constant. (top) Population and release sizes used in the simulations below. For the case $k = 1$, we use our usual population size of $N = 500$ with an initial release of $i = 16$ drive homozygotes. According to Eq. (2.5), the effective total population and release sizes in this case are $N_e = 250$ and $i_e = 8$. For other values of k, we use values of N and i which maintain constant effective population and release sizes: $N = N_e(2k+6)/4$ and $i = i_e(2k+6)/4$. These values are plotted: N (light blue) and i (dark blue). (bottom) Peak drive distributions assuming values of N and i as in the above plot. All employ $P = 0.5$, $f = 0.9$, and neutral resistance. Each distribution includes 5000 simulations.

The first number, n, varies according to a variety of environmental and ecological factors[13,78], so we assume a large but reasonable value in order to avoid underestimating our parameter k. For this, we assume that $n \approx 186$, which is roughly the highest value observed by Hammond *et al.* in the CRISPR drive study[13] and is in line with previous field work[78].

To estimate the survival probability for each egg to adulthood, we employ the method and pa-

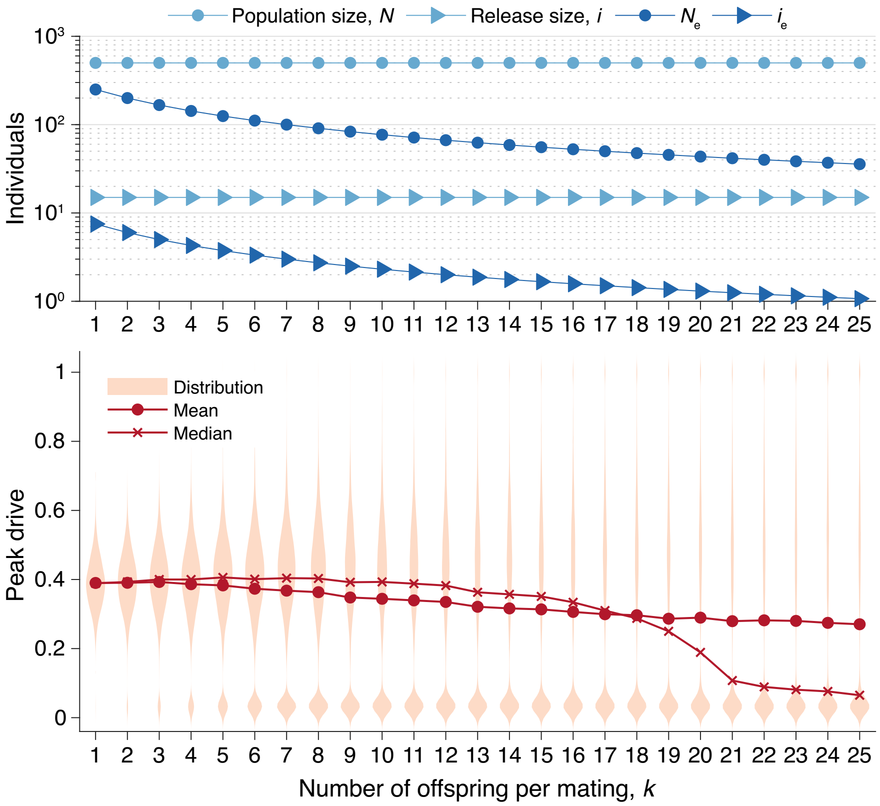


Figure 2.6: Peak drive distributions for varying numbers of offspring per mating with census population and actual release sizes held constant. (top) Population and release sizes used in the simulations below. Actual population size, N (light blue, circles) and actual release size, i (light blue, triangles). Note that $N = 500$ and $i = 15$ are constant. Effective values calculated via Eq. (2.5): population size, N_e (dark blue, circles) and release size, i_e (dark blue, triangles). (bottom) Peak drive distributions for simulations using indicated values of k and population and release sizes as depicted above. Compare with Fig. 2.5 which holds the *effective* population and release sizes constant, whereas here we hold the *census* population and release sizes constant. All simulations employ $P = 0.5, f = 0.9$, and neutral resistance. Each distribution includes 5000 simulations.

rameters presented by Deredec *et al.*[52] Each egg goes through three juvenile stages before reaching adulthood—the egg stage, the larva stage, and the pupae stage. We denote the probabilities of surviving each of these stages by θ_0, θ_L, and θ_P, respectively. The probability of a particular egg reaching adulthood is then $p = \theta_0\theta_L\theta_P$. These parameters were estimated to be $\theta_0 = 0.831, \theta_L = 0.076$, and $\theta_P = 0.831$. Thus we have $p = 0.0525$.

Given this formulation, the number of eggs laid per mating event which reach adulthood is distributed according to $\text{Bin}(n, p)$. We take the mean of this distribution to obtain:

$$k \approx np = 9.76.$$

Therefore, while *An. gambiae* females exhibit large egg batch sizes, the value of k for our model is much lower—indeed, low enough that the central tendency of the peak drive distribution remains roughly unchanged in Fig. 2.6.

2.5.7 Effect of varying fitness and homing efficiency

Above, we study various values of the homing efficiency, P, but we perform less exploration of the parameters governing drive fitness, f, and resistance cost, s. This is motivated primarily by the abundance of data for the former—see Table A.1—and the lack of data for the latter parameters.

In addition, we have assumed throughout that death rates are identical for the various genotypes, while reproductive events occur with probabilities proportional to fitness. On the other hand, some drive constructs might behave the opposite way: reducing fitness by increasing an organism's death rate, while leaving its birth rate unchanged.

In this section we explore these three effects: (i) varying drive fitness across its entire range, (ii) varying the fitness cost of resistance across its entire range, and (iii) modifying the model so that death rates are affected by fitness, rather than birth rates.

To begin, we consider our standard model for fitness and study drive spread across the entire range of values for drive fitness, f, and homing efficiency, P. In particular, we consider 51 values of each parameter: $P \in [0, 1]$ and $f \in [0.5, 1]$, both evenly spaced, for a total of 2,601 parameter pairs. For each pair, the average peak drive is calculated over 100 simulations, and the results are shown in Fig. 2.7, left.

We find that maximum drive frequencies of greater than 0.3 are common across a wide range of drive fitness values. In particular, for our lower-bound estimate of empirical drive efficiency ($P = 0.5$), drives can confer fitness costs as high as 20% before the peak drive drops below 0.3. For more typical empirical efficiencies ($P > 0.8$), the peak drive is typically greater than 0.5 even for costly drives ($f \approx 0.7$), and low-cost drives ($f > 0.9$) have peak drive of greater than 0.9.

We next modified our standard well-mixed model in the following way. Recall that the model involves choosing two parents to mate, then choosing an individual to die and be replaced by the parents' offspring. In our standard model, the two parents are chosen to reproduce with probabilities proportional to their fitnesses, and an individual is chosen to die uniformly. In our modified model, we choose the two parents uniformly and then choose the individual to die with probability proportional to the inverse of its fitness. Results from the modified model are shown in Fig. 2.7, right and are nearly identical to the results from the standard model.

In both cases, it is important to note that the peak drive and likelihood of invasion deemed socially acceptable for accidental release would likely be lower than those discussed above. With this in mind, our simulations suggest that if a drive is predicted to invade by deterministic models (*i.e.*, if it lies above the boundary in Fig. 2.7), then it will almost certainly reach a maximum frequency greater than 0.1. While acceptable levels of peak drive are as-yet unknown and will likely vary between species, applications, jurisdictions and so on, spread to this extent will likely surpass it.

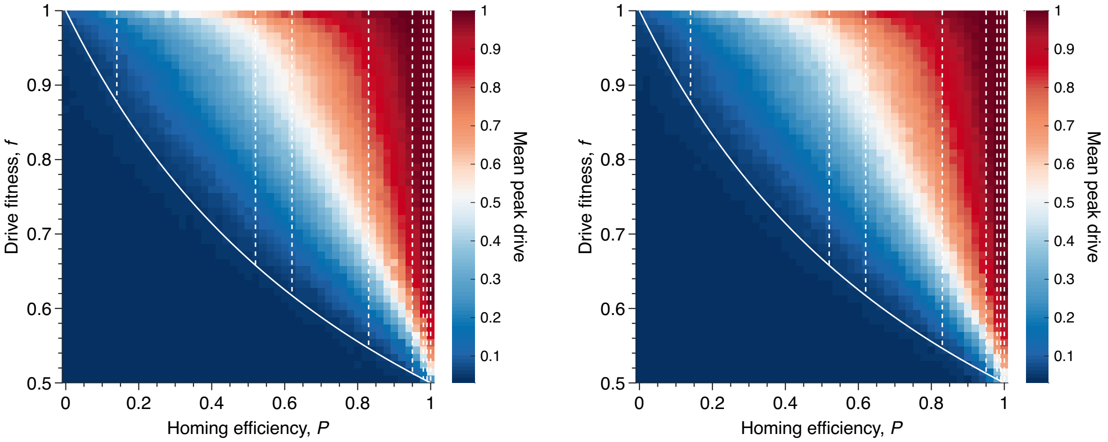


Figure 2.7: Mean peak drive for varying homing efficiency, P, and drive-individual fitness values, f (*i.e.*, individuals with genotypes WD, DD, and DR), assuming that fitness affects birth rate (left) or death rate (right). The left panel corresponds to our standard model, shown in Fig. 2.1C, while the right panel represents a modification: parents are chosen uniformly, and individuals die with probability proportional to the inverse of their fitness. The solid white line shows the boundary from Fig. 2.1B indicating whether the drive is predicted to invade by deterministic models. The drive is only expected to invade based on deterministic models if the fitness/homing efficiency pair lie above the boundary. The dashed white lines indicate the empirically measured homing efficiencies from Table A.1 and Fig. 2.1B. Each point in the grid (51×51) depicts an average of 100 simulations. Parameters used include a population size of 500, with an initial release of 15 drive homozygotes to ensure that trajectories establish. Neutral resistance is assumed throughout with no standing genetic variation.

Finally, we sought to understand the effect of varying the fitness cost associated with drive-resistance. Throughout the text above we have assumed that resistance is neutral, as this presumably represents the best case for containment. However, drive constructs developed for applications are likely to employ resistance-mitigating strategies, such as multiplex targeting of essential genes[39,41], which essentially increase the fitness cost associated with drive-resistance. Thus, we ran simulations varying drive-individual fitness, f, in the range $f \in [0.5, 1]$, and resistant-individual (RR) fitness in the range $[0, 1]$, assuming conservative drive efficiency, $P = 0.5$. In both dimensions we considered 51 parameter values, evenly spaced, for a total of 2,601 parameter pairs. For each pair, the average peak drive is calculated over 100 simulations, and the results are shown in Fig. 2.8.

We find qualitatively that there are two regimes, determined by the fitness cost of resistance, s (i.e., individuals with genotype RR have fitness $1 - s$), and the deterministic invasion condition, $f(1 + P) > 1$. In the figure, we assume that $P = 1/2$, so the deterministic invasion condition is simply $f > 2/3$. When the fitness cost of resistance, s, is sufficiently low ($s < 1/3$), then the dynamics are determined by the relationship between the fitness of drive individuals and the fitness of resistant individuals: if the fitness of drive individuals is greater than the fitness of resistant individuals, then the spread of the drive is dramatically improved—typically reaching fixation—compared to the baseline neutral-resistance case. However, if the fitness cost of resistance is sufficiently high ($s > 1/3$), then the improvement in drive spread brought about by increasing the cost of resistance saturates, since the drive can now be less costly than resistance ($f > 1 - s$) but also too costly to invade ($f < 2/3$). That is, for resistance costs higher than $1/3$, the mean peak drive as a function of drive fitness, f, remains essentially unchanged with increasing s, since the deterministic invasion condition can no longer be satisfied when the drive has fitness $f < 2/3$, no matter the cost of resistance.

2.5.8 Inbreeding

Since the drive functions only in heterozygotes, inbreeding in a population—which in effect reduces the frequency of heterozygotes—would be expected to impact drive invasiveness. Indeed, this has been shown in recent theoretical studies by Bull[63] and Drury *et al.*[68] Thus we here extend our well-mixed model to include inbreeding and study its effect.

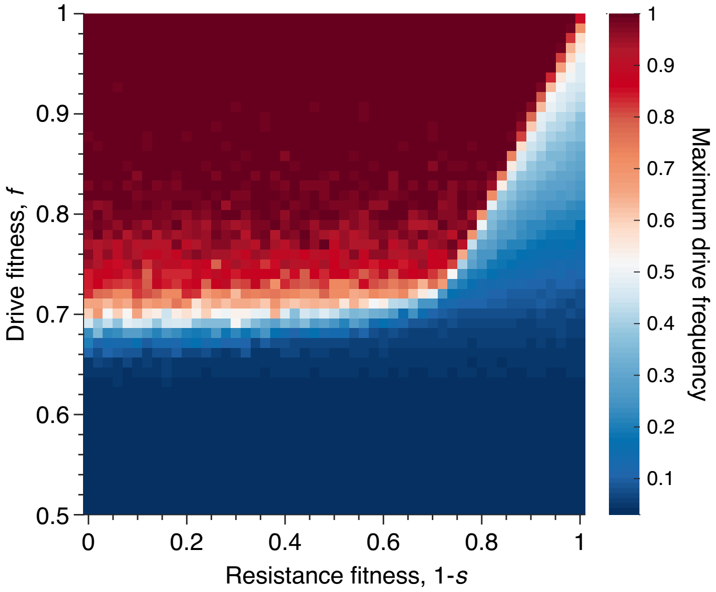


Figure 2.8: Mean peak drive for varying drive-individual fitness values, f, and resistant-individual (RR) fitness values, $1-s$, where s is the cost associated with resistance. Each point in the grid (51×51) depicts an average of 100 simulations. Parameters used include homing efficiency $P = 0.5$, population size of 500, with an initial release of 15 drive homozygotes to ensure that trajectories establish. Throughout we assume no standing genetic variation (i.e., the initial frequency of the resistant allele is 0).

For simplicity, we consider a partial selfing model. In each update step of our process (see Fig. 2.1C), we typically choose two parents for mating with probabilities proportional to their fitnesses. To include selfing, we instead choose the first parent as usual, with probability proportional to its fitness. We then choose the first parent as the second parent as well with probability s; or, with probability $1-s$, we choose a second parent from the remaining population, with probability proportional to its fitness. Note that the fitness of each offspring is determined entirely by its genotype and does not account for inbreeding depression. Implicitly, we thus consider the case of zero inbreeding depression. As this effect helps protect against drive invasion, we essentially consider the worst-case scenario for drive containment[63].

Using our extended model, we then computed peak drive distributions for values of s between 0 and 1 and for the three values of P explored above: $P = 0.15, 0.5, 0.9$. The results are shown in Fig. 2.9. We find that a fairly high degree of selfing is required to impact the peak drive distribution in a meaningful way. For highly effective drive, $P = 0.9$, the mass of the upper mode in the frequency distribution is larger than the lower mode until roughly $s \approx 0.75$. For conservative drive, $P = 0.5$, this occurs at roughly $s \approx 0.6$, and for ineffective drive there is little change, as the maximum frequency begins very near zero. To compare with previous results, we can consider the inbreeding coefficient rather than the selfing probability. In our model, the inbreeding coefficient, F, is given by $s/(2 - s)$. Thus highly effective drive can tolerate inbreeding of $F \approx 0.6$ and conservative drive can tolerate $F \approx 0.43$.

2.5.9 Comparison with deterministic model

To show that the deterministic ODE solutions provide reasonable approximations to the typical behavior of our stochastic mode, we overlay numerical solutions to the ODEs for the systems studied in Fig. 2.1D of the main text. The results are shown in Fig. 2.10.

Throughout we have assumed that resistance is neutral with respect to the wild-type. This assumption is biologically realizable as resistance is conferred by changing sequence homology to the drive's gRNA—something that could be achieved with synonymous codon substitutions, for example. In practice, some resistance mutations could be costly and those that are neutral could be rare. However, assuming resistance is always neutral represents the worst-case scenario for drive invasiveness, as resistance can increase in frequency without being selected against with respect to the

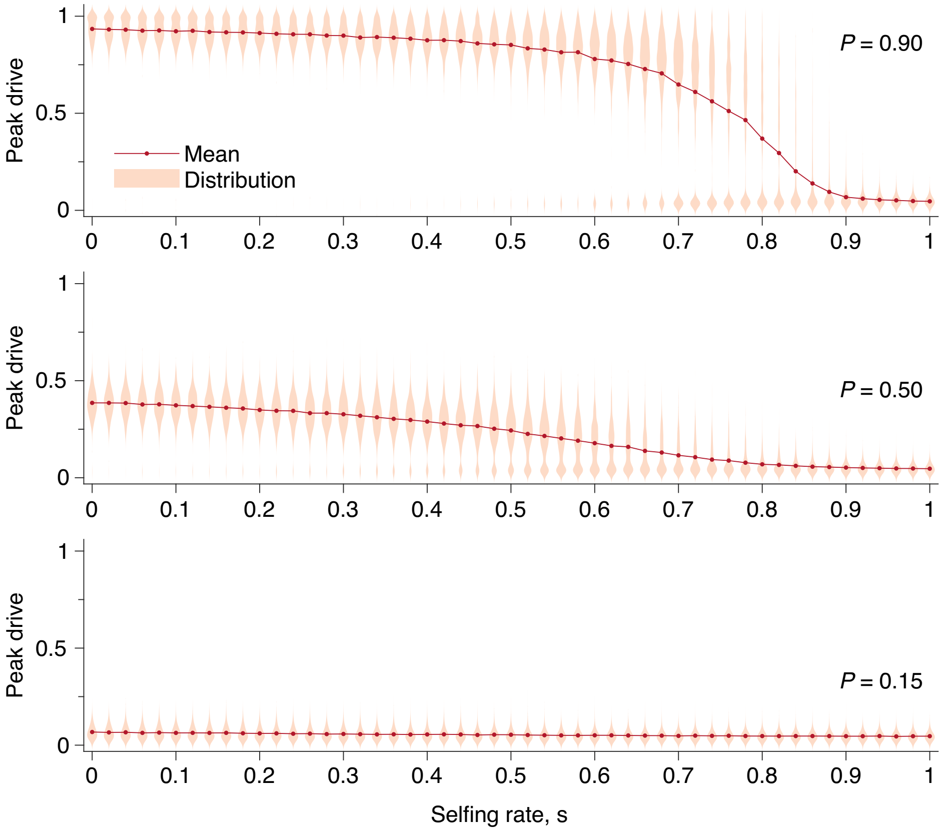


Figure 2.9: Peak drive distributions and means for varying selfing rates in our partial selfing model. (top) Effective drive, $P = 0.9$. (middle) conservative drive, $P = 0.5$, and (bottom) constitutive drive, $P = 0.15$. Each distribution comprises 1000 simulations. Parameters used include a population size of 500 with an initial release of 15 drive homozygotes. Neutral resistance is assumed throughout with no standing genetic variation, and the offspring number per mating is $k = 1$.

wild-type.

When resistance is no longer assumed to be neutral, other interesting dynamics can occur[81]. In particular, when resistance is costly with respect to the wild-type, but not so costly as the drive and its cargo, the dynamics resemble the Rock-Paper-Scissors game. This allows the drive to avoid extinction indefinitely.

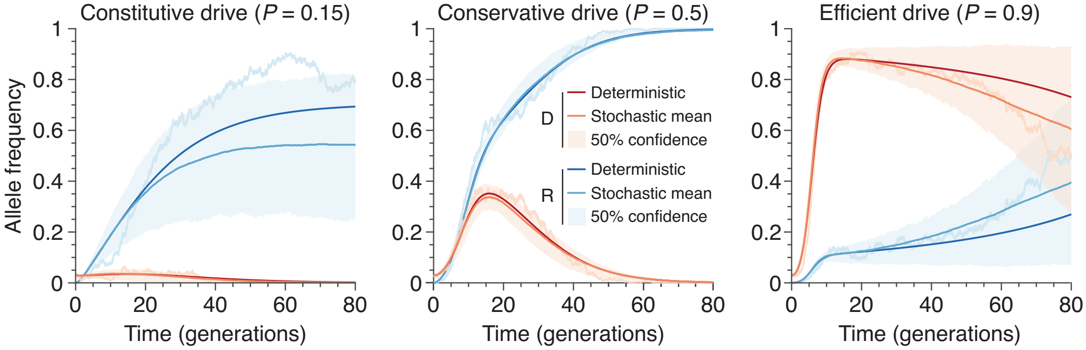


Figure 2.10: Finite-population simulations of 15 drive individuals released into a wild population of size 500, assuming low ($P = 0.5$) or high ($P = 0.9$) homing efficiencies, as well as a low-efficiency, constitutively active system ($P = 0.15$). Deterministic results (dark lines) and means of 10^3 simulations (medium lines), individual sample simulations (light lines), and 50% confidence intervals (shaded). Drive frequencies red and resistant-allele frequencies blue.

2.5.10 Analytic formulae for the escape probability in structured populations

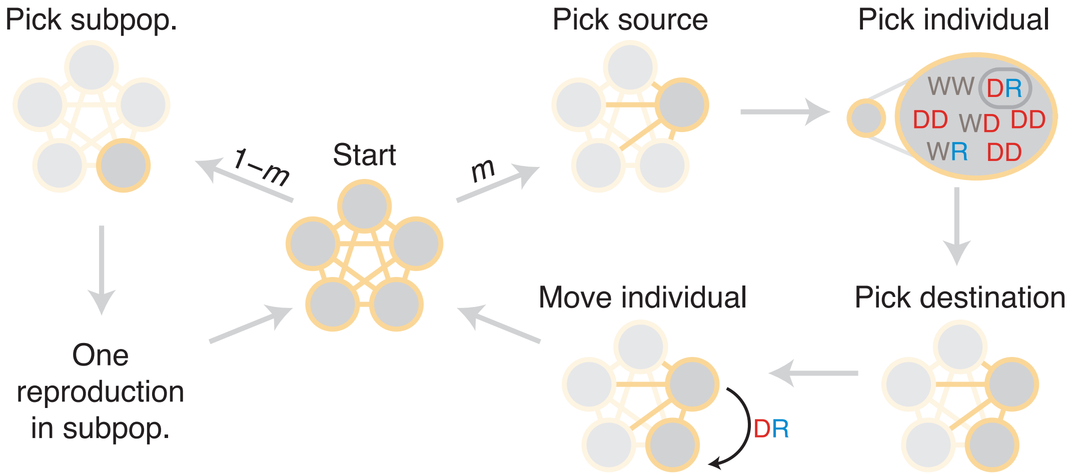


Figure 2.11: Diagram of simulation scheme. In each time step, a migration occurs with probability m, or a mating happens with probability $1 - m$. If a migration occurs, a source population is chosen randomly proportional to its size; an individual is chosen uniformly at random, then a destination is chosen uniformly at random, and the individual is moved. If a mating occurs, the dynamics proceed as in the well-mixed case for a particular subpopulation (Fig. 2.1C).

We consider a deme structured population, where each subpopulation has size N and there are n

demes. We define a Moran-type process, where in each time step either a reproduction or migration event takes place (illustrated in Fig. 2.11). A reproduction event occurs with probability $1 - m$ and a migration event occurs otherwise. If a reproduction occurs, then a subpopulation is selected proportional to the square of its total fitness. Next, two individuals in the subpopulation are selected proportional to their fitnesses and they produce an offspring according to the mechanism above. Finally, another individual from the subpopulation is chosen uniformly at random for death. If a migration event occurs, then an individual is selected uniformly at random and migrates to a new subpopulation uniformly at random. We denote the proportion of genotype α at time t in the initial subpopulation by P_t^α.

The process begins with i drive homozygotes and $N - i$ wild-type homozygotes in a single subpopulation. The remaining subpopulations consist only of wild-type homozygotes. Let $\mathcal{E}$ be the event that the frequency of drive alleles reaches 10% in a subpopulation other than where the drive was released, given that i drive homozygotes were released in the initial subpopulation. We assume that i is small with respect to N.

As an aside, note that the choice of 10% is arbitrary—any other percentage (less than the peak drive in the deterministic model, c) would be equivalent if N is large enough. This is clear from Fig. 2.1E, where either the drive does not invade and so peak drive is roughly equal to the initial frequency or the drive does invade and the peak drive is close to c. This claim is equivalent to stating that the probability that the drive starting at frequency c_0 attains frequency c_1 (such that $c_0 < c_1 < c$) before going extinct tends to 1. This behavior is typical of Moran-type models, since the extinction probability of i drive homozygotes rapidly approaches 0, even in an infinite population, as i

increases[76]. Specifically, if we have $i = c_0 N$, then the extinction probability approaches 0 as N becomes large, and moreover, if the drive does not go extinct, then it behaves almost deterministically and will reach frequency c and thus also c_1.

Returning to approximating the probability of $\mathcal{E}$, note that for $\mathcal{E}$ to take place a drive allele has to migrate from the initial subpopulation *and* this allele has to survive stochastic fluctuations and avoid extinction in its new subpopulation. The drive alleles do not last indefinitely in the initial population. We denote the random time at which the drive alleles go extinct by T. As long as the initial drives do not go extinct due to stochastic fluctuations, the frequency of the drive increases rapidly, as it outcompetes the wild-type. Concurrently, resistant alleles are produced that eventually push the drive to extinction. This means that the drive has a finite time to migrate to other subpopulations. Although this process is stochastic it shows fairly deterministic behavior once there are a sufficient number of drive alleles (see Fig. 2.10)—that is, if the drive avoids immediate extinction. Let $e_{i,j}$, be the probability that the drive survives stochastic fluctuations and avoids immediate extinction when starting with i drive homozygotes and j heterozygotes. Implicitly, here we are assume that $e_{i,j}$ does not depend on whether the heterozygotes are wild-type or resistant heterozygotes. Note that when i or j are $\mathcal{O}(N)$, $e_{i,j}$ is approximately 1, so when $i, j \ll N$, we assume that the probability that the drive migrates is approximately 0. Moreover, since the drive will almost certainly go extinct, there is some time where the frequency of drive alleles is again much less than $\mathcal{O}(N)$. We also assume here that the probability that the drive migrates is approximately 0.

At each time step, there is a small probability that the drive migrates from the initial population and invades another subpopulation. To calculate, we first condition on the non-extinction of the initial i drive homozygotes. Second, we note that if the drive does not migrate and avoid extinction in another subpopulation, then it does not do so at any particular time t. Third, we assume that these events for each t are approximately independent. Finally, we numerically solve a deterministic ODE system representing the dynamics[39] to approximate the probability that the drive does not migrate at time t. Thus,

$$\begin{aligned}
\mathbb{P}\{\mathcal{E}\} &= \mathbb{P}\{\mathcal{E} \mid \text{drive avoids extinction}\}e_{i,0} + \mathbb{P}\{\mathcal{E} \mid \text{drive does not avoid extinction}\}(1-e_{i,0}) \\
&\approx \mathbb{P}\{\mathcal{E} \mid \text{drive avoids extinction}\}e_{i,0} \\
&\approx e_{i,0}\left(1 - \prod_{t=1}^{T} \mathbb{P}\{\text{drive does not migrate and invade at time } t\}\right) \\
&= e_{i,0}\left(1 - \prod_{t=1}^{T} \left(1 - \mathbb{P}\{\text{drive invades} \mid \text{drive migrates at time } t\}\mathbb{P}\{\text{drive migrates at time } t\}\right)\right) \\
&= e_{i,0}\left(1 - \prod_{t=1}^{T} \left(1 - me_{1,0}\mathbb{E}P_t^{DD} - me_{0,1}(\mathbb{E}P_t^{WD} + \mathbb{E}P_t^{DR})\right)\right),
\end{aligned}$$

since if the drive avoids extinction it will invade. Now we substitute the ODE solution $p_t^{\alpha\beta}$ for $\mathbb{E}P_t^{\alpha\beta}$ in the above expression to find that

$$\begin{aligned}
\mathbb{P}\{\mathcal{E}\} &\approx e_{i,0}\left(1 - \exp\left(N\int_0^{T/(1-\lambda)} \mathrm{d}t \log\left(1 - \lambda e_{1,0}p_{(1-\lambda)t}^{DD} - \lambda e_{0,1}\left(p_{(1-\lambda)t}^{WD} + p_{(1-\lambda)t}^{DR}\right)\right)\right)\right) \\
&\approx e_{i,0}\left(1 - \exp\left(\frac{N}{1-\lambda}\int_0^{T} \mathrm{d}t \log(1 - \lambda e_{1,0}p_t^{DD} - \lambda e_{0,1}(p_t^{WD} + p_t^{DR}))\right)\right).
\end{aligned}$$

Here, we approximated the product with an integral and used a change of variables.

Note that if $m = \mathcal{O}(1/T)$ and heuristically we replace $\mathbb{E}P_t^{\alpha}$ in the above expressions with its time average, denoted ϕ^{α}, then

$$\begin{aligned}
&e_{i,0}\left[1-\prod_{t=1}^{T}\left(1-me_{1,0}\mathbb{E}P_t^{DD}-me_{0,1}(\mathbb{E}P_t^{WD}+\mathbb{E}P_t^{DR})\right)\right]\\
&\approx e_{i,0}\left[1-\left(1-\frac{e_{1,0}\phi^{DD}+e_{0,1}(\phi^{WD}+\phi^{DR})}{T}\right)^{T}\right]\\
&\approx e_{i,0}\left[1-\exp\left(-e_{1,0}\phi^{DD}+e_{0,1}(\phi^{WD}+\phi^{DR})\right)\right].
\end{aligned}$$

Thus, when the migration rate is on the order of the inverse of the drive extinction time, the invasion probability is order 1.

3

Daisy-chain gene drives for local alteration

3.1 Foreword

Taken together, the previous chapters suggest that standard CRISPR-based gene drive systems could be difficult to contain within particular populations. While this is a benefit in some situations—for example, malaria eradication, where the goal would be to alter as many populations as possible—it could be a drawback in others. In this chapter, we analyze the dynamics of a new CRISPR-based gene drive design, called “daisy-chain gene drive”, which we hypothesized would offer a balance between spread and containment.

This work was truly an interdisciplinary, team effort. I built and analyzed the mathematical models presented here with Jason Olejarz; John Min, Kevin Esvelt and Andrea Smidler ran preliminary simulations; John Min and Kevin Esvelt designed the library of gRNAs; Joanna Buchthal and Alejandro Chavez performed experiments to analyze the activity of the gRNAs; Erika DeBenedictis created a publicly available web applet for visualizing the model; and we all benefited greatly from insight, advising and support from George Church, Martin Nowak, and Kevin Esvelt.

This chapter first appeared as a preprint on *bioRxiv* (Ref. 65) and is currently under review: Charleston Noble*, John Min*, Jason Olejarz, Joanna Buchthal, Alejandro Chavez, Andrea L. Smidler, Erika A. DeBenedictis, George M. Church, Martin A. Nowak, Kevin M. Esvelt. Daisy-chain gene drives for the alteration of local populations. *bioRxiv* (2016). (*equal contribution)

3.2 Introduction

RNA-guided gene drive systems based on CRISPR nucleases could be used to spread many types of genetic alterations through sexually reproducing species[41]. These systems function by “homing”, or the conversion of heterozygotes to homozygotes in the germline, which renders offspring more likely to inherit the gene drive element and the accompanying alteration than via normal Mendelian inheritance[28] (Fig. 3.1A). To date, gene drive systems based on Cas9 have been demonstrated in yeast[37], fruit flies[38,67], and two species of mosquito[5,13]. Suggested applications include eliminating vector-borne and parasitic diseases, promoting sustainable agriculture, and enabling ecological conservation by curtailing or removing invasive species[41].

The self-propagating nature of standard RNA-guided gene drive systems renders the technology uniquely suited to addressing large-scale ecological problems, but the high likelihood of spread to most populations of the target species[40,76] tremendously complicates discussions of whether and how to proceed with any given intervention[86]. Technologies capable of unilaterally altering the shared environment require broad public support. Because people will not be able to opt-out of technologies intended to alter the shared environment, ethical gene drive research and development should be openly guided by the communities and nations that depend on the potentially affected ecosystems. Unfortunately, attaining this level of engagement becomes progressively more challenging as the size of the affected region increases. Candidate applications that will affect multiple nations could be delayed indefinitely due to a lack of agreement, particularly given the possibility that it may not be possible to conduct safely contained field trials[40,76].

A method of preventing gene drive systems from spreading indefinitely would greatly simplify community-directed development and deployment while also enabling safe field testing. Existing theoretical self-exhausting strategies[64,85] can locally spread cargo genes nearly to fixation if sufficiently many organisms ($>30\%$ of the local population) are released, while "threshold-dependent" drive systems such as those employing underdominance[87] will spread to fixation in small and geographically isolated subpopulations if organisms are released in an amount exceeding the threshold for population takeover (typically $\approx 50\%$). Toxin-based underdominance approaches are promising and have been demonstrated in fruit flies[24,25], though they cannot directly suppress populations. All of these approaches involve releasing comparatively large numbers of organisms, which may not be politically, economically, or environmentally feasible for some applications.

A way to construct highly efficient yet locally confined RNA-guided drive systems could enable many potential applications for which neither self-propagating invasive drive systems nor existing local drives are suitable. Here we describe "daisy drive", a powerful yet self-exhausting form of local drive based on CRISPR-mediated homing in which the drive components are separated into an interdependent daisy-chain. We additionally report newly characterized guide RNA sequences required for evolutionary stability and safe use.

3.3 Design and modeling

A daisy drive system consists of a linear series of genetic elements arranged such that each element drives the next in the chain (Fig. 3.1B). The final element in the chain, which carries the "cargo", is driven to higher and higher frequencies in the population by the earlier elements in the chain. No element can drive itself (Fig. 3.1C). The bottom element is lost from the population over time, causing the next element to cease driving and be lost in turn. This process continues along the chain until, eventually, the population returns to its wild-type state (Fig. 3.1B).

The simplest form of daisy drive—a two-element chain—is obtained by separating CRISPR gene drive components such that the cargo-carrying element, designated 'A', exhibits drive only in the presence of an unlinked, non-driving element, 'B' (Fig. B.1). These "split drives" have been

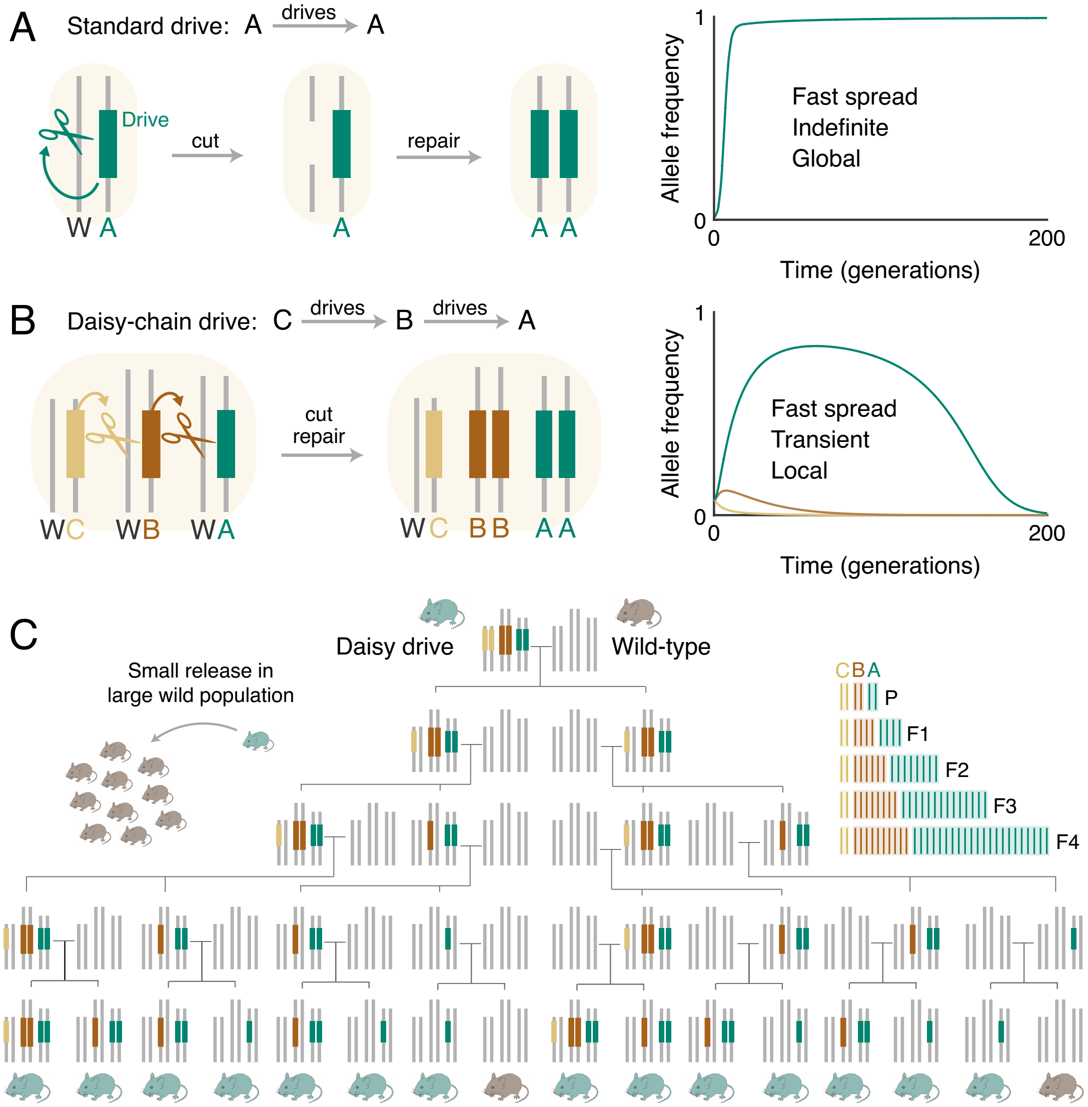


Figure 3.1: Comparison of self-propagating and daisy-chain gene drive. (A) Self-propagating CRISPR gene drives distort inheritance in a self-propagating manner by converting wild-type (W) alleles to drive alleles in heterozygous germline cells. (B) A "daisy drive" system consists of a linear chain of serially dependent, unlinked drive elements; in this example, A, B, and C are on separate chromosomes. Elements at the base of the chain cannot drive and are successively lost over time via natural selection, limiting overall spread. (C) Family tree resulting from the release of a single daisy drive organism in a resident wild-type population in the absence of selection. On the right is a graphical depiction of the total number of alleles per generation. Throughout, chromosome illustrations represent genotypes in germline cells.

described[41], demonstrated[37], and recommended[42] as a stringent laboratory confinement strategy. Because any accidental release would involve only a small number of organisms carrying the B element, the driving effect experienced by the A element—and thus its spread—would be negligible in a large population[37]. As long as the cargo confers a fitness cost to the host organism, both elements will eventually disappear due to natural selection.

We hypothesized that the spread of the cargo-carrying element, A, could be enhanced to useful levels by adding more elements to the base of the daisy chain. To explore this idea, we formulated a deterministic model that considers the evolution of a large population of diploid organisms affected by a daisy drive system with elements spread across n loci (Sections 3.9.1 and 3.9.2). At each locus there are three alleles, the wild-type (W), the corresponding daisy drive element (D) and an allele that is resistant to the effects of the upstream daisy element (R). Such resistant alleles could exist before release in the form of standing genetic variation, or they could be created through misrepair following drive-mediated cleavage or by de novo mutation[39,62,67].

To model the effects of daisy drive in individuals, we make a few assumptions: (i) Daisy drive alleles cut their targeted wild-type alleles with probability 1[5,67,88]; (ii) Drive and resistant alleles are immune to drive-mediated cutting; (iii) Cutting is followed by homologous repair (HR) with probability H, leading to duplication of whatever allele is present at the homologous chromosome, or by nonhomologous end-joining (NHEJ) with probability $1 - H$, resulting in production of a new resistant allele. While we model the rates of outcomes following cutting, we do not vary the cutting efficiency. If the cutting rate were diminished, we expect our results to remain qualitatively similar but with lengthened timescales and perhaps decreased maximum spread.

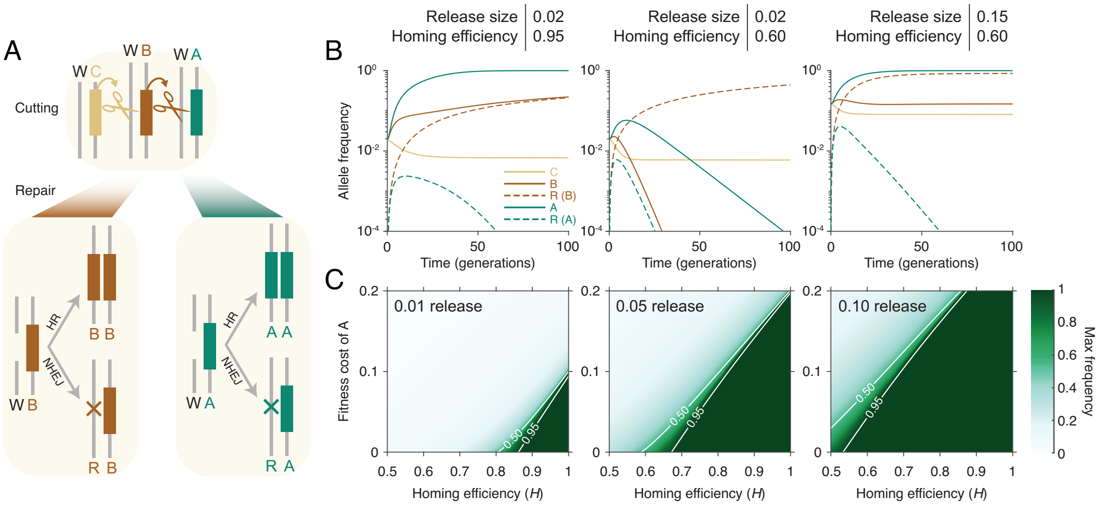


Figure 3.2: Dynamics of CBA daisy-chain gene drive systems. (A) After being cut by an upstream daisy allele, a wild-type allele is repaired either by homologous recombination (HR), creating a second copy of the other allele at the locus, or by nonhomologous repair (e.g., NHEJ), leading to generation of a resistant allele. This process occurs in the germ line and is independent at each locus. We assume that resistance at the cargo locus, A, is dominant lethal if inherited. (B) A highly efficient daisy drive (95% homing efficiency) with an 8% fitness cost for the cargo element seeded at 2%, spreads the cargo nearly to fixation (left). A low-efficiency drive (60%) with the same initial release size no longer allows drive spread (middle). Increasing the release size of the inefficient drive to 15% again allows cargo spread to near fixation (right). (C) The maximum frequency achieved by cargo alleles as a function of the homing efficiency and the cargo fitness cost, for release sizes of 1% (left), 5% (middle), and 10% (right). Throughout, we assume a 0.01% fitness cost for C and B elements and neutral resistant alleles at the C and B loci.

The effect of a daisy drive element at a particular locus (e.g. B) depends on the genotype at the next locus in the daisy-chain (Fig. 3.2A). If that genotype is DD, DR, or RR, then no cutting occurs and the genotype remains unchanged. If the genotype is WW then both wild-type alleles are cut until the locus is converted to RR. Similarly, WR is converted to RR. However, if the genotype is WD, then the W allele is converted to D with probability H, or to R with probability $1 - H$. We assume that standard Mendelian segregation occurs after conversion, so that, for example, individuals initially WD at a locus produce D gametes with probability $(1 + H)/2$ or R gametes with probability $(1 - H)/2$, assuming a daisy allele exists at the previous locus to facilitate the conversion. Finally, we assume that all loci undergo inheritance independently (i.e., all elements are unlinked, ideally on different chromosomes), so that the total probability of an individual producing a gamete of a particular haplotype is the product of its individual-locus inheritance probabilities. Details can be found in Section 3.9.2.2, with gamete-production probabilities explicitly written in Eq. (3.7).

To model selection dynamics, we assume that each daisy drive element confers a dominant fitness cost, c_i, on its host organism. Furthermore, we assume that resistance at every upstream (non-cargo) locus is neutral, while resistance at the cargo locus is dominant lethal. The latter requirement can be attained by targeting a haploinsufficient essential gene with the cargo element while including a genetically recoded copy in the drive construct[39,41]. All costs are assumed to be independent. (See Section 3.9.2.2 for further details. Fitness calculations are performed via Eq. (3.6).)

While the requirement of dominant lethality for resistance at the cargo locus might seem prohibitively difficult to achieve, it is worth noting that recent experimental studies support the feasibility of this approach. In a study of CRISPR-Cas9 gene drive in yeast, DiCarlo *et al.* constructed a drive targeting an essential gene, ABD1, while including a recoded copy in the drive construct, and no obvious impact on fitness was observed compared to wild-type strains[37]. Furthermore, Ostrov *et al.* employed genetic recoding to successfully eliminate seven codons from 91% of essential genes in *E. coli*, leading to an overall fitness cost of less than 10%[57]. Models predict that cutting multiple sites within genes important for fitness is required for a drive system to affect an entire population[39,69], and recent experiments featuring a two-gRNA drive element in fruit flies appear to provide evidence for simultaneous and reliable cutting by more than one gRNA[88].

Gene drive dynamics are sensitive to homing efficiency (H) and fitness cost. In the four species examined, homing efficiency has ranged from 37% to 99%, with almost all the range stemming from variation across experiments in fruit flies. The rate was over 99% for each of the many drive systems tested in yeast[37], 99.8% for the drive system in *An. stephensi*[5], 87.3% to 99.7% for the three drive systems in *An. gambiae*[13], and 37% to 95% for the three drive systems in the fruit fly, which varied with genetic background[38,67]. Fitness costs have not been rigorously measured, but costs associated with non-cargo daisy drive elements are expected to be much lower than typical cargoes[89,90] because they will only encode guide RNAs. Potentially costly off-target cutting is minimal when using high-fidelity Cas9 variants[91,92]. If the target gene is haploinsufficient for proliferative gametogenesis, the cost may approach zero and the homing rate 100% in some species (Fig. B.2).

We studied a three-element daisy drive system (CBA) via numerical simulation (Fig. 3.2). As expected, arbitrarily high frequencies of the cargo element, A, can be achieved by varying the release frequency. However, the system displays high sensitivity to the homing rate and cargo cost. In particular, moderate release sizes (>10% of the resident population) are required to drive costly cargoes if homing efficiency is on the lower end of observed drive systems (≈60%).

We next explored the effects of adding additional elements to the daisy drive system as a potential means of increasing potency. We observe that longer chains lead to much stronger drive (Fig. 3.3A). At a homing efficiency of 95% per daisy drive element, six- and seven-element systems driving a cargo with a 10% cost could be released at frequencies as low as 1% and still exceed 99% frequency in fewer than 20 generations. On a per-organism basis, these are 10 to 1000-fold more efficient than simply releasing organisms with the cargo, depending on the homing efficiency (Fig. B.3).

Adjusting the model to include repeated releases in every subsequent generation, we observed that daisy drives can readily alter local populations if repeatedly released in very small numbers, although the benefit of repeated release is lost when the repeated release size becomes large ($>5\%$) (Fig. 3.3B). This may be useful for applications that must affect large geographic regions over extended periods of time, as well as for local eradication campaigns[93]. (More accurately, we simulated a continuous release of engineered individuals into a wild population for convenience in doing the simulations; see Section 3.9.2.3 for details on this implementation.)

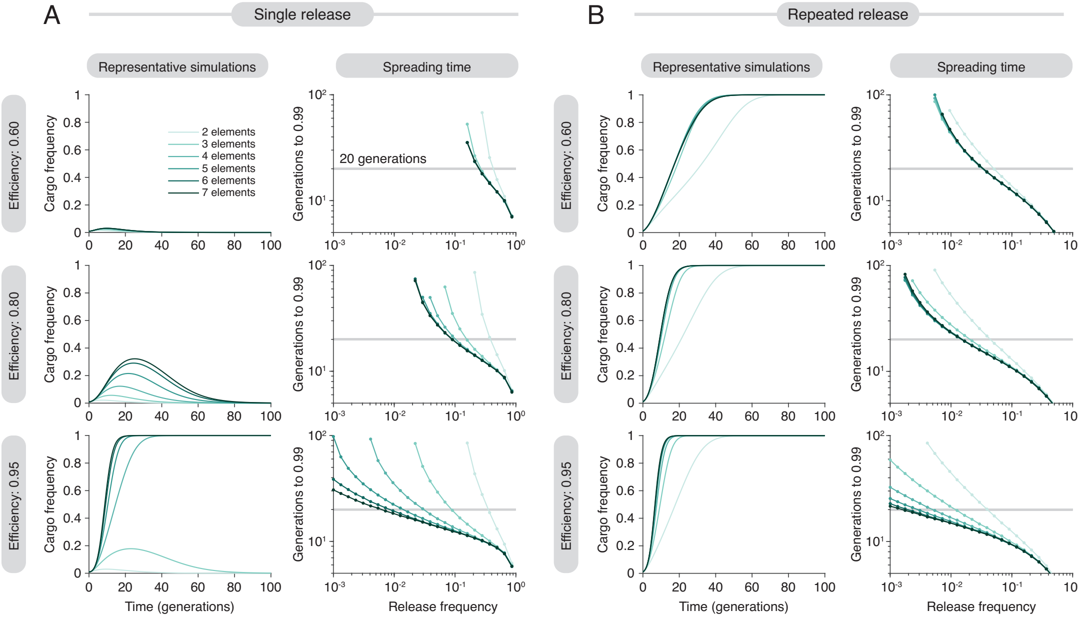


Figure 3.3: Quantitative evaluation of cargo spread in a single population, for single and repeated releases. (A) Results assuming a single release of daisy drive organisms in a wild population. (left) Representative simulations assuming a 1% release. (right) Time to achieve 99% frequency for varying release frequency. (B) Results assuming a constant rate of release of daisy drive organisms. (left) Representative simulations, assuming an initial 1% release with a subsequent release rate of 1% per generation (see Section 3.9.2.3 for details). (right) Time to 99% frequency with varying release rate, which we set as both the initial release frequency as well as the subsequent continuous release frequency, indicated by the horizontal axis. (See Section 3.9.2.3 for details on continuous release.) All simulations assume a 10% cargo cost, 0.01% cost per upstream element, and 60% (top), 80% (middle), or 95% (bottom) homing efficiencies.

Given that the cargo element could achieve arbitrarily high frequencies in a population, we next asked how long the cargo might persist after attaining a high frequency. Thorough quantitative analysis of this point will be an important direction for future work, but as a first step we here sought to understand qualitatively how each of our model parameters impacts this persistence time. To accomplish this, we returned to our basic 3-element (CBA) model and performed the following procedure: (i) We chose a particular set of parameter values such that the drive could attain at least 50% frequency across a range of nearby values for each parameter. (ii) We then varied each parameter individually while measuring the number of generations that the cargo element remained above 50% frequency, thus isolating the effect of each parameter.

The results of this analysis are shown in Fig. B.4. Overall, we find that the persistence time (i.e., the number of generations above 50% frequency) varies significantly across plausible ranges for the parameters in our model. The most dramatic effect is observed by varying the fitness cost of resistance at the cargo element, s. We find that, roughly, if s is less than c, the fitness cost of the cargo element, then the cargo is unlikely to achieve near-fixation, while if $s > c$, then resistance is more deleterious than the cargo itself, and the cargo can remain in the population indefinitely barring mutations that inactivate its function. Regarding the other parameters, we find that the persistence time is inversely proportional to c and more robust to small perturbations in the homing efficiency, H, release frequency, and fitness cost, d, associated with upstream elements (C, B).

Finally, we considered the potential for daisy drive systems to affect local populations of invasive species on islands or other regions with limited gene flow. To study the extent of spread between populations, we formulated a metapopulation model consisting of N populations connected by pairwise gene flow rates in a directed-graph-based structure (Sections 3.9.3 and 3.9.4). Within each population, we assume random mating with selection and germline dynamics identical to those described in the single-population model above.

To begin our analysis of this model, we studied a particular case consisting of 5 equally-sized populations connected in a chain, with each population exchanging individuals with its neighboring populations immediately before and/or after it in the chain (Section 3.9.5). We further assumed gene flow rates of 10^{-2} between each pair of neighboring populations.

Given this population structure, we compared three scenarios, each beginning with a release of engineered individuals in the population at the beginning of the chain (Fig. 3.4): (1) A three-element (CBA) daisy-chain drive; (2) A standard self-propagating drive element designed with multiple gRNAs to mitigate resistance (adapting the model from Chapter 1 and Ref. 39; see Section 3.9.5.2 for details); (3) An inundative release of engineered alleles that do not drive at all. (This scenario was simulated using the same model as in scenario 2, as described in Section 3.9.5.2, except we set the cutting rate, q, to zero so that standard Mendelian inheritance occurs.)

To ensure that the three scenarios were comparable, we employed identical parameters where applicable. In the two drive scenarios (1 and 2) we assumed a moderate 80% homing efficiency, 15% release size and 10% fitness cost for the cargo element (as well as perfect cutting efficiency, as described above). Additionally, for daisy drive, we continued assuming a low fitness cost for the C and B elements (0.01%). For the inundative release scenario, we assumed an identical 10% dominant fitness cost for the engineered element, but we set the release size to 99.9%.

Results for these three initial scenarios can be found in Fig. 3.4. For daisy-chain drive, we find that the cargo element can be driven to near-fixation in its initial-release population while attaining significant frequency (≈ 0.8) in the second population, low frequency in the third population (≈ 0.2) and only negligible frequencies in the subsequent populations. Moreover, transience of the cargo element is ensured in the initial population by influx of wild-type individuals. This constitutes a mechanism for transience that cannot be captured by our single-population model; therefore, we would expect our persistence time results discussed above and presented in Fig. B.4 to be substantially different in this more realistic multiple-population context. In contrast, the self-propagating drive rapidly spreads to near-fixation in all populations.

We then further analyzed inter-population spread in this model via numerical simulation, and additional results can be found in Fig. B.5. Specifically, we varied the migration rate between 10^{-4} and 10^{-1} for each of the three scenarios described above and measured the maximum frequency achieved by each allele across 500-generation simulations. We find that, for migration rates below 10^{-2} (the value assumed in Fig. 3.4), maximum daisy-chain cargo frequency in the second population decreases roughly linearly with the migration rate, whereas self-propagating drive approaches fixation in all populations even for very low migration rates. Notably, the resistant allele at the B locus can exhibit high frequencies in multiple populations due to its assumed low fitness cost; however, this effect could potentially be mitigated by engineering that element to select against resistance in the same way as the A element.

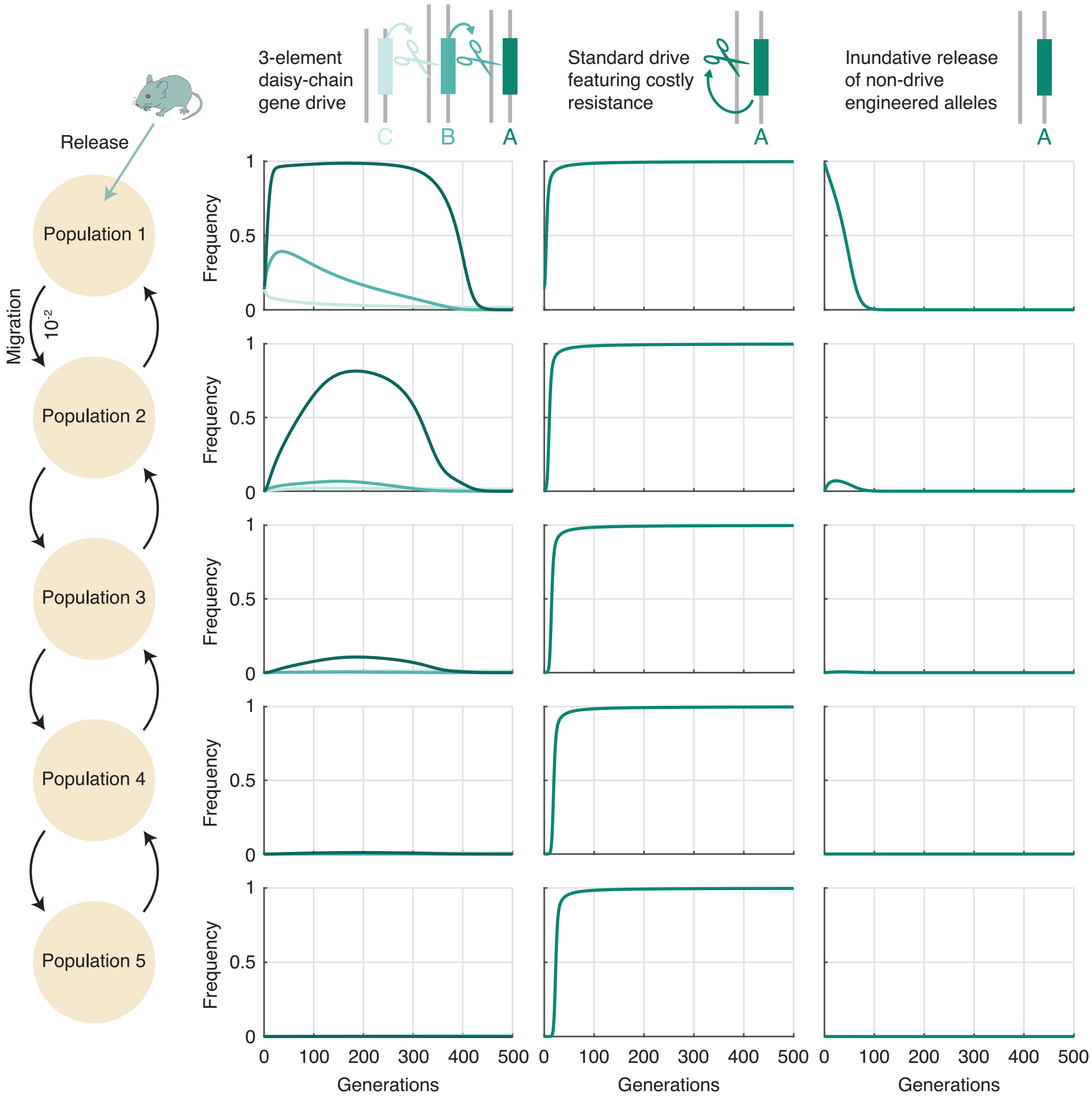


Figure 3.4: Modeling daisy drive containment in a system of populations connected by gene flow. (left) Illustration of the population structure: five populations with equal sizes are connected in a chain, and each neighboring pair has bidirectional gene flow with rate 10^{-2} in each direction. The three figure panel columns then correspond to the three scenarios described in the text: (left) CBA daisy-chain drive, (middle) self-propagating ("standard") drive with multiple gRNAs targeting an essential gene, as in Chapter 1 and Ref. 39, (right) non-drive inundative release. Frequencies over time are indicated for each allele in each of the populations. Drive-based simulations (daisy-chain and standard) assume 80% homing efficiency, 10% dominant cargo element fitness cost and 15% release frequency. Daisy-chain drive simulations further assume 0.01% upstream element (C, B) fitness cost. Inundative release simulations assume 10% dominant fitness cost and 99.9% release. See Section 3.9.5 for details.

3.4 Evolutionary stability and CRISPR multiplexing

Despite these promising theoretical results, current technological limitations preclude the safe use of daisy drive systems. Specifically, a recombination event that moves one or more guide RNAs within an upstream element of the chain into any downstream element could convert a linear daisy drive chain into a self-propagating 'necklace' anticipated to spread to populations worldwide (Fig. 3.5A).

One way to reliably prevent such events is to eliminate regions of homology between the elements. Promoter homology can be removed by using different U6, H1, or tRNA promoters to express the required guide RNAs[94–96]; if there are insufficient promoters then each can drive expression of multiple guide RNAs using tRNA[97,98] or miRNA processing[99–101]. However, each element must still encode multiple guide RNAs >80 base pairs in length in order to prevent the creation of drive-resistant alleles, precluding safe and stable daisy drive designs.

One alternative is to use a distinct orthogonal CRISPR system for every daisy element[102] (Fig. B.6). Unfortunately, it is more difficult to find multiple promoters suitable for nuclease expression than for gRNA expression, and the fitness cost is likely higher than an equivalent gRNA element. We accordingly sought to identify highly active guide RNA sequences for *S. pyogenes* Cas9 with minimal homology to one another that could enable safe daisy drive using only a single CRISPR nuclease.

We compared known tracrRNA, crRNA, and alternative sgRNA sequences for CRISPR systems related to that of *S. pyogenes* to identify bases tolerant of variation[103,104] within the sequence of the most commonly used sgRNA (Figs. 3.5B-C, B.7). We then created dozens of sgRNA variants designed to be as divergent from one another as possible. Assaying these using a sensitive tdTomato-based transcriptional activation reporter in human cell culture identified 15 different sgRNAs with activities comparable to the self-propagating version (Fig. 3.5D). Activity increased with the length of the first stem in agreement with other reports[105] (Figs. B.8 and B.9). This set of minimally homologous sgRNAs can be used to construct stable daisy drive systems of up to 5 elements with 4 sgRNAs per driving element, and will also facilitate multiplexed Cas9 targeting in the laboratory by permitting the commercial synthesis of DNA fragments encoding many sequence-divergent guide RNAs. Future studies will need to examine the stability of the resulting daisy drive systems in large populations of animal models.

Importantly, our divergent guide RNAs will also enable self-propagating CRISPR gene drive elements to overcome the problem of instability caused by including multiple repetitive guide RNA sequences in the drive cassette[106], which is needed to overcome drive-resistant alleles[39,69]. Using non-repetitive guides may consequently allow stable and efficient self-propagating drive systems to affect every organism in the target population.

3.5 Construction and deployment

On a practical level, researchers need only construct one ‘generic’ daisy drive strain per species that could subsequently be loaded with any desired cargo. This generic daisy drive system, which would typically harbor the nuclease gene in the A position, could be used in three different ways.

First, one or more “effector” elements carrying cargo genes and guide RNAs sufficient to drive themselves in the presence of nuclease could be added directly to the generic daisy drive strain. In

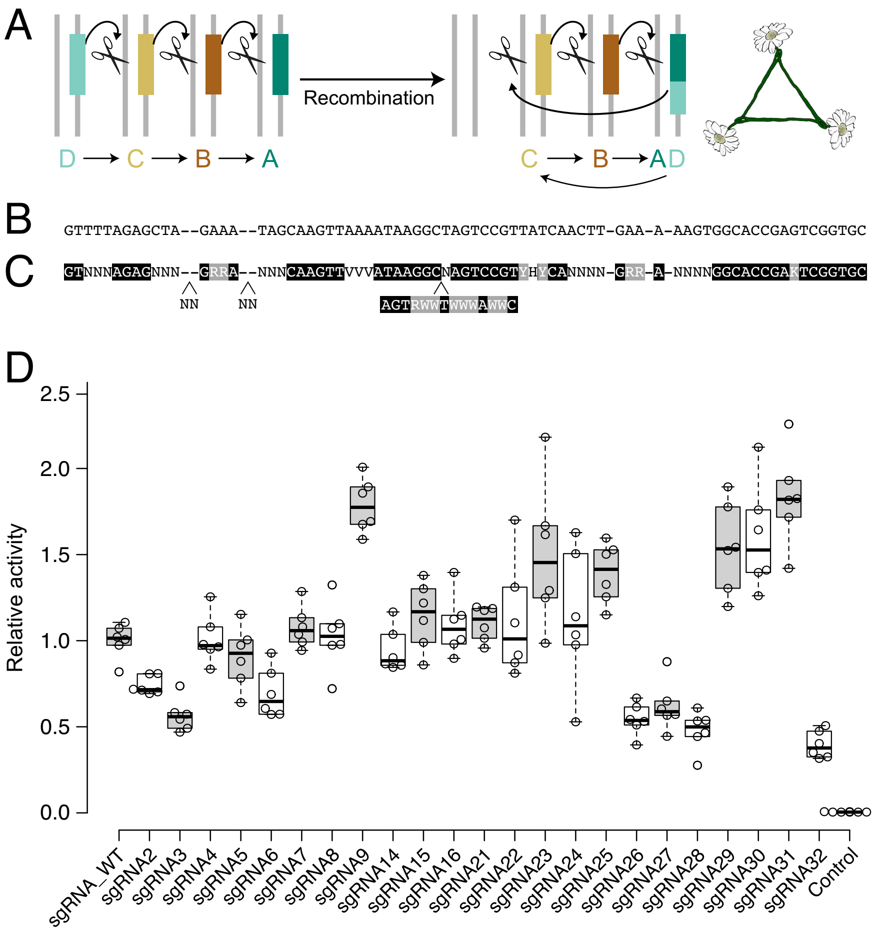


Figure 3.5: Preventing the formation of "daisy necklaces". (A) Any recombination event that moves a guide RNA from one element to another, where it will be reliably copied, could create a "daisy necklace" capable of self-propagating drive. (B) Because promoters can be changed, repetition of the conserved guide RNA sequence is a key problem. (C) Using existing data, we generated a template identifying candidate positions presumed tolerant of sequence changes. (D) Relative activities of candidate guide RNAs generated from the template were assayed using a Cas9 transcriptional activator screen using a tdTomato reporter in human cells.

this configuration, the nuclease-encoding element would become the B element with the effector(s) in the A position. These daisy-drive organisms would then be mass-produced and released in a single-strain, single-stage approach.

Second, the generic daisy drive strain could be released in the target region alongside organisms carrying effector elements already present from releases in adjacent areas. Matings in the wild would then combine the daisy-chain and effector elements, allowing more precise control in spreading the effector cargo into new areas.

Third, the generic daisy drive strain could be released without an effector, and the spread of the nuclease gene could be monitored. This would allow for precise prediction and tuning of the region affected before a later release of strains carrying effector elements to initiate the desired effect. If necessary, the extent of nuclease spread could be adjusted by releasing wild-type or more daisy-drive organisms to fine-tune the areas affected, allowing a level of control not afforded by classic gene drive architectures, albeit one that is imperfect due to stochastic migration. Superior control might be obtained by coupling daisy drive to underdominance to limit dispersion of the alteration to areas in which it is already in the majority[107].

3.6 Field trials and safeguards

Ecological problems such as malaria are so widely distributed geographically that addressing them may require self-propagating CRISPR-based gene drive systems. However, alteration drive systems of this type arguably cannot be tested in field trials without a substantial risk of eventual international spread[40,76], and future models may demonstrate that the same is true of self-propagating suppression drive. Daisy drive systems, which are capable of mimicking the molecular effects of any self-propagating drive on a local level, may offer a potential solution.

Notably, daisy drive systems might be used to directly suppress target populations by imposing a genetic load or by sex-biasing the local population, exactly as would equivalent self-propagating CRISPR-based drive systems. For example, a daisy drive that disrupts female fertility genes, such as those recently identified in malarial mosquitoes[13], might encode the basal element of the daisy chain on the Y chromosome or an equivalent male-specific locus, thereby ensuring that most male offspring preferentially inherit the complete daisy suppression drive system and enabling outcrossing to wild females during production (Fig. B.10). As with a Y-linked suppression element[108], such males should suffer no direct fitness costs from the genetic load relative to competing wild-type males.

Finally, scientists currently have few attractive options for controlling unauthorized or accidentally-released CRISPR-based gene drive systems. While it is possible to overwrite genome-level alterations and undo phenotypic changes using immunizing reversal drives[41], these countermeasures must necessarily spread to the entire population in order to immunize them against the unwanted drive system; strategies based on pure reversal drives[37] or variations such as gene drive 'brakes'[109] should only slow it down. In contrast, daisy drive systems may be powerful enough to eliminate all copies of an unwanted self-propagating drive system via local immunizing reversal, population suppression, or both (conceptually illustrated in Fig. B.11). Feasibility, especially in species with high dispersal rates, should be investigated by modeling and metapopulation experiments.

3.7 Discussion

RNA-guided gene drives based on CRISPR have generated considerable excitement as a potential means of addressing otherwise intractable ecological problems. While experiments have raced ahead at a rapid pace, the high likelihood of international spread once released into a wild population may prove a formidable barrier given the need for public support and international regulatory approval, which may not be achievable if the proposed system cannot be safely tested in the field. These ethical and diplomatic complications are most acute for drive systems aiming to solve the most urgent humanitarian problems, including malaria, schistosomiasis, dengue, and other vector-borne and parasitic diseases, as the lack of international agreement could significantly delay releases.

Similarly, the potential for RNA-guided drive systems to be released accidentally or unilaterally has led to many calls for caution and expressions of alarm, not least from scientists in the vanguard of the field[41–43]. Any such event could have potentially devastating consequences for public trust and support for future interventions.

In contrast, our results suggest that daisy drive systems might be safely developed in the laboratory, assessed in the field, and deployed to accomplish transient alterations that should minimally impact other nations or jurisdictions. They might be used to locally duplicate the effects of a self-propagating drive system for safe field studies, to efficiently alter entire local populations with limited gene flow such as those on islands, or to accomplish transient changes to pockets of mainland populations.

However, it is essential to note that daisy drive alone cannot prevent the spread of engineered genes into adjacent populations[110]. Addressing this problem further could require, for example, triggering a threshold-dependent drive system after the daisy drive has been exhausted to actively eliminate engineered alleles from adjacent populations where they are in the minority[107].

By using molecular constraints to limit generational and geographic spread, daisy drive approaches could expand the scope of ecological engineering by enabling local communities to make decisions concerning their own local environments.

3.8 Experimental methods

The biological experiments performed for this study were designed and carried out by John Min, Joanna Buchthal and Alejandro Chavez, with advising from Kevin Esvelt and George Church. For completeness (as this work has been submitted for publication together in its entirety), I present the results of these experiments in Section 3.4, as well as in Figs. 3.5, B.8, and B.9, and the related methods follow in this Section (Sec. 3.8).

3.8.1 Guide RNA design

We examined existing data on crRNA and tracrRNA sequences from closely related Cas9 systems (from Fig. S2 of Ref. 103 and Fig. 4 of Ref. 111) by multiple sequence alignment[112,113], as well as the crystal structure of *S. pyogenes* Cas9 in complex with sgRNA, to construct a template specifying bases most likely to tolerate mutations. The template is shown in Fig. 3.5C, and the tracrRNA multiple sequence alignment is shown in Fig. B.7. We used this template to design a set of 20 sgRNA sequences sharing no more than 17bp of homology with one another. Activity assays (see below) with two replicates identified sequence changes harmful to activity. These experiments suggested that the large insertion found in sgRNAs from closely related bacteria was well-tolerated in only one case. Additional sgRNAs lacking this feature were designed to preserve the 17bp homology limit across the set. All candidates were then assayed to identify those with sufficiently high activity. Future experiments requiring additional highly divergent sgRNAs, such as daisy suppression drives in which the A element encodes many guide RNAs that disrupt multiple recessive fertility genes at multiple sites, will require a more comprehensive library-based approach to activity profiling.

3.8.2 Measuring guide RNA activity

HEK293T cells were grown in Dulbecco's Modified Eagle Medium (Life Technologies) fortified with 10% FBS (Life Technologies) and Penicillin/Streptomycin (Life Technologies). Cells were incubated at a constant temperature of 37°C with 5% CO2. In preparation for transfection, cells were split into 24-well plates, divided into approximately 50,000 cells per well. Cells were transfected using 2ul of Lipofectamine 2000 (Life Technologies) with 200ng of dCas9 activator plasmid, 25ng of guide RNA plasmid, 60ng of reporter plasmid and 25ng of EBFP2 expressing plasmid.

Fluorescent transcriptional activation reporter assays were performed using a modified version of addgene plasmid #47320, a reporter expressing a tdTomato fluorescent protein adapted to contain an additional gRNA binding site 100bp upstream of the original site. gRNAs were co-transfected with reporter, dCas9-VPR, a tripartite transcriptional activator fused to the C-terminus of nuclease-null Streptococcus pyogenes Cas9, and an EBFP2 expressing control plasmid into HEK293T cells. 48 hours post-transfection, cells were analyzed by flow cytometry. In order to exclusively analyze transfected cells, cells with less than 10^3 arbitrary units of EBFP2 fluorescence were ignored. The preliminary screen of the initial 20 designs was performed with only two replicates to identify critical bases. Experiments evaluating the final set of sgRNA sequences were performed with six biological replicates.

3.9 Supplementary model details

In this Section, we develop the mathematical models used for numerical simulations throughout this chapter. We begin with a very simple model of daisy-chain gene drive and then successively extend it until concluding with the versions used for simulations. We begin in Section 3.9.1 by presenting a simple model for daisy-chain gene drive without resistant alleles. This description begins with a simple 2-element model (Section 3.9.1.1), which is extended to include n elements in Section 3.9.1.2. Then, in Section 3.9.2, we extend the 2-element and n-element models from Section 3.9.1 to include resistant alleles, resulting in the model used for all single-population simulations throughout this chapter. In Section 3.9.3, we then extend the model with resistance (Sec. 3.9.2) to include two distinct populations connected by gene flow, and this model is then extended to N populations in Section 3.9.4. Lastly, in Section 3.9.5 we explicitly write the equations for the special case of the model from Section 3.9.4 wherein 5 islands are connected in a chain, results from which are presented in Figs. 3.4, B.5 and related discussion.

3.9.1 Evolutionary dynamics of a daisy drive construct

To begin, we describe a model for a daisy drive system consisting of only two elements (i.e., B and A), with only wild-type and drive alleles at each locus (i.e., no resistance). This simple case demonstrates the principles behind daisy drive engineering and illustrates the modeling approaches we employ in the more complex scenarios. We then describe a daisy drive system with an arbitrary number of elements in Section 3.9.1.2.

3.9.1.1 Model for a 2-element daisy drive

We consider a wild population of diploid organisms and focus on two loci, "1" and "2". The wild-type alleles at the two loci are 1_W and 2_W, and we denote by $1_{WW}2_{WW}$ the genotype of an individual that is homozygous for both.

Using CRISPR genome editing technology, one can engineer what we refer to as "daisy" alleles at both loci (1_D and 2_D). They function as follows. The 1_D allele effects cutting of the 2_W allele in an individual's germline. We assume that the two loci are independent and that a single copy of 1_D always induces cutting of the 2_W allele.

In addition, we assume for now (relaxed in Section 3.9.2) that the W allele at the second locus is a haploinsufficient essential gene that is targeted with multiple gRNAs (as described in Chapter 1) and that the 2_D allele contains a genetically recoded copy of of the wild-type allele. Therefore, 2_{WW}, 2_{DW} and 2_{DD} genotypes are all viable, but a loss-of-function variant of the 2_W allele—which can result from drive-mediated cutting without successful homing, due to the multiple gRNA assumption—is dominant lethal.

Due to this multiple gRNA/haploinsuffient assumption, if an individual has genotype $1_{WD}2_{WW}$ or $1_{DD}2_{WW}$, then the drive allele at the first locus cuts and disrupts both wild-type alleles at the second locus, resulting in nonviable gametes. If an individual has genotype $1_{WD}2_{WD}$ or $1_{DD}2_{WD}$, then the drive allele at the first locus cuts the wild-type allele at the second locus, and one of two things can happen. If a homing event occurs, then the drive allele at the second locus is successfully copied into the position of the shredded wild-type allele, resulting in gametes that necessarily have

Genotype	$1_W 2_W$	$1_W 2_D$	$1_D 2_W$	$1_D 2_D$
$1_{WW} 2_{WW}$	1	0	0	0
$1_{WW} 2_{WD}$	$\frac{1}{2}F$	$\frac{1}{2}F$	0	0
$1_{WW} 2_{DD}$	0	F	0	0
$1_{WD} 2_{WW}$	0	0	0	0
$1_{WD} 2_{WD}$	0	$\frac{1}{2}HF$	0	$\frac{1}{2}HF$
$1_{WD} 2_{DD}$	0	$\frac{1}{2}F$	0	$\frac{1}{2}F$
$1_{DD} 2_{WW}$	0	0	0	0
$1_{DD} 2_{WD}$	0	0	0	HF
$1_{DD} 2_{DD}$	0	0	0	F

Table 3.1: Gamete production table showing the relative rates at which individuals of each genotype (rows) produce gametes of each haplotype (columns).

the drive allele at the second locus. If a homing event does not occur, then the resulting gametes are nonviable. This results in super-Mendelian inheritance of the 2_D allele in a 1_D-mediated fashion. Importantly, the 1_D allele undergoes standard inheritance and does not facilitate its own spread similarly.

(Notice that in this simplified treatment, we do not explicitly study evolution with a resistant allele, as described in the main text. This simplified model illustrates the principle behind daisy drive engineering without concern for complications arising from emergence of resistance. In Section 3.9.2, we introduce resistance into the model.)

To see how the daisy drive works, consider Table 3.1, which is understood as follows:

Gametes of haplotype $1_W 2_W$ are produced in the following ways:

- $1_{WW} 2_{WW}$ individuals produce only $1_W 2_W$ gametes. We set the rate of production of $1_W 2_W$ gametes by $1_{WW} 2_{WW}$ individuals to be 1.
- $1_{WW} 2_{WD}$ individuals produce gametes with a wild-type allele at the second locus with probability $1/2$. There is a fitness effect, F, due to the payload of the drive allele at the second locus. So $1_{WW} 2_{WD}$ individuals produce $1_W 2_W$ gametes at relative rate $F/2$.

Gametes of haplotype $1_W 2_D$ are produced in the following ways:

- $1_{WW}2_{WD}$ individuals produce gametes with a drive allele at the second locus with probability $1/2$. There is a fitness effect, F, due to the payload of the drive allele at the second locus. So $1_{WW}2_{WD}$ individuals produce $1_W 2_D$ gametes at relative rate $F/2$.
- $1_{WW}2_{DD}$ individuals produce only $1_W 2_D$ gametes. There is a fitness effect, F, due to the payload of the drive allele at the second locus. So $1_{WW}2_{DD}$ individuals produce $1_W 2_D$ gametes at relative rate F.
- $1_{WD}2_{WD}$ individuals produce gametes with a wild-type allele at the first locus with probability $1/2$. The action of the drive allele at the first locus is to cut the wild-type allele at the second locus, and homing occurs with probability H. There is a fitness effect, F, due to the payload of the drive allele at the second locus. So $1_{WD}2_{WD}$ individuals produce $1_W 2_D$ gametes at relative rate $HF/2$.
- $1_{WD}2_{DD}$ individuals produce gametes with a wild-type allele at the first locus with probability $1/2$. There is a fitness effect, F, due to the payload of the drive allele at the second locus. So $1_{WD}2_{DD}$ individuals produce $1_W 2_D$ gametes at relative rate $F/2$.

Gametes of haplotype $1_D 2_D$ are produced in the following ways:

- $1_{WD}2_{WD}$ individuals produce gametes with a drive allele at the first locus with probability $1/2$. The action of the drive allele at the first locus is to cut the wild-type allele at the second locus, and homing occurs with probability H. There is a fitness effect, F, due to the payload of the drive allele at the second locus. So $1_{WD}2_{WD}$ individuals produce $1_D 2_D$ gametes at relative rate $HF/2$.
- $1_{WD}2_{DD}$ individuals produce gametes with a drive allele at the first locus with probability $1/2$. There is a fitness effect, F, due to the payload of the drive allele at the second locus. So $1_{WD}2_{DD}$ individuals produce $1_D 2_D$ gametes at relative rate $F/2$.
- $1_{DD}2_{WD}$ individuals have only the drive allele at the first locus. The action of the drive allele at the first locus is to cut the wild-type allele at the second locus, and homing occurs with probability H. There is a fitness effect, F, due to the payload of the drive allele at the second locus. So $1_{DD}2_{WD}$ individuals produce $1_D 2_D$ gametes at relative rate HF.

- $1_{DD}2_{DD}$ individuals produce only 1_D2_D gametes. There is a fitness effect, F, due to the payload of the drive allele at the second locus. So $1_{DD}2_{DD}$ individuals produce 1_D2_D gametes at relative rate F.

(Notice that if H is interpreted as the homing probability and F is interpreted as the fitness effect due to the drive payload, then Table 3.1 is naturally interpreted as describing drive that occurs in the embryo. That is, individuals with at least one copy of the drive allele at the first locus and a single copy of the drive allele at the second locus shred the wild-type allele at the second locus during embryonic development. And if homing does not occur, then the resulting, mature individuals are nonviable since the W (or D) allele is haploinsufficient. But Table 3.1 also effectively describes the production of gametes in the case of meiotic drive. The subtle distinction in that case would be that, if cutting occurs and homing does not follow, then $1_{WD}2_{WD}$ and $1_{DD}2_{WD}$ individuals produce a nonzero amount of gametes with a mutilated wild-type allele at the second locus. But when those gametes pair with any other gamete, the resulting individuals are necessarily nonviable, and so, effectively, $1_{WD}2_{WD}$ and $1_{DD}2_{WD}$ individuals only produce gametes with a drive allele at the second locus.)

Using these rules, we can formally express the rates at which the four types of gametes are produced in the population. We denote by $g(z)$ the rate (with implicit time-dependence) at which gametes with haplotype z are produced by individuals in the population.

$$
\begin{aligned}
g(1_W 2_W) &= x(1_{WW} 2_{WW}) + \frac{1}{2} F x(1_{WW} 2_{WD}) \\
g(1_W 2_D) &= \frac{1}{2} F x(1_{WW} 2_{WD}) + F x(1_{WW} 2_{DD}) + \frac{1}{2} H F x(1_{WD} 2_{WD}) + \frac{1}{2} F x(1_{WD} 2_{DD}) \\
g(1_D 2_W) &= 0 \\
g(1_D 2_D) &= \frac{1}{2} H F x(1_{WD} 2_{WD}) + \frac{1}{2} F x(1_{WD} 2_{DD}) + H F x(1_{DD} 2_{WD}) + F x(1_{DD} 2_{DD})
\end{aligned}
$$

Here, $x(z)$ is the frequency of individuals with genotype z.

The selection dynamics are then modeled by the following system of equations:

$$
\begin{aligned}
\dot{x}(1_{WW} 2_{WW}) &= & g(1_W 2_W)^2 - \psi^2 x(1_{WW} 2_{WW}) \\
\dot{x}(1_{WW} 2_{WD}) &= & 2 g(1_W 2_W) g(1_W 2_D) - \psi^2 x(1_{WW} 2_{WD}) \\
\dot{x}(1_{WW} 2_{DD}) &= & g(1_W 2_D)^2 - \psi^2 x(1_{WW} 2_{DD}) \\
\dot{x}(1_{WD} 2_{WW}) &= & 2 g(1_W 2_W) g(1_D 2_W) - \psi^2 x(1_{WD} 2_{WW}) \\
\dot{x}(1_{WD} 2_{WD}) &= & 2 g(1_W 2_D) g(1_D 2_W) + 2 g(1_W 2_W) g(1_D 2_D) - \psi^2 x(1_{WD} 2_{WD}) \\
\dot{x}(1_{WD} 2_{DD}) &= & 2 g(1_W 2_D) g(1_D 2_D) - \psi^2 x(1_{WD} 2_{DD}) \\
\dot{x}(1_{DD} 2_{WW}) &= & g(1_D 2_W)^2 - \psi^2 x(1_{DD} 2_{WW}) \\
\dot{x}(1_{DD} 2_{WD}) &= & 2 g(1_D 2_W) g(1_D 2_D) - \psi^2 x(1_{DD} 2_{WD}) \\
\dot{x}(1_{DD} 2_{DD}) &= & g(1_D 2_D)^2 - \psi^2 x(1_{DD} 2_{DD})
\end{aligned}
$$

Here, an overdot denotes the time derivative, d/dt. Throughout this Chapter, we omit explicitly writing the time dependence of our dynamical quantities. Note that this formulation assumes random mating, i.e., that two random gametes come together to form an individual. Also note that products $g(y)g(z)$ represent the pairings of different gametes. At any given time, we require that the total number of individuals sums to one:

$$\sum_z x(z) = 1$$

To enforce this density constraint, we set

$$\psi = g(1_W 2_W) + g(1_W 2_D) + g(1_D 2_W) + g(1_D 2_D)$$

3.9.1.2 Model for an n-element daisy drive

We can apply the same engineering to a daisy drive chain of arbitrary length, n, where the drive allele at one locus induces cutting of the wild-type allele at the next locus in the sequence. To describe this mathematically, it is helpful to generalize our notation.

Consider a daisy drive construct with only two loci, as in Section 3.9.1.1. We use a "1" bit to denote a wild-type allele, and we use a "0" bit to denote a daisy drive allele. To represent genotypes, we introduce vectors $a = (a_1, a_2)$ and $b = (b_1, b_2)$, where each $a_1, a_2, b_1, b_2 \in \{0, 1\}$. We construct these vectors such that a_1 and b_1 represent the two alleles at the first locus, while a_2 and b_2 represent the two alleles at the second locus. A full genotype is then a list of the two vectors, $[a, b]$. We write the nine possible genotypes for a two-element drive system as:

$$\begin{aligned}
1_{WW}2_{WW} &= [(1,1),(1,1)] \\
1_{WW}2_{WD} &= [(1,1),(1,0)] \\
1_{WW}2_{DD} &= [(1,0),(1,0)] \\
1_{WD}2_{WW} &= [(1,1),(0,1)] \\
1_{WD}2_{WD} &= [(1,1),(0,0)] \\
1_{WD}2_{DD} &= [(1,0),(0,0)] \\
1_{DD}2_{WW} &= [(0,1),(0,1)] \\
1_{DD}2_{WD} &= [(0,1),(0,0)] \\
1_{DD}2_{DD} &= [(0,0),(0,0)]
\end{aligned}$$

Notice that if an individual is heterozygous at a particular locus, then this notation allows for two ways of writing the alleles at that locus. For example, genotype $1_{WD}2_{WD}$ can be written in any one of four equivalent ways: $[(1,1),(0,0)]$, $[(0,0),(1,1)]$, $[(1,0),(0,1)]$, or $[(0,1),(1,0)]$.

When modeling daisy drives with a large number of loci, it is helpful to adopt shorthand notation. To do this, we extend the lengths of a and b to be equal to the number of loci, n. That is, we let $a = (a_1,\ldots,a_n)$ and $b = (b_1,\ldots,b_n)$, where each $a_i, b_j \in \{0,1\}$. For example, the genotype $1_{WW}2_{DD}3_{WD}$ can be written $[a,b] = [(1,0,1),(1,0,0)]$ or, equivalently, $[a,b] = [(1,0,0),(1,0,1)]$.

We denote by x_{ab} the frequency of individuals with genotype $[a,b]$. We denote by g_b the rate at which gametes with haplotype b are produced. For an n-element daisy drive, g_b is given by

$$
\begin{aligned}
g_b =& \sum_{\alpha,\beta} x_{\alpha\beta} F^{1-\alpha_n\beta_n} \\
&\times \prod_{i=1}^{n} \left\{ \delta_{\alpha_i b_i}\delta_{\beta_i b_i}\left[\delta_{0,b_i} + \alpha_{i-1}\beta_{i-1}\delta_{1,b_i}\right] + (1-\delta_{\alpha_i\beta_i})\left[\frac{\alpha_{i-1}\beta_{i-1}}{2} + (1-\alpha_{i-1}\beta_{i-1})H\delta_{0,b_i}\right]\right\}
\end{aligned}
\tag{3.1}
$$

Here, we have defined $\alpha_0 = \beta_0 = 1$. δ_{ij} is the Kronecker delta, defined by $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$. In Equations (3.1), in the sum over α, β when enumerating genotypes, heterozygous loci ($\alpha_i \neq \beta_i$) are each counted once, so there is no double-counting. g_b is linear in each $x_{\alpha\beta}$, where all genotypes $[\alpha, \beta]$ are summed over.

We understand the terms in the factors in brackets as follows. Consider just a single factor in brackets for a particular value of i.

- If $\alpha_i = \beta_i = b_i = 0$, then individuals of genotype $[\alpha, \beta]$ have two identical copies of allele 0 at the i^{th} locus, and those individuals create only gametes with allele 0 at position i.
- If $\alpha_i = \beta_i = b_i = 1$ and $\alpha_{i-1}\beta_{i-1} = 1$, then individuals of genotype $[\alpha, \beta]$ have two identical copies of allele 1 at the i^{th} locus and no copy of allele 0 at the $(i-1)^{\text{th}}$ locus, and those individuals create only gametes with allele 1 at position i.
- If $\alpha_i \neq \beta_i$ and $\alpha_{i-1}\beta_{i-1} = 1$, then individuals of genotype $[\alpha, \beta]$ have a single copy of allele b_i at the i^{th} locus, and without any action from the daisy drive, those individuals create gametes with allele b_i and allele $(1 + (-1)^{b_i})/2$ at position i in equal proportion.
- If $\alpha_i \neq \beta_i$ and $\alpha_{i-1}\beta_{i-1} = 0$, then individuals of genotype $[\alpha, \beta]$ have a single copy of allele b_i at the i^{th} locus, and the daisy drive allele at the $(i-1)^{\text{th}}$ locus cuts the wild-type allele at the i^{th} locus. Homing then occurs with probability H, and gametes with allele 0 at position i are created.

The prefactor $F^{1-\alpha_n\beta_n}$ is the fitness cost associated with the payload. It appears if there is at least one copy of the daisy drive allele at the last position, n, in the daisy chain.

The selection dynamics for an n-element daisy drive are modeled by the following equations:

$$\dot{x}_{ab} = \sum_{\alpha} g_\alpha \sum_{\beta} g_\beta \prod_{i=1}^{n} [\delta_{a_i b_i}\delta_{\alpha_i a_i}\delta_{\beta_i b_i} + (1-\delta_{a_i b_i})(1-\delta_{\alpha_i \beta_i})] - \psi^2 x_{ab} \tag{3.2}$$

In Equations (3.2), the haplotypes α and β are summed independently. There is one such equation for each possible genotype $[a, b]$.

We make sense of Equations (3.2) as follows. Each pair of gametes g_α and g_β makes a new individual.

- If $a_i = b_i = \alpha_i = \beta_i$, then gametes of haplotypes α and β pair to make only individuals with genotype $[a_i, b_i]$ at locus i.
- If $a_i \neq b_i$ and $\alpha_i \neq \beta_i$, then gametes of haplotypes α and β pair to make only individuals with genotype $[a_i, b_i]$ at locus i.

We impose the density constraint

$$\sum_{a,b} x_{ab} = 1 \tag{3.3}$$

As already noted for Equations (3.1), in the sum over a, b when enumerating genotypes, heterozygous loci ($a_i \neq b_i$) are each counted once, so there is no double-counting. We use the following identity:

$$\sum_{a,b} \prod_{i=1}^{n} [\delta_{a_i b_i}\delta_{\alpha_i a_i}\delta_{\beta_i b_i} + (1-\delta_{a_i b_i})(1-\delta_{\alpha_i \beta_i})] = 1$$

The form of ψ that enforces the density constraint is

$$\psi = \sum_{\alpha} g_\alpha \tag{3.4}$$

3.9.2 Evolutionary dynamics of daisy drive resistance

Thus far in Section 3.9, we have assumed that there are exactly two alleles at each daisy drive locus: the daisy drive element, D, and the corresponding wild-type, W. However, additional alleles could arise in various ways: standing genetic variation, de novo mutation, or misrepair after cutting could all result in alleles with mismatches between the engineered guide RNAs and their corresponding recognition sequences. Such alleles would be *resistant* to the future effects of daisy-mediated cutting.

Our previous consideration of only two classes of allele was motivated by our presumed biological design: each daisy element was to target a highly conserved essential gene using multiple guide RNAs, and the corresponding daisy drive construct was to contain a genetically recoded copy of the target gene. Under these assumptions, we would expect low rates of standing genetic variation and *de novo* mutation, and targets resulting from misrepair would almost certainly produce nonviable offspring.

However, these assumptions are fairly restrictive. It could be difficult, in practice, to locate highly conserved regions, recode essential genes, and design multiple guide RNAs for every daisy element in a large chain, particularly in time-sensitive situations, such as responding to release of a rogue drive. Thus, in this section, we relax these earlier assumptions by extending our model to account for drive-resistant alleles.

3.9.2.1 Model for a 2-element daisy drive with resistance

We begin by considering the special case of two daisy drive elements, as in Section 3.9.1.1 above. The relevant loci are denoted 1 and 2 as before. Now, however, there are three alleles: the wild-type, W, the drive element, D, and a resistant allele, R, which is immune to the effects of the drive. We assume that resistant alleles primarily arise as the result of misrepair following cutting events (standing genetic variation could be accounted for by simply varying the initial frequency of the R allele). Because only the second locus is acted upon by the drive, we ignore resistance at the first locus.

Now, we consider the case where there is at least one drive element at the first locus (e.g., an individual with genotype 1_{WD} or 1_{DD}). Then there are six cases, depending on the genotype at the second locus:

- WW: The drive element cuts at both W alleles until both are resistant to further cutting. The individual thus converts to genotype 2_{RR} at this locus, and all gametes contain the 2_R allele.
- WD: The drive element cuts at the W allele. Subsequent repair occurs by homologous recombination with probability H, or by nonhomologous end-joining with probability $1 - H$. In the former case, the individual converts to genotype 2_{DD} and all gametes have the 2_D allele. In the latter case, the individual converts to 2_{DR} and produces gametes with 2_D or 2_R alleles with equal proportions.
- WR: The drive element cuts at the W allele. Subsequent repair by either repair pathway results in a resistant allele, so the individual converts to genotype 2_{RR}. Thus, all gametes produced contain the 2_R allele.
- DD: No cutting occurs, so all gametes contain the 2_D allele.
- DR: No cutting occurs, so gametes are produced containing the 2_D or 2_R allele with equal proportions.

Genotype	$1_W 2_W$	$1_W 2_D$	$1_W 2_R$	$1_D 2_W$	$1_D 2_D$	$1_D 2_R$	Fitness
$1_{WW}2_{WW}$	1	0	0	0	0	0	1
$1_{WW}2_{WD}$	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0	0	F
$1_{WW}2_{WR}$	$\frac{1}{2}$	0	$\frac{1}{2}$	0	0	0	K
$1_{WW}2_{DD}$	0	1	0	0	0	0	F
$1_{WW}2_{DR}$	0	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0	FK
$1_{WW}2_{RR}$	0	0	1	0	0	0	K
$1_{WD}2_{WW}$	0	0	$\frac{1}{2}$	0	0	$\frac{1}{2}$	G
$1_{WD}2_{WD}$	0	$\frac{1+H}{4}$	$\frac{1-H}{4}$	0	$\frac{1+H}{4}$	$\frac{1-H}{4}$	FG
$1_{WD}2_{WR}$	0	0	$\frac{1}{2}$	0	0	$\frac{1}{2}$	GK
$1_{WD}2_{DD}$	0	$\frac{1}{2}$	0	0	$\frac{1}{2}$	0	FG
$1_{WD}2_{DR}$	0	$\frac{1}{4}$	$\frac{1}{4}$	0	$\frac{1}{4}$	$\frac{1}{4}$	FGK
$1_{WD}2_{RR}$	0	0	$\frac{1}{2}$	0	0	$\frac{1}{2}$	GK
$1_{DD}2_{WW}$	0	0	0	0	0	1	G
$1_{DD}2_{WD}$	0	0	0	0	$\frac{1+H}{2}$	$\frac{1-H}{2}$	FG
$1_{DD}2_{WR}$	0	0	0	0	0	1	GK
$1_{DD}2_{DD}$	0	0	0	0	1	0	FG
$1_{DD}2_{DR}$	0	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$	FGK
$1_{DD}2_{RR}$	0	0	0	0	0	1	GK

Table 3.2: Gamete production probabilities and genotype fitnesses for two-element daisy drive with resistant alleles.

- RR: No cutting occurs, so all gametes contain the 2_R allele.

The cases above describe the production probabilities of the various alleles. But what are their effects on fitness? We assume that the payload element, 2_D, confers a dominant fitness cost, c; the upstream drive element, 1_D, confers a dominant fitness cost, d; and the resistant allele confers a dominant fitness cost, s. We assume that the all-wild-type individual has maximum fitness 1, so that $0 \leq c, d, s \leq 1$. We then define the shorthand notation $F = 1 - c$, $G = 1 - d$, and $K = 1 - s$. These assumptions are summarized in Table 3.2.

Using these rules, we can formally express the rates at which the six types of gametes are produced in the population. We denote by $g(z)$ the rate (with implicit time-dependence) at which gametes with haplotype z are produced by individuals in the population.

$$\begin{aligned}
g(1_W 2_W) &= x(1_{WW} 2_{WW}) + \frac{1}{2} F x(1_{WW} 2_{WD}) + \frac{1}{2} K x(1_{WW} 2_{WR}) \\
g(1_W 2_D) &= \frac{1}{2} F x(1_{WW} 2_{WD}) + F x(1_{WW} 2_{DD}) + \frac{1}{2} F K x(1_{WW} 2_{DR}) + \frac{1+H}{4} F G x(1_{WD} 2_{WD}) \\
&\quad + \frac{1}{2} F G x(1_{WD} 2_{DD}) + \frac{1}{4} F G K x(1_{WD} 2_{DR}) \\
g(1_W 2_R) &= \frac{1}{2} K x(1_{WW} 2_{WR}) + \frac{1}{2} F K x(1_{WW} 2_{DR}) + K x(1_{WW} 2_{RR}) + \frac{1}{2} G x(1_{WD} 2_{WW}) \\
&\quad + \frac{1-H}{4} F G x(1_{WD} 2_{WD}) + \frac{1}{2} G K x(1_{WD} 2_{WR}) + \frac{1}{4} F G K x(1_{WD} 2_{DR}) \\
&\quad + \frac{1}{2} G K x(1_{WD} 2_{RR}) \\
g(1_D 2_W) &= 0 \\
g(1_D 2_D) &= \frac{1+H}{4} F G x(1_{WD} 2_{WD}) + \frac{1}{2} F G x(1_{WD} 2_{DD}) + \frac{1}{4} F G K x(1_{WD} 2_{DR}) \\
&\quad + \frac{1+H}{2} F G x(1_{DD} 2_{WD}) + F G x(1_{DD} 2_{DD}) + \frac{1}{2} F G K x(1_{DD} 2_{DR}) \\
g(1_D 2_R) &= \frac{1}{2} G x(1_{WD} 2_{WW}) + \frac{1-H}{4} F G x(1_{WD} 2_{WD}) + \frac{1}{2} G K x(1_{WD} 2_{WR}) \\
&\quad + \frac{1}{4} F G K x(1_{WD} 2_{DR}) + \frac{1}{2} G K x(1_{WD} 2_{RR}) + G x(1_{DD} 2_{WW}) \\
&\quad + \frac{1-H}{2} F G x(1_{DD} 2_{WD}) + G K x(1_{DD} 2_{WR}) + \frac{1}{2} F G K x(1_{DD} 2_{DR}) \\
&\quad + G K x(1_{DD} 2_{RR})
\end{aligned}$$

Here, $x(z)$ is the frequency of individuals with genotype z.

The selection dynamics are then modeled by the following system of equations:

$$
\begin{aligned}
\dot{x}(1_{WW}2_{WW}) &= g(1_W2_W)^2 - \psi^2 x(1_{WW}2_{WW})\\
\dot{x}(1_{WW}2_{WD}) &= 2g(1_W2_W)g(1_W2_D) - \psi^2 x(1_{WW}2_{WD})\\
\dot{x}(1_{WW}2_{WR}) &= 2g(1_W2_W)g(1_W2_R) - \psi^2 x(1_{WW}2_{WR})\\
\dot{x}(1_{WW}2_{DD}) &= g(1_W2_D)^2 - \psi^2 x(1_{WW}2_{DD})\\
\dot{x}(1_{WW}2_{DR}) &= 2g(1_W2_D)g(1_W2_R) - \psi^2 x(1_{WW}2_{DR})\\
\dot{x}(1_{WW}2_{RR}) &= g(1_W2_R)^2 - \psi^2 x(1_{WW}2_{RR})\\
\dot{x}(1_{WD}2_{WW}) &= 2g(1_W2_W)g(1_D2_W) - \psi^2 x(1_{WD}2_{WW})\\
\dot{x}(1_{WD}2_{WD}) &= 2g(1_W2_D)g(1_D2_W) + 2g(1_W2_W)g(1_D2_D) - \psi^2 x(1_{WD}2_{WD})\\
\dot{x}(1_{WD}2_{WR}) &= 2g(1_W2_R)g(1_D2_W) + 2g(1_W2_W)g(1_D2_R) - \psi^2 x(1_{WD}2_{WR})\\
\dot{x}(1_{WD}2_{DD}) &= 2g(1_W2_D)g(1_D2_D) - \psi^2 x(1_{WD}2_{DD})\\
\dot{x}(1_{WD}2_{DR}) &= 2g(1_W2_D)g(1_D2_R) + 2g(1_W2_R)g(1_D2_D) - \psi^2 x(1_{WD}2_{DR})\\
\dot{x}(1_{WD}2_{RR}) &= 2g(1_W2_R)g(1_D2_R) - \psi^2 x(1_{WD}2_{RR})\\
\dot{x}(1_{DD}2_{WW}) &= g(1_D2_W)^2 - \psi^2 x(1_{DD}2_{WW})\\
\dot{x}(1_{DD}2_{WD}) &= 2g(1_D2_W)g(1_D2_D) - \psi^2 x(1_{DD}2_{WD})\\
\dot{x}(1_{DD}2_{WR}) &= 2g(1_D2_W)g(1_D2_R) - \psi^2 x(1_{DD}2_{WR})\\
\dot{x}(1_{DD}2_{DD}) &= g(1_D2_D)^2 - \psi^2 x(1_{DD}2_{DD})\\
\dot{x}(1_{DD}2_{DR}) &= 2g(1_D2_D)g(1_D2_R) - \psi^2 x(1_{DD}2_{DR})\\
\dot{x}(1_{DD}2_{RR}) &= g(1_D2_R)^2 - \psi^2 x(1_{DD}2_{RR})
\end{aligned}
$$

Note that this formulation assumes random mating as before, i.e., that two random gametes come together to form an individual. Also note that products $g(y)g(z)$ represent the pairings of different gametes. At any given time, we require that the total number of individuals sums to one:

$$\sum_z x(z) = 1$$

To enforce this density constraint, we set

$$\psi = g(1_W 2_W) + g(1_W 2_D) + g(1_W 2_R) + g(1_D 2_W) + g(1_D 2_D) + g(1_D 2_R)$$

3.9.2.2 Model for an n-element daisy drive with resistance

As in Section 3.9.1.2 above, we now apply the same concept to a daisy drive chain of arbitrary length, n. To describe this mathematically, we return to and amend our previous notation for an n-element system.

Consider a daisy drive construct with only two loci, as in Section 3.9.2.1. We use "W" to denote a wild-type allele, "D" to denote a daisy drive allele, and "R" to denote a resistant allele. To represent genotypes, we introduce vectors $a = (a_1, a_2)$ and $b = (b_1, b_2)$, where each $a_1, a_2, b_1, b_2 \in \{W, D, R\}$. We construct these vectors such that a_1 and b_1 represent the two alleles at the first locus, while a_2 and b_2 represent the two alleles at the second locus. A full genotype is then a list of the two vectors, $[a, b]$.

Below are a few examples of this naming convention applied to the genotypes of the two-element system:

$$1_{WW}2_{WW} = [(W, W), (W, W)]$$

$$1_{WW}2_{WD} = [(W, W), (W, D)]$$

$$1_{WW}2_{DD} = [(W, D), (W, D)]$$

$$1_{WW}2_{DR} = [(W, D), (W, R)]$$

$$1_{WD}2_{WW} = [(W, W), (D, W)]$$

$$1_{WD}2_{WD} = [(W, W), (D, D)]$$

To consider daisy drives of arbitrary length, we extend the lengths of the vectors a and b to be equal to the number of loci, n. That is, we let $a = (a_1, \ldots, a_n)$ and $b = (b_1, \ldots, b_n)$, where each $a_i, b_j \in \{W, D, R\}$. Again, notice that if an individual is heterozygous at a particular locus, then this notation allows for two ways of writing the alleles at that locus. For example, the genotype $1_{DD}2_{RR}3_{DR}$ can be written $[a, b] = [(D, R, D), (D, R, R)]$ or, equivalently, $[a, b] = [(D, R, R), (D, R, D)]$.

We denote by x_{ab} the frequency of individuals with genotype $[a, b]$. We denote by g_b the rate at which gametes with haplotype b are produced. For an n-element daisy drive, g_b is given by

$$g_b = \sum_{\alpha,\beta} x_{\alpha\beta} f(\alpha, \beta) p_{\alpha,\beta}(b) \tag{3.5}$$

Here we have used shorthand notation: $f(\alpha, \beta)$ is the fitness of an individual with genotype $[\alpha, \beta]$, and $p_{\alpha,\beta}(b)$ is the probability that an individual with genotype $[\alpha, \beta]$ produces a gamete with haplotype b. Notice that this is the same form as our Equations (3.1) above, with the fitness and gamete production components clearly identified.

The fitness of an $[\alpha, \beta]$ individual, $f(\alpha, \beta)$, is given by:

$$f(\alpha,\beta) = \prod_{i=1}^{n} F_i^{1-(1-\delta_{\alpha_i,D})(1-\delta_{\beta_i,D})} K_i^{1-(1-\delta_{\alpha_i,R})(1-\delta_{\beta_i,R})} \tag{3.6}$$

Here, $F_i = 1 - c_i$, where c_i is the fitness cost associated with the ith daisy drive element. Similarly, $K_i = 1 - s_i$, where s_i is the fitness cost of resistance at the ith position. δ_{ij} is the Kronecker delta, defined by $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$. This formulation assumes dominance of each fitness cost and mutual independence of all costs, as in the two-element system in Section 3.9.2.1 above.

Although the above formulation allows us to assign arbitrary costs at each position, we make the following simplifying assumptions in our simulations:

- The cost of resistance at upstream (non-payload) elements is zero: $K_1 = \cdots = K_{n-1} = 1$.
- All upstream (non-payload) drive elements have identical associated fitness costs: $F_1 = \cdots = F_{n-1} = 1 - d$.
- We define a cost, s, associated with resistance to the payload element: $K_n = 1 - s$.
- We define a cost, c, associated with the payload element itself: $F_n = 1 - c$.

Then, the probability, $p_{\alpha,\beta}(b)$, of an $[\alpha,\beta]$ individual producing gamete b is given by:

$$\begin{aligned} p_{\alpha,\beta}(b) = \prod_{i=1}^{n} \Bigg\{ & \left(1-\gamma^{D}_{\alpha_{i-1},\beta_{i-1}}(0)\right) \\ & \times \Big[\delta_{b_i,R}\gamma^{W}_{\alpha_i,\beta_i}(2) + \delta_{b_i,R}\gamma^{R}_{\alpha_i,\beta_i}(1)\gamma^{W}_{\alpha_i,\beta_i}(1) \\ & \quad + \delta_{b_i,R}\gamma^{R}_{\alpha_i,\beta_i}(2) + \frac{1}{2}\delta_{b_i,R}\gamma^{R}_{\alpha_i,\beta_i}(1)\gamma^{D}_{\alpha_i,\beta_i}(1) + \frac{1-H}{2}\delta_{b_i,R}\gamma^{W}_{\alpha_i,\beta_i}(1)\gamma^{D}_{\alpha_i,\beta_i}(1) \\ & \quad + \delta_{b_i,D}\gamma^{D}_{\alpha_i,\beta_i}(2) + \frac{1}{2}\delta_{b_i,D}\gamma^{D}_{\alpha_i,\beta_i}(1)\gamma^{R}_{\alpha_i,\beta_i}(1) + \frac{1+H}{2}\delta_{b_i,D}\gamma^{D}_{\alpha_i,\beta_i}(1)\gamma^{W}_{\alpha_i,\beta_i}(1)\Big] \\ & + \gamma^{D}_{\alpha_{i-1},\beta_{i-1}}(0) \\ & \times \Big[\delta_{b_i,W}\gamma^{W}_{\alpha_i,\beta_i}(2) + \frac{1}{2}\delta_{b_i,W}\gamma^{W}_{\alpha_i,\beta_i}(1)\gamma^{D}_{\alpha_i,\beta_i}(1) + \frac{1}{2}\delta_{b_i,W}\gamma^{W}_{\alpha_i,\beta_i}(1)\gamma^{R}_{\alpha_i,\beta_i}(1) \\ & \quad + \delta_{b_i,D}\gamma^{D}_{\alpha_i,\beta_i}(2) + \frac{1}{2}\delta_{b_i,D}\gamma^{D}_{\alpha_i,\beta_i}(1)\gamma^{R}_{\alpha_i,\beta_i}(1) + \frac{1}{2}\delta_{b_i,D}\gamma^{D}_{\alpha_i,\beta_i}(1)\gamma^{W}_{\alpha_i,\beta_i}(1) \\ & \quad + \delta_{b_i,R}\gamma^{R}_{\alpha_i,\beta_i}(2) + \frac{1}{2}\delta_{b_i,R}\gamma^{R}_{\alpha_i,\beta_i}(1)\gamma^{W}_{\alpha_i,\beta_i}(1) + \frac{1}{2}\delta_{b_i,R}\gamma^{R}_{\alpha_i,\beta_i}(1)\gamma^{D}_{\alpha_i,\beta_i}(1)\Big]\Bigg\} \end{aligned} \tag{3.7}$$

Here, we use shorthand notation, $\gamma^{c}_{\alpha_i,\beta_i}(k)$, to count the number of a particular allele at a particular locus: we define $\gamma^{c}_{\alpha_i,\beta_i}(k) = 1$ if there are k copies ($k = 0,1,2$) of allele c ($c \in \{W,D,R\}$) at position i in an individual with genotype $[\alpha,\beta]$. Otherwise, $\gamma^{c}_{\alpha_i,\beta_i}(k) = 0$. This is given by:

$$\gamma^{c}_{\alpha_i,\beta_i}(k) = \delta_{k,0}\left[(1-\delta_{\alpha_i,c})(1-\delta_{\beta_i,c})\right] + \delta_{k,1}\left[\delta_{\alpha_i,c}(1-\delta_{\beta_i,c}) + \delta_{\beta_i,c}(1-\delta_{\alpha_i,c})\right] + \delta_{k,2}\left[\delta_{\alpha_i,c}\delta_{\beta_i,c}\right].$$

For example, $\gamma^{W}_{\alpha_i,\beta_i}(2) = 1$ if there are two copies of a wild-type allele at position i in an $[\alpha,\beta]$ individual; otherwise $\gamma^{W}_{\alpha_i,\beta_i}(2) = 0$. We also define $\alpha_0 = \beta_0 = W$.

We understand Equations (3.7) as follows. Inheritance at each locus is independent, so the total probability $p_{\alpha,\beta}(b)$ is the product of inheritance probabilities at each individual position. Consider locus i. There are two possibilities. Either there is a daisy drive allele at the previous locus, which entails $\gamma^D_{\alpha_{i-1},\beta_{i-1}}(0) = 0$. (This eliminates the sum in the second pair of square brackets.) Or there is no daisy drive allele at the previous locus, which entails $\gamma^D_{\alpha_{i-1},\beta_{i-1}}(0) = 1$. (This eliminates the sum in the first pair of square brackets.)

If there is a daisy drive allele at the previous locus, then the value of the factor in the product of Equations (3.7) depends on the genotype at the current locus:

- $(\alpha_i, \beta_i) = (W, W)$. This entails $\gamma^W_{\alpha_i,\beta_i}(2) = 1$. Only R alleles are produced at locus i. Thus, the factor is 1 if $\delta_{b_i,R} = 1$. Otherwise, it is zero.
- $(\alpha_i, \beta_i) = (W, D)$. This entails $\gamma^W_{\alpha_i,\beta_i}(1)\gamma^D_{\alpha_i,\beta_i}(1) = 1$. By the action of the drive, D alleles are produced at locus i with probability $(1 + H)/2$, or R alleles are produced at locus i with probability $(1 - H)/2$. So if $\delta_{b_i,D} = 1$, then the factor is $(1 + H)/2$. Or if $\delta_{b_i,R} = 1$, then the factor is $(1 - H)/2$. Otherwise, it is zero.
- $(\alpha_i, \beta_i) = (W, R)$. This entails $\gamma^W_{\alpha_i,\beta_i}(1)\gamma^R_{\alpha_i,\beta_i}(1) = 1$. Only R alleles are produced at locus i. Thus, the factor is 1 if $\delta_{b_i,R} = 1$. Otherwise, it is zero.
- $(\alpha_i, \beta_i) = (D, D)$. This entails $\gamma^D_{\alpha_i,\beta_i}(2) = 1$. Only D alleles are produced at locus i. Thus, the factor is 1 if $\delta_{b_i,D} = 1$. Otherwise, it is zero.
- $(\alpha_i, \beta_i) = (D, R)$. This entails $\gamma^D_{\alpha_i,\beta_i}(1)\gamma^R_{\alpha_i,\beta_i}(1) = 1$. Here, D and R alleles are produced at locus i in equal proportions. Thus, the factor is $1/2$ if $\delta_{b_i,D} = 1$ or if $\delta_{b_i,R} = 1$. Otherwise, it is zero.
- $(\alpha_i, \beta_i) = (R, R)$. This entails $\gamma^R_{\alpha_i,\beta_i}(2) = 1$. Only R alleles are produced at locus i. Thus, the factor is 1 if $\delta_{b_i,R} = 1$. Otherwise, it is zero.

Similarly, if there is no daisy drive allele at the previous locus, then the value of the factor in the product of Equations (3.7) depends on the genotype at the current locus. However, because there is no drive, the inheritance probabilities are simply Mendelian:

- $(\alpha_i, \beta_i) = (W, W)$. This entails $\gamma^W_{\alpha_i,\beta_i}(2) = 1$. Only W alleles are produced at locus i. Thus, the factor is 1 if $\delta_{b_i,W} = 1$. Otherwise, it is zero.
- $(\alpha_i, \beta_i) = (W, D)$. This entails $\gamma^W_{\alpha_i,\beta_i}(1)\gamma^D_{\alpha_i,\beta_i}(1) = 1$. There is no drive action, so W alleles and D alleles are produced at locus i in equal proportions. Thus, if $\delta_{b_i,W} = 1$ or $\delta_{b_i,D} = 1$, then the factor is $1/2$. Otherwise, it is zero.
- $(\alpha_i, \beta_i) = (W, R)$. This entails $\gamma^W_{\alpha_i,\beta_i}(1)\gamma^R_{\alpha_i,\beta_i}(1) = 1$. Here, W alleles and R alleles are produced at locus i in equal proportions. Thus, if $\delta_{b_i,W} = 1$ or $\delta_{b_i,R} = 1$, then the factor is $1/2$. Otherwise, it is zero.
- $(\alpha_i, \beta_i) = (D, D)$. This entails $\gamma^D_{\alpha_i,\beta_i}(2) = 1$. Only D alleles are produced at locus i. Thus, the factor is 1 if $\delta_{b_i,D} = 1$. Otherwise, it is zero.
- $(\alpha_i, \beta_i) = (D, R)$. This entails $\gamma^D_{\alpha_i,\beta_i}(1)\gamma^R_{\alpha_i,\beta_i}(1) = 1$. Here, D alleles and R alleles are produced at locus i in equal proportions. Thus, the factor is $1/2$ if $\delta_{b_i,D} = 1$ or $\delta_{b_i,R} = 1$. Otherwise, it is zero.
- $(\alpha_i, \beta_i) = (R, R)$. This entails $\gamma^R_{\alpha_i,\beta_i}(2) = 1$. Only R alleles are produced at locus i. Thus, the factor is 1 if $\delta_{b_i,R} = 1$. Otherwise, it is zero.

The selection dynamics for an n-element daisy drive are then modeled by the following equations:

$$\dot{x}_{ab} = \sum_\alpha g_\alpha \sum_\beta g_\beta \prod_{i=1}^n \delta^{\alpha_i\beta_i}_{a_ib_i} - \psi^2 x_{ab} \tag{3.8}$$

Here, as shorthand notation, we define

$$\begin{aligned}\delta^{\alpha_i\beta_i}_{a_ib_i} &= \delta_{a_ib_i}\delta_{\alpha_ia_i}\delta_{\beta_ib_i}\\ &\quad + \gamma^W_{a_i,b_i}(1)\gamma^D_{a_i,b_i}(1)\gamma^W_{\alpha_i,\beta_i}(1)\gamma^D_{\alpha_i,\beta_i}(1)\\ &\quad + \gamma^W_{a_i,b_i}(1)\gamma^R_{a_i,b_i}(1)\gamma^W_{\alpha_i,\beta_i}(1)\gamma^R_{\alpha_i,\beta_i}(1)\\ &\quad + \gamma^D_{a_i,b_i}(1)\gamma^R_{a_i,b_i}(1)\gamma^D_{\alpha_i,\beta_i}(1)\gamma^R_{\alpha_i,\beta_i}(1)\end{aligned}$$

In Equations (3.8), the haplotypes α and β are summed independently. There is one such equation for each possible genotype $[a, b]$.

We impose the density constraint

$$\sum_{a,b} x_{ab} = 1. \tag{3.9}$$

We use the following identity:

$$\sum_{a,b} \prod_{i=1}^{n} \delta_{a_i b_i}^{\alpha_i \beta_i} = 1$$

And, as before, the form of ψ that enforces the density constraint is

$$\psi = \sum_{\alpha} g_{\alpha}. \tag{3.10}$$

3.9.2.3 Continuous release

To model a continuous release of individuals carrying the daisy drive construct into a population, we use the following equations:

$$\dot{x}_{ab} = \sum_{\alpha} g_{\alpha} \sum_{\beta} g_{\beta} \prod_{i=1}^{n} \delta_{a_i b_i}^{\alpha_i \beta_i} + C_{ab} - \left(\psi^2 + \sum_{\alpha,\beta} C_{\alpha\beta} \right) x_{ab} \tag{3.11}$$

A nonzero value of C_{ab} models a flow of individuals of genotype $[a, b]$ into the population. Equations (3.11) are thus a generalization of Equations (3.8). ψ is given by Equation (3.10), and the density constraint, Equation (3.9), holds at all times.

3.9.3 Two-population model for an n-element daisy drive with resistance

We now extend the model from Section 3.9.2.3 to include a simple spatial component: two populations connected by gene flow.

3.9.3.1 Two-population model without gene flow

First, we consider two populations whose evolutionary dynamics are decoupled. We denote by x_{ab} the frequency of individuals with genotype $[a, b]$ among individuals in the target population, and we denote by y_{ab} the frequency of individuals with genotype $[a, b]$ among individuals in the mainland population. We denote by $g_b^{(T)}$ the rate at which gametes with haplotype b are produced in the target population, and we denote by $g_b^{(M)}$ the same for the mainland population. For an n-element daisy drive, $g_b^{(T)}$ and $g_b^{(M)}$ are given by

$$\begin{aligned} g_b^{(T)} &= \sum_{\alpha,\beta} x_{\alpha\beta} f(\alpha, \beta) p_{\alpha,\beta}(b) \\ g_b^{(M)} &= \sum_{\alpha,\beta} y_{\alpha\beta} f(\alpha, \beta) p_{\alpha,\beta}(b) \end{aligned} \tag{3.12}$$

Here, $f(\alpha, \beta)$ is the fitness of the genotype $[\alpha, \beta]$, and $p_{\alpha,\beta}(b)$ is the probability that an individual of genotype $[\alpha, \beta]$ produces a gamete with haplotype b. These two quantities are given by Equations (3.6) and (3.7), respectively.

Equations (3.12) are essentially identical to Equations (3.5), except we assume that only individuals in the target population contribute to the target population gamete pool and similarly for the mainland. Thus, the difference between Equations (3.12) and Equations (3.5) arises from the separation of the two populations via $g_b^{(T)}$, $g_b^{(M)}$, $x_{\alpha\beta}$, and $y_{\alpha\beta}$.

The selection dynamics for an n-element daisy drive system in two populations are then modeled by the following equations:

$$\dot{x}_{ab} = \sum_{\alpha} g_{\alpha}^{(T)} \sum_{\beta} g_{\beta}^{(T)} \prod_{i=1}^{n} \delta_{a_i b_i}^{\alpha_i \beta_i} + C_{ab}^{(T)} - \left(\left(\psi^{(T)} \right)^2 + \sum_{\alpha,\beta} C_{\alpha\beta}^{(T)} \right) x_{ab}$$

$$\dot{y}_{ab} = \sum_{\alpha} g_{\alpha}^{(M)} \sum_{\beta} g_{\beta}^{(M)} \prod_{i=1}^{n} \delta_{a_i b_i}^{\alpha_i \beta_i} + C_{ab}^{(M)} - \left(\left(\psi^{(M)} \right)^2 + \sum_{\alpha,\beta} C_{\alpha\beta}^{(M)} \right) y_{ab}$$

Notice that each population experiences selection dynamics identical to the single-population model given by Equations (3.11). A nonzero value of $C_{ab}^{(T)}$ models a flow of individuals of genotype $[a, b]$ into the target population, and a nonzero value of $C_{ab}^{(M)}$ models a flow of individuals of genotype $[a, b]$ into the mainland population.

The density constraints are

$$\sum_{a,b} x_{ab} = 1$$

$$\sum_{a,b} y_{ab} = 1$$

To enforce these density constraints, we set

$$\psi^{(T)} = \sum_{\alpha} g_{\alpha}^{(T)}$$

$$\psi^{(M)} = \sum_{\alpha} g_{\alpha}^{(M)}$$

3.9.3.2 Two-population model with gene flow

Next, we assume that there is a nonzero rate of migration of individuals from the target population to the mainland population and vice versa. For notational clarity, we define new frequency variables. We denote by X_{ab} (with an uppercase X) the frequency of individuals with genotype $[a, b]$ among individuals in the target population when there is migration, and we denote by Y_{ab} (with an uppercase Y) the frequency of individuals with genotype $[a, b]$ among individuals in the mainland population when there is migration. We denote by $G_b^{(T)}$ (with an uppercase G) the rate at which gametes with haplotype b are produced in the target population when there is migration, and we denote by $G_b^{(M)}$ (with an uppercase G) the same for the mainland population when there is migration. $G_b^{(T)}$ and $G_b^{(M)}$ are given by

$$\begin{aligned} G_b^{(T)} &= \sum_{\alpha,\beta} X_{\alpha\beta} f(\alpha, \beta) p_{\alpha,\beta}(b) \\ G_b^{(M)} &= \sum_{\alpha,\beta} Y_{\alpha\beta} f(\alpha, \beta) p_{\alpha,\beta}(b) \end{aligned} \tag{3.13}$$

Here, $f(\alpha, \beta)$ is the fitness of the genotype $[\alpha, \beta]$, and $p_{\alpha,\beta}(b)$ is the probability that an individual of genotype $[\alpha, \beta]$ produces a gamete with haplotype b. These two quantities are given by Equations (3.6) and (3.7), respectively.

We assume that, over a given time interval, the number of individuals migrating in each direction is equal, so that the population sizes of the target and the mainland each remain constant. The rate of migration is quantified by the parameter r. We also denote by R the fraction of all individuals that are on the target. (Similarly, $1 - R$ is the fraction of all individuals that are on the mainland.) The selection dynamics for an n-element daisy drive system in two populations that are connected by gene flow are then modeled by the following equations:

$$\begin{aligned}\dot{X}_{ab} &= \sum_{\alpha} G_{\alpha}^{(T)} \sum_{\beta} G_{\beta}^{(T)} \prod_{i=1}^{n} \delta_{a_i b_i}^{\alpha_i \beta_i} + C_{ab}^{(T)} + \frac{r}{R}\left(Y_{ab} - X_{ab}\right) - \left(\left(\Psi^{(T)}\right)^2 + \sum_{\alpha,\beta} C_{\alpha\beta}^{(T)}\right) X_{ab} \\ \dot{Y}_{ab} &= \sum_{\alpha} G_{\alpha}^{(M)} \sum_{\beta} G_{\beta}^{(M)} \prod_{i=1}^{n} \delta_{a_i b_i}^{\alpha_i \beta_i} + C_{ab}^{(M)} + \frac{r}{1-R}\left(X_{ab} - Y_{ab}\right) - \left(\left(\Psi^{(M)}\right)^2 + \sum_{\alpha,\beta} C_{\alpha\beta}^{(M)}\right) Y_{ab}\end{aligned} \tag{3.14}$$

The density constraints are

$$\sum_{a,b} X_{ab} = 1$$

$$\sum_{a,b} Y_{ab} = 1$$

To enforce these density constraints, we set $\Psi^{(T)}$ (with an uppercase Ψ) and $\Psi^{(M)}$ (with an uppercase Ψ) to equal

$$\Psi^{(T)} = \sum_{\alpha} G_{\alpha}^{(T)}$$

$$\Psi^{(M)} = \sum_{\alpha} G_{\alpha}^{(M)}$$

3.9.4 N-population model for an n-element daisy drive with resistance

The above treatment is readily extended to a population that consists of N islands. Denote the frequency of individuals of genotype $[a, b]$ on island ℓ (for $1 \leq \ell \leq N$) as $X_{ab}^{(\ell)}$. Gametes with haplotype b are produced on island ℓ at rate $G_b^{(\ell)}$, where $G_b^{(\ell)}$ is given by

$$G_b^{(\ell)} = \sum_{\alpha,\beta} X_{\alpha\beta}^{(\ell)} f(\alpha,\beta) p_{\alpha,\beta}(b)$$

The rate of migration of individuals between islands ℓ and ω is quantified by the parameter $r_{\ell\omega} = r_{\omega\ell}$. The fraction of all individuals in the population that are on island ℓ is denoted by R_ℓ. The dynamics of $X_{ab}^{(\ell)}$ are given by

$$\dot{X}_{ab}^{(\ell)} = \sum_{\alpha} G_\alpha^{(\ell)} \sum_{\beta} G_\beta^{(\ell)} \prod_{i=1}^{n} \delta_{a_i b_i}^{\alpha_i \beta_i} + C_{ab}^{(\ell)} + \sum_{\substack{\omega=1 \\ \omega\neq\ell}}^{N} \frac{r_{\ell\omega}}{R_\ell} \left(X_{ab}^{(\omega)} - X_{ab}^{(\ell)} \right) - \left(\left(\Psi^{(\ell)} \right)^2 + \sum_{\alpha,\beta} C_{\alpha\beta}^{(\ell)} \right) X_{ab}^{(\ell)} \tag{3.15}$$

The density constraints are

$$\sum_{a,b} X_{ab}^{(\ell)} = 1$$

To enforce these density constraints, we set $\Psi^{(\ell)}$ (with an uppercase Ψ) to equal

$$\Psi^{(\ell)} = \sum_{\alpha} G_\alpha^{(\ell)}$$

3.9.5 Particular case: Daisy-chain versus self-propagating drives on five islands

It is instructive to contrast the evolutionary dynamics of a daisy-chain gene drive with a self-propagating gene drive, where in both cases the evolution occurs in a population consisting of five islands. For simplicity, we assume that individuals are only exchanged between nearby islands, i.e., there is gene flow between islands 1 and 2, between islands 2 and 3, between islands 3 and 4, and between islands 4 and 5. We further assume that these rates of gene flow are all equal, and we assume that each island has the same number of individuals.

In this section, we present the equations necessary to perform simulations of the evolutionary dynamics for each of these scenarios.

3.9.5.1 5-population model for an n-element daisy drive

For modeling the dynamics of a daisy-chain gene drive on five islands, we use Equations (3.15). Substituting $r_{12}/R_1 = r_{21}/R_2 = r_{23}/R_2 = r_{32}/R_3 = r_{34}/R_3 = r_{43}/R_4 = r_{45}/R_4 = r_{54}/R_5 = r$, and setting all other migration rates equal to zero, we obtain

$$\begin{aligned}
\dot{X}^{(1)}_{ab} &= \sum_{\alpha} G^{(1)}_{\alpha} \sum_{\beta} G^{(1)}_{\beta} \prod_{i=1}^{n} \delta^{\alpha_i \beta_i}_{a_i b_i} + C^{(1)}_{ab} + r\left(X^{(2)}_{ab} - X^{(1)}_{ab}\right) - \left(\left(\Psi^{(1)}\right)^2 + \sum_{\alpha,\beta} C^{(1)}_{\alpha\beta}\right) X^{(1)}_{ab} \\
\dot{X}^{(2)}_{ab} &= \sum_{\alpha} G^{(2)}_{\alpha} \sum_{\beta} G^{(2)}_{\beta} \prod_{i=1}^{n} \delta^{\alpha_i \beta_i}_{a_i b_i} + C^{(2)}_{ab} + r\left(X^{(3)}_{ab} + X^{(1)}_{ab} - 2X^{(2)}_{ab}\right) - \left(\left(\Psi^{(2)}\right)^2 + \sum_{\alpha,\beta} C^{(2)}_{\alpha\beta}\right) X^{(2)}_{ab} \\
\dot{X}^{(3)}_{ab} &= \sum_{\alpha} G^{(3)}_{\alpha} \sum_{\beta} G^{(3)}_{\beta} \prod_{i=1}^{n} \delta^{\alpha_i \beta_i}_{a_i b_i} + C^{(3)}_{ab} + r\left(X^{(4)}_{ab} + X^{(2)}_{ab} - 2X^{(3)}_{ab}\right) - \left(\left(\Psi^{(3)}\right)^2 + \sum_{\alpha,\beta} C^{(3)}_{\alpha\beta}\right) X^{(3)}_{ab} \\
\dot{X}^{(4)}_{ab} &= \sum_{\alpha} G^{(4)}_{\alpha} \sum_{\beta} G^{(4)}_{\beta} \prod_{i=1}^{n} \delta^{\alpha_i \beta_i}_{a_i b_i} + C^{(4)}_{ab} + r\left(X^{(5)}_{ab} + X^{(3)}_{ab} - 2X^{(4)}_{ab}\right) - \left(\left(\Psi^{(4)}\right)^2 + \sum_{\alpha,\beta} C^{(4)}_{\alpha\beta}\right) X^{(4)}_{ab} \\
\dot{X}^{(5)}_{ab} &= \sum_{\alpha} G^{(5)}_{\alpha} \sum_{\beta} G^{(5)}_{\beta} \prod_{i=1}^{n} \delta^{\alpha_i \beta_i}_{a_i b_i} + C^{(5)}_{ab} + r\left(X^{(4)}_{ab} - X^{(5)}_{ab}\right) - \left(\left(\Psi^{(5)}\right)^2 + \sum_{\alpha,\beta} C^{(5)}_{\alpha\beta}\right) X^{(5)}_{ab}
\end{aligned}$$

3.9.5.2 5-population model for a self-propagating drive

The equations for modeling the dynamics of a self-propagating gene drive on five islands are based on the model presented in Chapter 1 (and, correspondingly, in the Supplementary Materials for Ref. 39). For more details and descriptions, please see the writing therein.

For a self-propagating gene drive, consider that there are $\mathcal{N}$ guide RNAs. There are the drive allele, D, $\mathcal{N}$ "costly" resistant alleles, R_i (with $1 \leq i \leq \mathcal{N}$), $\mathcal{N}$ "neutral" resistant alleles, S_i (with $1 \leq i \leq \mathcal{N}$), and the wild-type allele, S_0.

We use $X_{ab}^{(\ell)}$ to denote the frequency of individuals of genotype $[a, b]$ on island ℓ. The rates at which each of the $2\mathcal{N} + 2$ types of gametes are produced on island ℓ are given by

$$\begin{aligned}
F_D^{(\ell)} &= f_{DD} X_{DD}^{(\ell)} + \sum_{k=1}^{\mathcal{N}} p_{R_k D, D} f_{R_k D} X_{R_k D}^{(\ell)} + \sum_{k=0}^{\mathcal{N}} p_{S_k D, D} f_{S_k D} X_{S_k D}^{(\ell)} \\
F_{S_i}^{(\ell)} &= \sum_{k=0}^{\mathcal{N}} \frac{1 + \delta_{ki}}{2} f_{S_k S_i} X_{S_k S_i}^{(\ell)} + \frac{1}{2} \sum_{k=1}^{\mathcal{N}} f_{R_k S_i} X_{R_k S_i}^{(\ell)} + \sum_{k=0}^{i} p_{S_k D, S_i} f_{S_k D} X_{S_k D}^{(\ell)} \\
F_{R_i}^{(\ell)} &= \sum_{k=1}^{\mathcal{N}} \frac{1 + \delta_{ki}}{2} f_{R_k R_i} X_{R_k R_i}^{(\ell)} + \frac{1}{2} \sum_{k=0}^{\mathcal{N}} f_{R_i S_k} X_{R_i S_k}^{(\ell)} \\
&\quad + \sum_{k=1}^{i} p_{R_k D, R_i} f_{R_k D} X_{R_k D}^{(\ell)} + \sum_{k=0}^{i-1} p_{S_k D, R_i} f_{S_k D} X_{S_k D}^{(\ell)}
\end{aligned}$$

From conservation of probability, we have

$$\begin{aligned}
p_{R_k D, D} + \sum_{i=k}^{\mathcal{N}} p_{R_k D, R_i} &= 1 \\
p_{S_k D, D} + \sum_{i=k}^{\mathcal{N}} p_{S_k D, S_i} + \sum_{i=k+1}^{\mathcal{N}} p_{S_k D, R_i} &= 1
\end{aligned}$$

Since type $R_{\mathcal{N}}D$ and type $S_{\mathcal{N}}D$ individuals are fully resistant to being manipulated by the drive construct, they show standard Mendelian segregation in their production of gametes, and we have

$$p_{R_{\mathcal{N}}D,R_{\mathcal{N}}} = p_{S_{\mathcal{N}}D,S_{\mathcal{N}}} = \frac{1}{2}$$

For modeling the dynamics of a self-propagating gene drive on five islands, we use the following equations:

The dynamics of individuals of genotype DD on each island are given by

$$\begin{aligned}
\dot{X}_{DD}^{(1)} &= \left(F_D^{(1)}\right)^2 + C_{DD}^{(1)} + r\left(X_{DD}^{(2)} - X_{DD}^{(1)}\right) - \left(\left(\Psi^{(1)}\right)^2 + \sum_{\alpha,\beta} C_{\alpha\beta}^{(1)}\right) X_{DD}^{(1)} \\
\dot{X}_{DD}^{(2)} &= \left(F_D^{(2)}\right)^2 + C_{DD}^{(2)} + r\left(X_{DD}^{(3)} + X_{DD}^{(1)} - 2X_{DD}^{(2)}\right) - \left(\left(\Psi^{(2)}\right)^2 + \sum_{\alpha,\beta} C_{\alpha\beta}^{(2)}\right) X_{DD}^{(2)} \\
\dot{X}_{DD}^{(3)} &= \left(F_D^{(3)}\right)^2 + C_{DD}^{(3)} + r\left(X_{DD}^{(4)} + X_{DD}^{(2)} - 2X_{DD}^{(3)}\right) - \left(\left(\Psi^{(3)}\right)^2 + \sum_{\alpha,\beta} C_{\alpha\beta}^{(3)}\right) X_{DD}^{(3)} \\
\dot{X}_{DD}^{(4)} &= \left(F_D^{(4)}\right)^2 + C_{DD}^{(4)} + r\left(X_{DD}^{(5)} + X_{DD}^{(3)} - 2X_{DD}^{(4)}\right) - \left(\left(\Psi^{(4)}\right)^2 + \sum_{\alpha,\beta} C_{\alpha\beta}^{(4)}\right) X_{DD}^{(4)} \\
\dot{X}_{DD}^{(5)} &= \left(F_D^{(5)}\right)^2 + C_{DD}^{(5)} + r\left(X_{DD}^{(4)} - X_{DD}^{(5)}\right) - \left(\left(\Psi^{(5)}\right)^2 + \sum_{\alpha,\beta} C_{\alpha\beta}^{(5)}\right) X_{DD}^{(5)}
\end{aligned}$$

The dynamics of individuals of genotype R_iD on each island are given by

$$\dot{X}^{(1)}_{R_iD} = 2F^{(1)}_{R_i}F^{(1)}_{D} + C^{(1)}_{R_iD} + r\left(X^{(2)}_{R_iD} - X^{(1)}_{R_iD}\right) - \left(\left(\Psi^{(1)}\right)^2 + \sum_{\alpha,\beta} C^{(1)}_{\alpha\beta}\right) X^{(1)}_{R_iD}$$

$$\dot{X}^{(2)}_{R_iD} = 2F^{(2)}_{R_i}F^{(2)}_{D} + C^{(2)}_{R_iD} + r\left(X^{(3)}_{R_iD} + X^{(1)}_{R_iD} - 2X^{(2)}_{R_iD}\right) - \left(\left(\Psi^{(2)}\right)^2 + \sum_{\alpha,\beta} C^{(2)}_{\alpha\beta}\right) X^{(2)}_{R_iD}$$

$$\dot{X}^{(3)}_{R_iD} = 2F^{(3)}_{R_i}F^{(3)}_{D} + C^{(3)}_{R_iD} + r\left(X^{(4)}_{R_iD} + X^{(2)}_{R_iD} - 2X^{(3)}_{R_iD}\right) - \left(\left(\Psi^{(3)}\right)^2 + \sum_{\alpha,\beta} C^{(3)}_{\alpha\beta}\right) X^{(3)}_{R_iD}$$

$$\dot{X}^{(4)}_{R_iD} = 2F^{(4)}_{R_i}F^{(4)}_{D} + C^{(4)}_{R_iD} + r\left(X^{(5)}_{R_iD} + X^{(3)}_{R_iD} - 2X^{(4)}_{R_iD}\right) - \left(\left(\Psi^{(4)}\right)^2 + \sum_{\alpha,\beta} C^{(4)}_{\alpha\beta}\right) X^{(4)}_{R_iD}$$

$$\dot{X}^{(5)}_{R_iD} = 2F^{(5)}_{R_i}F^{(5)}_{D} + C^{(5)}_{R_iD} + r\left(X^{(4)}_{R_iD} - X^{(5)}_{R_iD}\right) - \left(\left(\Psi^{(5)}\right)^2 + \sum_{\alpha,\beta} C^{(5)}_{\alpha\beta}\right) X^{(5)}_{R_iD}$$

The dynamics of individuals of genotype S_iD on each island are given by

$$\dot{X}^{(1)}_{S_iD} = 2F^{(1)}_{S_i}F^{(1)}_{D} + C^{(1)}_{S_iD} + r\left(X^{(2)}_{S_iD} - X^{(1)}_{S_iD}\right) - \left(\left(\Psi^{(1)}\right)^2 + \sum_{\alpha,\beta} C^{(1)}_{\alpha\beta}\right) X^{(1)}_{S_iD}$$

$$\dot{X}^{(2)}_{S_iD} = 2F^{(2)}_{S_i}F^{(2)}_{D} + C^{(2)}_{S_iD} + r\left(X^{(3)}_{S_iD} + X^{(1)}_{S_iD} - 2X^{(2)}_{S_iD}\right) - \left(\left(\Psi^{(2)}\right)^2 + \sum_{\alpha,\beta} C^{(2)}_{\alpha\beta}\right) X^{(2)}_{S_iD}$$

$$\dot{X}^{(3)}_{S_iD} = 2F^{(3)}_{S_i}F^{(3)}_{D} + C^{(3)}_{S_iD} + r\left(X^{(4)}_{S_iD} + X^{(2)}_{S_iD} - 2X^{(3)}_{S_iD}\right) - \left(\left(\Psi^{(3)}\right)^2 + \sum_{\alpha,\beta} C^{(3)}_{\alpha\beta}\right) X^{(3)}_{S_iD}$$

$$\dot{X}^{(4)}_{S_iD} = 2F^{(4)}_{S_i}F^{(4)}_{D} + C^{(4)}_{S_iD} + r\left(X^{(5)}_{S_iD} + X^{(3)}_{S_iD} - 2X^{(4)}_{S_iD}\right) - \left(\left(\Psi^{(4)}\right)^2 + \sum_{\alpha,\beta} C^{(4)}_{\alpha\beta}\right) X^{(4)}_{S_iD}$$

$$\dot{X}^{(5)}_{S_iD} = 2F^{(5)}_{S_i}F^{(5)}_{D} + C^{(5)}_{S_iD} + r\left(X^{(4)}_{S_iD} - X^{(5)}_{S_iD}\right) - \left(\left(\Psi^{(5)}\right)^2 + \sum_{\alpha,\beta} C^{(5)}_{\alpha\beta}\right) X^{(5)}_{S_iD}$$

The dynamics of individuals of genotype R_iS_j on each island are given by

$$\dot{X}^{(1)}_{R_iS_j} = 2F^{(1)}_{R_i}F^{(1)}_{S_j} + C^{(1)}_{R_iS_j} + r\left(X^{(2)}_{R_iS_j} - X^{(1)}_{R_iS_j}\right) - \left(\left(\Psi^{(1)}\right)^2 + \sum_{\alpha,\beta} C^{(1)}_{\alpha\beta}\right) X^{(1)}_{R_iS_j}$$

$$\dot{X}^{(2)}_{R_iS_j} = 2F^{(2)}_{R_i}F^{(2)}_{S_j} + C^{(2)}_{R_iS_j} + r\left(X^{(3)}_{R_iS_j} + X^{(1)}_{R_iS_j} - 2X^{(2)}_{R_iS_j}\right) - \left(\left(\Psi^{(2)}\right)^2 + \sum_{\alpha,\beta} C^{(2)}_{\alpha\beta}\right) X^{(2)}_{R_iS_j}$$

$$\dot{X}^{(3)}_{R_iS_j} = 2F^{(3)}_{R_i}F^{(3)}_{S_j} + C^{(3)}_{R_iS_j} + r\left(X^{(4)}_{R_iS_j} + X^{(2)}_{R_iS_j} - 2X^{(3)}_{R_iS_j}\right) - \left(\left(\Psi^{(3)}\right)^2 + \sum_{\alpha,\beta} C^{(3)}_{\alpha\beta}\right) X^{(3)}_{R_iS_j}$$

$$\dot{X}^{(4)}_{R_iS_j} = 2F^{(4)}_{R_i}F^{(4)}_{S_j} + C^{(4)}_{R_iS_j} + r\left(X^{(5)}_{R_iS_j} + X^{(3)}_{R_iS_j} - 2X^{(4)}_{R_iS_j}\right) - \left(\left(\Psi^{(4)}\right)^2 + \sum_{\alpha,\beta} C^{(4)}_{\alpha\beta}\right) X^{(4)}_{R_iS_j}$$

$$\dot{X}^{(5)}_{R_iS_j} = 2F^{(5)}_{R_i}F^{(5)}_{S_j} + C^{(5)}_{R_iS_j} + r\left(X^{(4)}_{R_iS_j} - X^{(5)}_{R_iS_j}\right) - \left(\left(\Psi^{(5)}\right)^2 + \sum_{\alpha,\beta} C^{(5)}_{\alpha\beta}\right) X^{(5)}_{R_iS_j}$$

The dynamics of individuals of genotype R_iR_j on each island are given by

$$\dot{X}^{(1)}_{R_iR_j} = (2-\delta_{ij})F^{(1)}_{R_i}F^{(1)}_{R_j} + C^{(1)}_{R_iR_j} + r\left(X^{(2)}_{R_iR_j} - X^{(1)}_{R_iR_j}\right) - \left(\left(\Psi^{(1)}\right)^2 + \sum_{\alpha,\beta} C^{(1)}_{\alpha\beta}\right) X^{(1)}_{R_iR_j}$$

$$\dot{X}^{(2)}_{R_iR_j} = (2-\delta_{ij})F^{(2)}_{R_i}F^{(2)}_{R_j} + C^{(2)}_{R_iR_j} + r\left(X^{(3)}_{R_iR_j} + X^{(1)}_{R_iR_j} - 2X^{(2)}_{R_iR_j}\right) - \left(\left(\Psi^{(2)}\right)^2 + \sum_{\alpha,\beta} C^{(2)}_{\alpha\beta}\right) X^{(2)}_{R_iR_j}$$

$$\dot{X}^{(3)}_{R_iR_j} = (2-\delta_{ij})F^{(3)}_{R_i}F^{(3)}_{R_j} + C^{(3)}_{R_iR_j} + r\left(X^{(4)}_{R_iR_j} + X^{(2)}_{R_iR_j} - 2X^{(3)}_{R_iR_j}\right) - \left(\left(\Psi^{(3)}\right)^2 + \sum_{\alpha,\beta} C^{(3)}_{\alpha\beta}\right) X^{(3)}_{R_iR_j}$$

$$\dot{X}^{(4)}_{R_iR_j} = (2-\delta_{ij})F^{(4)}_{R_i}F^{(4)}_{R_j} + C^{(4)}_{R_iR_j} + r\left(X^{(5)}_{R_iR_j} + X^{(3)}_{R_iR_j} - 2X^{(4)}_{R_iR_j}\right) - \left(\left(\Psi^{(4)}\right)^2 + \sum_{\alpha,\beta} C^{(4)}_{\alpha\beta}\right) X^{(4)}_{R_iR_j}$$

$$\dot{X}^{(5)}_{R_iR_j} = (2-\delta_{ij})F^{(5)}_{R_i}F^{(5)}_{R_j} + C^{(5)}_{R_iR_j} + r\left(X^{(4)}_{R_iR_j} - X^{(5)}_{R_iR_j}\right) - \left(\left(\Psi^{(5)}\right)^2 + \sum_{\alpha,\beta} C^{(5)}_{\alpha\beta}\right) X^{(5)}_{R_iR_j}$$

The dynamics of individuals of genotype S_iS_j on each island are given by

$$\dot{X}^{(1)}_{S_iS_j} = (2-\delta_{ij})F^{(1)}_{S_i}F^{(1)}_{S_j} + C^{(1)}_{S_iS_j} + r\left(X^{(2)}_{S_iS_j} - X^{(1)}_{S_iS_j}\right) - \left(\left(\Psi^{(1)}\right)^2 + \sum_{\alpha,\beta} C^{(1)}_{\alpha\beta}\right) X^{(1)}_{S_iS_j}$$

$$\dot{X}^{(2)}_{S_iS_j} = (2-\delta_{ij})F^{(2)}_{S_i}F^{(2)}_{S_j} + C^{(2)}_{S_iS_j} + r\left(X^{(3)}_{S_iS_j} + X^{(1)}_{S_iS_j} - 2X^{(2)}_{S_iS_j}\right) - \left(\left(\Psi^{(2)}\right)^2 + \sum_{\alpha,\beta} C^{(2)}_{\alpha\beta}\right) X^{(2)}_{S_iS_j}$$

$$\dot{X}^{(3)}_{S_iS_j} = (2-\delta_{ij})F^{(3)}_{S_i}F^{(3)}_{S_j} + C^{(3)}_{S_iS_j} + r\left(X^{(4)}_{S_iS_j} + X^{(2)}_{S_iS_j} - 2X^{(3)}_{S_iS_j}\right) - \left(\left(\Psi^{(3)}\right)^2 + \sum_{\alpha,\beta} C^{(3)}_{\alpha\beta}\right) X^{(3)}_{S_iS_j}$$

$$\dot{X}^{(4)}_{S_iS_j} = (2-\delta_{ij})F^{(4)}_{S_i}F^{(4)}_{S_j} + C^{(4)}_{S_iS_j} + r\left(X^{(5)}_{S_iS_j} + X^{(3)}_{S_iS_j} - 2X^{(4)}_{S_iS_j}\right) - \left(\left(\Psi^{(4)}\right)^2 + \sum_{\alpha,\beta} C^{(4)}_{\alpha\beta}\right) X^{(4)}_{S_iS_j}$$

$$\dot{X}^{(5)}_{S_iS_j} = (2-\delta_{ij})F^{(5)}_{S_i}F^{(5)}_{S_j} + C^{(5)}_{S_iS_j} + r\left(X^{(4)}_{S_iS_j} - X^{(5)}_{S_iS_j}\right) - \left(\left(\Psi^{(5)}\right)^2 + \sum_{\alpha,\beta} C^{(5)}_{\alpha\beta}\right) X^{(5)}_{S_iS_j}$$

The density constraints are

$$X^{(\ell)}_{DD} + \sum_{i=1}^{\mathcal{N}} X^{(\ell)}_{R_iD} + \sum_{i=0}^{\mathcal{N}} X^{(\ell)}_{S_iD} + \sum_{i=1}^{\mathcal{N}}\sum_{j=0}^{\mathcal{N}} X^{(\ell)}_{R_iS_j} + \sum_{i=1}^{\mathcal{N}}\sum_{j=1}^{i} X^{(\ell)}_{R_iR_j} + \sum_{i=0}^{\mathcal{N}}\sum_{j=0}^{i} X^{(\ell)}_{S_iS_j} = 1$$

To enforce these density constraints, we set

$$\Psi^{(\ell)} = F^{(\ell)}_D + \sum_{i=1}^{\mathcal{N}} F^{(\ell)}_{R_i} + \sum_{i=0}^{\mathcal{N}} F^{(\ell)}_{S_i}$$

4

Discussion

The goal of this dissertation has been to better understand the evolutionary dynamics of CRISPR-based gene drive systems, including potential problems and possible solutions.

In summary, progress was made on both fronts. In Chapter 1, I presented a mathematical model for CRISPR-based gene drive systems with resistance encoded at the drive locus and found that this form of resistance could, in fact, be a major obstacle to the long-term evolutionary stability of

CRISPR gene drive elements. I similarly studied the dynamics of an alternative design, and the results suggested that it could mitigate the stability problem posed by this form of resistance, at least in principle under the various assumptions of the model.

In Chapter 2, I then turned to the question of how effectively a basic CRISPR-based gene drive system might spread following accidental or otherwise unauthorized release, despite resistance. This necessitated a different modeling approach that could capture the effect of stochastic fluctuations when the drive is rare. The results suggested that even simple drive systems without optimization to mitigate the evolution of resistance (as studied in Chapter 1) could spread to significant frequencies in wild populations following small releases. I considered a variety of mitigating factors that have been observed empirically and found that the results were robust to each of these factors. In light of these findings, I noted the importance of adhering to previously proposed safety protocols in experimental design and recommended a great deal of further experimental and theoretical research before field trials are considered in wild populations.

In Chapter 3, I studied an alternative gene drive system called "daisy-chain gene drive" that seemed, intuitively, more amenable to containment—and, therefore, a potential technical solution to the problem of inadvertent spread studied in Chapter 2. As in the previous chapters, I constructed a mathematical model to clarify our thinking about the system and to study its dynamics given our assumptions. The results suggested that daisy-chain gene drive systems could be capable of attaining high frequency in a local population following a small release—making them potentially useful—while exhibiting low spread in subsequent populations connected by gene flow—making them potentially safer than standard drive designs. This was, of course, a preliminary examination of daisy-chain gene drive, and a large volume of future experimental and theoretical work will be required to evaluate the system's dynamics in real-world populations.

As for next steps, there are at least two promising future directions for this work. One could seek to better understand the drive systems considered here or develop entirely new drive systems.

In further studying existing drive systems, there is great potential for future work to use experimental and modeling methods to iteratively inform and refine each other. On the experimental front, next steps will involve carefully characterizing the effects of drive systems in application-relevant organisms—beyond the proof-of-principle experiments conducted to date—particularly considering within-organism drive dynamics, including drive efficiency, fitness effects, common avenues that lead to different forms of resistance and off-target cutting effects, as well as higher-level behavioral and ecological effects. On the other hand, additional modeling will be important for carefully designing informative experiments, while also producing tentative projections for dynamics of these systems in wild populations—which is the best that can be done prior to field trials. In modeling efforts of this type, a variety of factors not considered in this dissertation will need to be carefully considered, including species-specific behavior, interspecies interactions, and environmental features.

Besides studying existing systems, it will be useful for future work to also consider entirely new systems. A benefit of the modeling approaches used in this dissertation is that they can be used to rapidly prototype new designs—on paper—and determine whether they show enough promise to dedicate experimental resources to their construction and further study. In this effort, designs to solve known problems, such as containment and evolutionary stability, will continue to be promising areas of study.

Finally, perhaps the most important direction for future work in the field more broadly is to better understand whether, when and where to use these systems at all—a discussion in which technical details, as discussed in this dissertation, are only one factor. Moreover, this discussion will require a tremendous, conscientious effort spanning many stakeholder groups, from local communities to ethicists and policymakers. The approach of this dissertation has been to make no claims of whether CRISPR gene drive systems should be used in the wild, but rather to proceed with the assumption that their potential impact and the likelihood of their future use at least warrant their careful study. In my estimation, whether CRISPR gene drive systems see an eventual application or not, the efforts of the field will have been extremely worthwhile.

A

Review of CRISPR gene drive experiments

In Table A.1, we present empirical homing efficiencies for all CRISPR gene drive constructs reported to date. These studies varied in multiple ways: they studied different organisms; they used different methods for counting drive constructs (ranging from direct genetic measurement, such as quantitative PCR, to indirectly observing visible phenotypes), and they sometimes observed differential inheritance rates between sexes, possibly due to differences in male and female gamete characteristics.

Given this complexity, we elaborate here on the specific data we selected for review to produce Table A.1 and the reasoning for our choices.

Organism	Ref.	System name	Efficiency
Yeast	[37]	ade2::sgRNA	$> 99\%$
		ade2::sgRNA+URA3	100%
		sgRNA+ABD1	100%
		cas9+sgRNA	$> 99\%$
		ADE2+sgRNA+cas9	$> 99\%$
Fruit flies	[38]	γ-MCR	97%
	[67]	nanos	62%
		vasa	52%
		additional nanos	40%–62%
		additional vasa	37%–53%
Mosquitoes	[5]	AsMCRkh2 (male)	98%
		AsMCRkh2 (female)	14%
	[13]	AGAP011377	83%
		AGAP005958	95%
		AGAP007280	99%

Table A.1: Empirical homing efficiencies for all CRISPR gene drive systems published to date.

To begin, all studies performed some variation of producing drive/wild-type heterozygotes (DW), followed by counting the number which converted their wild-type allele to a drive allele. There were two main approaches.

1. Some constructs acted in the early embryo, in which case WW and DD individuals were mated to produce offspring which were initially WD. Observations were then made of adult genotypes. DD individuals must have undergone drive conversion, while WD individuals must have avoided conversion. Without drive, all adults are expected to be WD, but with drive, all are expected to be DD.

2. Other constructs acted in the germline of adults, so that adult WD individuals produce D gametes more often than chance under the effects of drive. To study these constructs, WD

individuals were mated with WW individuals. Without drive, half of adults should be WW, and half should be WD. With drive, however, all adults should be WD.

To employ a consistent strategy across the studies, we calculate two numbers for each drive construct: (i) the total number of initial alleles counted which were drives or were subject to drive, T, and (ii) the total number of resulting drive alleles, D. The homing efficiency can then be calculated in the following way:

$$P = \frac{2D}{T} - 1$$

Notice that if drive is perfectly efficient ($P = 1$), we have $D/T = 1$, *i.e.*, there are twice as many drive alleles as starting heterozygotes, while under standard inheritance ($P = 0$), the number of drive alleles is unchanged from the initial heterozygous state, $D/T = 1/2$. Below, we explain our calculations of these quantities for Table A.1.

Yeast, DiCarlo et al., (2015)

The study by DiCarlo *et al.* studied 5 distinct gene drive systems in yeast[37]. We address each distinct system in subsections below.

1. ade2::sgRNA

This is the basic split drive system containing only a guide RNA. Its design is depicted in Fig. 2B, and it is described on pp. 1250-1251, with results pictured in Fig. 2D and Fig. 4. Drive abundances were measured via colony counting (Fig. 2D), obtaining absolute colony numbers, and via qPCR (Fig. 4), obtaining relative abundances of drive alleles. By the colony counting method, the drive efficiency is measured at 100% ($D = T = 72$). By the qPCR method, $> 99\%$ of alleles counted from offspring were drive alleles, so $D > 0.99T$. Therefore:

$$P > 0.99$$

Strictly speaking, the inequality $D > 0.99T$ entails $P > 0.98$, but we set this to $P > 0.99$ because the qPCR results were indistinguishable from 100%. We make a similar approximation below for systems 4 and 5.

2. ADE2::sgRNA+URA3

This system aimed to test whether an associated 'cargo' gene could be spread with the minimal drive element. Its design is depicted in Fig. 3a, and results are shown in Fig. 3b. The related experiment measured drive presence via a visible phenotype (red pigment). In total, 60 haploids were red, or $D = 60$, out of 60 total alleles, $T = 60$. Thus:

$$P = 1$$

3. sgRNA+ABD1

The sgRNA+ABD1 drive system tested the ability to target a recoded essential gene. Its design is depicted in Fig. 3c, and results are discussed in the text (first full paragraph on pp. 1252). The presence of the drive was measured via sequencing of the ABD1 locus. In total, 72 haploids were found to have the drive, $D = 72$, out of 72 counted, $T = 72$.

$$P = 1$$

4. cas9+sgRNA

The first example of an 'autonomous' drive in the paper, this system is depicted in Fig. 5a. It consisted of a gRNA and cas9 together targeting the ADE2 locus (recoded due to safety/containment considerations). The fractional abundance of drive allele was measured by performing qPCR on diploid offspring from wild-type/drive haploid matings; the corresponding data is found in Fig. 5b. The fractional abundance of the drive allele was measured to be $> 99\%$, so $P > 0.99$, as for the first construct above.

$$P > 0.99$$

5. ADE2+sgRNA+cas9

This system is DiCarlo *et al.*'s example of a 'reversal' drive, designed to target and overwrite the autonomous drive (cas9+sgRNA, directly above). The system is depicted in Fig. 5c. The drive efficiency was measured in the same way as that for the cas9+sgRNA drive (qPCR to calculate fractional abundance of the overwriting drive allele in diploid offspring from haploid matings). The fractional abundance was calculated to be $> 99\%$, so $P > 0.99$, as above.

$$P > 0.99$$

Fruit flies, Gantz and Bier, (2015)

Gantz and Bier constructed an X-linked drive construct targeting the (X-linked) *yellow* locus in *Drosophila melanogaster* and acting in the early embryo[38]. The drive functions to knock out the *yellow* gene, which produces a yellow-body phenotype, denoted $y-$, due to lack of black melanin pigment formation. The wild-type phenotype is referred to as $y+$. Females with $<\ 2$ ar $y+$, while females with 2 copies of the drive or males with 1 copy should appear $y-$. The related data is found in Fig. 2E and Table 1.

Two sets of crosses were performed: (i) drive-males with wild-type females, and (ii) drive-females with wild-type males. To tabulate the allele counts D and T, we discuss the two crosses separately.

First, cross (i): In this cross, male offspring could not have possibly inherited a drive allele nor received one through conversion. This is because the only allele they could have inherited from the drive-male parent was the Y chromosome, but the drive is X-linked. Thus we do not consider male offspring in the total. As for female offspring, these should inherit exactly one drive allele and one wild-type allele prior to conversion. Then the adult female individuals should appear $y-$ if and only if drive-mediated conversion was successful. Thus we add exactly two alleles for each female offspring toward the total allele count, while we add one or two drive alleles to the drive allele count if the adults are $y+$ or $y-$, respectively. This yields $D_{♂} = 40 \times 2 + 1 \times 1 = 81$ and $T_{♂} = 40 \times 2 + 1 \times 2 = 82$. The drive efficiency for this cross is $P_{♂} = 2D_{♂}/T_{♂} - 1 = 0.976$.

Second, cross (ii): In this cross, male offspring are again uninformative, since each should inherit exactly one drive allele from the female parent and one Y allele from the male wild-type parent.

Thus we ignore male offspring in our counting. Female offspring, on the other hand, should all begin as WD embryos, with $y+$ phenotypes. Then adults are $y-$ if and only if they have undergone drive-mediated conversion. Thus we count two alleles for every female offspring in the total, one drive allele per $y+$ adult and two drive alleles per $y-$ adult. This yields $D_♀ = 203{\times}2{+}1{\times}6 = 412$, and $T_♀ = 203 \times 2 + 6 \times 2 = 418$. The drive efficiency for this cross is thus $P_♀ = 2D_♀/T_♀ = 0.971$.

We then consider crosses (i) and (ii) together to calculate the overall drive efficiency. This yields:

$$P = 2\frac{D_♂ + D_♀}{T_♂ + T_♀} - 1 = 2\frac{81 + 412}{82 + 418} - 1 = 0.972$$

Fruit flies, Champer *et al.*, (2017)

Champer *et al.* constructed two CRISPR gene drive constructs in *D. melanogaster*[67]. The first resembled the *vasa* promoter-driven construct from Gantz *et al.*, discussed in the section immediately above. An important addition, however, was a DsRed fluorescent protein as payload in the drive construct, which allows the drive to be detected in heterozygotes, as its red fluorescent phenotype is dominant. The second construct used the *nanos* promoter, which has been shown to restrict drive function to the germline and is expected to produce less toxicity (and thus a lower fitness cost associated with the drive construct).

1. *VASA* CONSTRUCT

This construct was similar to the one studied by Gantz *et al.*, discussed above. The construct targets the X-linked *yellow* gene. Disruption of the gene produces a recessive yellow phenotype, while the drive itself carries a DsRed payload, producing a dominant red fluorescent eye phenotype. To assess the construct's homing efficiency, wild-type males were crossed with heterozygous DW females. In this setup, all progeny should exhibit the red eye phenotype if the drive is perfectly efficient, while roughly 50% of progeny should exhibit the red eye phenotype in the absence of conversion. Here we count toward the total number of drive or susceptible alleles one allele per male offspring and one allele per female offspring, since in either case only one allele is inherited from the drive parent. Toward the number of drive alleles, we count one per offspring if the offspring displays the DsRed phenotype and zero otherwise. This data is shown in Table 2B of the Champer *et al.* (2017) study. We count as follows: $D_{♀} = 909 + 4 = 913$ (*i.e.*, the number of drive alleles counted over female offspring), $T_{♀} = 909 + 4 + 316 = 1229$, $D_{♂} = 953$, $T_{♂} = 953 + 265 + 3 = 1221$. Then we obtain:

$$P = 2\frac{D_{♂} + D_{♀}}{T_{♂} + T_{♀}} - 1 = 2\frac{953 + 913}{1221 + 1229} - 1 = 0.523.$$

2. *NANOS* CONSTRUCT

This construct is essentially the same as the *vasa* construct, except that it uses a different promoter and targets a different sequence in the *yellow* gene (the coding sequence, rather than the promoter as in the previous construct). The data is found in Table 1B of the Champer *et al.* (2017) study. We count potential drive alleles and total alleles as above. Our count is as follows: $D_{♀} = 290 + 100 + 108 = 498$, $T_{♀} = 290+100+108+119+10+9 = 636$, $D_{♂} = 594$, $T_{♂} = 594+11+103+2 = 710$. We obtain:

$$P = 2\frac{D_{♂} + D_{♀}}{T_{♂} + T_{♀}} - 1 = 2\frac{594 + 498}{710 + 636} - 1 = 0.622.$$

Additional data

The constructs described above were then tested in a variety of additional *D. melanogaster* lines, detailed in Table 3 of that work. The authors' efficiency calculations are detailed in the S1 Dataset. For the vasa construct (2 lines), the minimum is $P = 0.37$, and the maximum is $P = 0.53$. For the nanos construct (7 lines), the minimum is $P = 0.40$, and the maximum is $P = 0.62$.

Mosquitoes, Gantz *et al.*, (2015)

In this study, Gantz *et al.* constructed an autonomous CRISPR-based gene drive system in the malaria vector mosquito *Anopheles stephensi*[5]. The construct comprises two effector genes with anti-*Plasmodium falciparum* activity, a dominant marker gene (DsRed), and the CRISPR components (Cas9 with a single gRNA), spanning roughly 17 kb. The construct targets the *kynurenine hydroxylase*^{white} (*kh*^{w}*)* locus, which has a recessive white-eye phenotype. The effect of this targeting is that drive/wild-type heterozygotes display a DsRed phenotype, while drive homozygotes display both DsRed and white eyes.

While this one construct was made and studied, it exhibited differential transmission between lines founded by drive males/wild-type females and drive-females/wild-type males. More specifically, lines in which drive alleles are inherited only through male parents display drastically higher drive efficiencies than lines in which the drive allele is inherited at some point via a female parent. To explain this discrepancy, the authors propose a model whereby in crosses between transgenic females and wild-type males, maternal deposition of Cas9 in eggs results in NHEJ-mediated disruption of the paternally derived wild-type chromosome in the early embryo. Crosses between transgenic males and wild-type females, on the other hand, do not see Cas9 deposited in the early embryo, and Cas9 cutting is better contained to the later germline, where HDR is more efficient.

To account for this discrepancy, we choose to consider these two cases separately and report homing efficiencies for each.

1. Transgenic male lines

Here we consider all offspring (larvae + adults) whose drive alleles (or potentially-inherited drive alleles) have been passed down only through male ancestors. This includes all offspring from the male-founder crosses in Table 1 of the main text (10.1 G2♂ and 10.2 G2♂), as well as crosses 6 and 8 in Table 2 (also Fig. 3). We choose to compile all alleles from each of these crosses together to calculate an average efficiency across all available data. Because the constructs are on autosomes, we treat male offspring and female offspring identically, and we count toward the total allele count, T, one allele from each offspring (since at most one drive allele can be inherited in each cross), and we count toward the drive allele total, D, one allele for each DsRed$^+$ individual observed, since this is a dominant marker for the drive. Finally, we consider both larvae and adults identically, as conversion

G3 crosses	D	T	Reference
10.1 G_2×WT, larval	829	832	Table S3
10.2 G_2×WT, larval	3060	3085	Table S4
10.1 G_2×WT, adult	833	836	Table S5
10.2 G_2×WT, adult	1258	1274	Table S6
Total	5980	6027	—

G_4 crosses	D	T	Reference
Cross 6, larval	949	955	Table S7
Cross 8, larval	609	628	Table S8
Cross 6, adult	882	888	Table S10
Cross 8, adult	565	583	Table S11
Total	3005	3054	—

Table A.2: Gantz *et al.*, *An. stephensi* transgenic male lines. (top) Phenotypes of G3 progeny. (bottom) Phenotypes of G_4 progeny.

is anticipated to have occurred before this stage, and results are similar between adults and larvae. Values of D and T for each cross are displayed in Table A.2.

To obtain an average efficiency for the construct, we sum the values of D and T across all crosses in Table A.2. We obtain:

$$P = 2\frac{8985}{9081} - 1 = 0.979.$$

2. Transgenic female lines

To understand the effect of maternal Cas9 deposition, we count all offspring (larvae + adults) from crosses such that the any (potentially) inherited drive allele has been inherited via a female parent at least once. This includes no G3 offspring, as the drive alleles present in G_2 parents were inherited from G_1 males. Thus we include only G_4 offspring of G3 parents, specifically Crosses 1-4, and as for

G$_4$ larvae	D	T	Reference
Cross 1	28	48	Table S7
Cross 2	332	635	Table S7
Cross 3	204	324	Table S8
Cross 4	372	632	Table S8
Total	936	1639	—

G$_4$ adults	D	T	Reference
Cross 1	19	35	Table S10
Cross 2	306	554	Table S10
Cross 3	169	272	Table S11
Cross 4	1430	2500	Table S11
Total	1924	3361	—

Table A.3: Gantz *et al.*, *An. stephensi* transgenic male lines. (left) Phenotypes of G$_4$ larvae. (right) Phenotypes of G$_4$ adults.

the transgenic male lines, we sum both larval and adult crosses. Values of D and T for each cross are displayed in Table A.3. Summing the values in Table A.3 yields:

$$P = 2\frac{2860}{5000} - 1 = 0.144.$$

Mosquitoes, Hammond *et al.*, (2015)

In this study, the authors construct three CRISPR-based gene drive systems in the malaria vector *An. gambiae*, each targeting a different gene with a recessive female sterility phenotype upon disruption[13]. These are examples of suppression drives whose purpose is to reduce or eradicate wild populations. Each drive construct carries a copy of Cas9, a single guide RNA, and red fluorescent protein (RFP) which has a dominant fluorescent phenotype. Each construct targets one of three female fertility genes, referred to as AGAP011377, AGAP005958, and AGAP007280, but otherwise they are identical.

To determine homing efficiency, drive-heterozygotes were crossed with wild-type homozygotes, and offspring were scored visually for the presence of the dominant marker RFP gene. Thus in our tabulations, we count one allele per individual toward the total, T, and we count one allele per RFP$^+$ individual toward the drive allele count, D. Furthermore, the outcrosses were performed over several generations. To obtain average homing efficiencies, we sum drive alleles and total alleles over G$_2$, G$_3$, G$_4$, and G$_5$ generations, when applicable. (Some constructs were tested over more generations than others.) This data is found in Table 2 in the study. Furthermore, we sum across male- and female-drive parent crosses, since we would expect these to behave identically with respect to homing, given that the female drive parents are capable of producing offspring.

1. AGAP011377

This construct was studied over generations G$_2$ to G$_5$ in Table 2. The total number of relevant alleles resulting from crosses between drive-male parents and wild-type females was $T_{♂} = 636+1631+1654+505 = 4426$, while the male drive total was $D_{♂} = 581+1442+1550+491 = 4064$. The female total was $T_{♀} = 60+92+142 = 294$, and the female drive total was $D_{♀} = 55+70+121 = 246$. The average efficiency is then:

$$P = 2\frac{D_{♂} + D_{♀}}{T_{♂} + T_{♀}} - 1 = 2\frac{4064 + 246}{4426 + 294} - 1 = 0.826.$$

2. AGAP005958

This construct was studied over generations G$_2$ and G$_3$. There were no offspring from female-drive crosses to wild-type due to the low fertility of these individuals. The total was $T = 1689 + 278 =$

1967, and the drive total was $D = 1654 + 268 = 1922$. The efficiency is thus:

$$P = 2\frac{D}{T} - 1 = 2\frac{1922}{1967} - 1 = 0.954.$$

3. AGAP007280

This construct was studied over generations G2 and G3. The male total was $T_{\male} = 1383 + 505 = 1888$, and the male drive total was $D_{\male} = 1377 + 499 = 1876$. The female total was $T_{\female} = 257$, and the female drive total was $D_{\female} = 255$. The efficiency is:

$$P = 2\frac{D_{\male} + D_{\female}}{T_{\male} + T_{\female}} - 1 = 2\frac{1876 + 255}{1888 + 257} - 1 = 0.987.$$

B

Supplementary figures

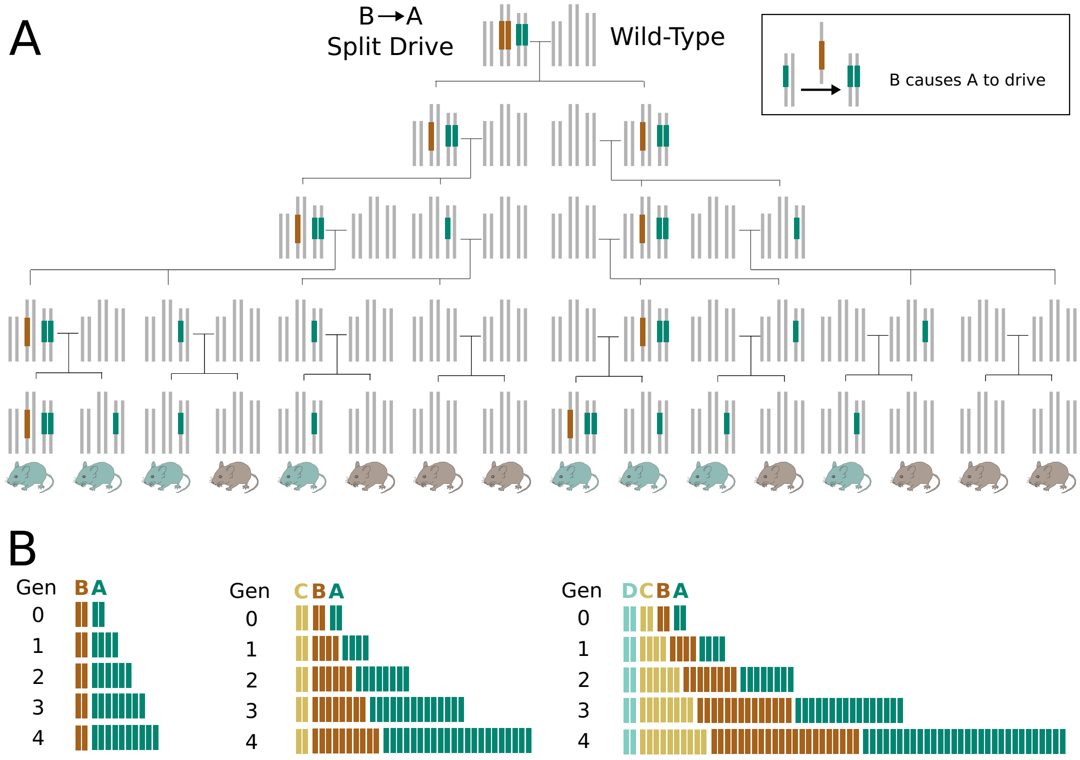


Figure B.1: B⟶A "split" drives and daisy drive family tree analysis. (A) Family tree resulting from a single-organism release of a B⟶A split drive in a large wild-type population in the absence of selection. (In reality, B elements would be deleterious and thus decline in frequency over time.) For comparison, a C⟶B⟶A daisy drive is shown in main text Fig. 1c. Green mice have at least one copy of the cargo A element, while grey mice have only the wild-type allele at that locus. (B) A graphical depiction of total alleles in a population per generation for B⟶A through D⟶C⟶B⟶A daisy drives. Throughout, chromosome illustrations represent genotypes of germline cells after drive has occurred.

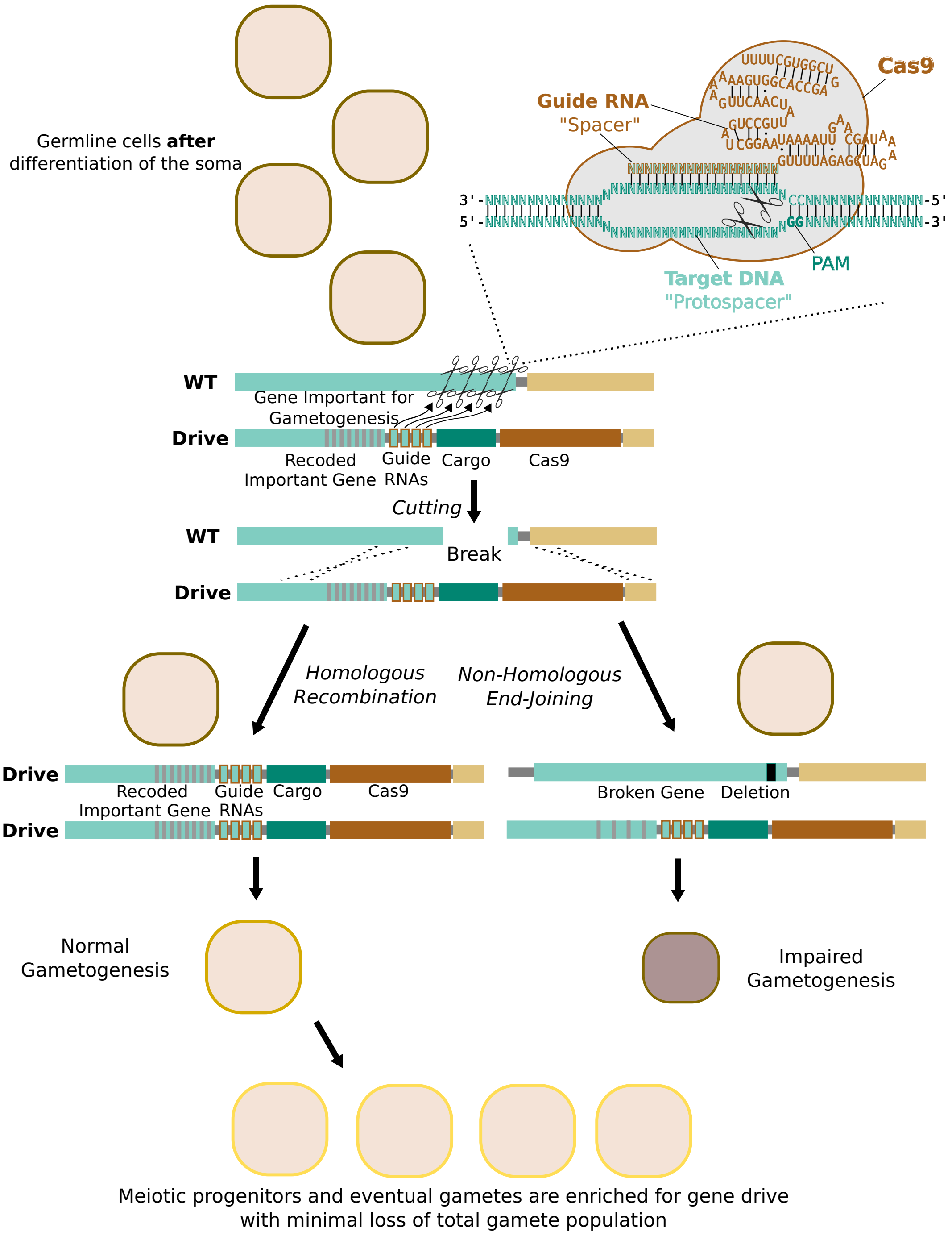

Germline cells **after** differentiation of the soma
Cas9
UUUUCGUGGCU
AAAGUGGCACCGA G
A GUUCAACU A
Guide RNA
"Spacer"
GUCCGUU
A UCGGAA UAAAAUU GAA CGAUAA
GUUUUAGAGCUAG A
NNNNNNNNNNNNNNNNNNNN
3'-NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCCNNNNNNNNNNNNNNN-5'
5'-NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGGNNNNNNNNNNNNNNN-3'
PAM
Target DNA
"Protospacer"
WT
Gene Important for Gametogenesis
Drive
Recoded Important Gene
Guide RNAs
Cargo
Cas9
Cutting
WT
Break
Drive
Homologous Recombination
Non-Homologous End-Joining
Drive
Recoded Important Gene
Guide RNAs
Cargo
Cas9
Drive
Broken Gene
Deletion
Normal Gametogenesis
Impaired Gametogenesis
Meiotic progenitors and eventual gametes are enriched for gene drive with minimal loss of total gamete population

Figure B.2: A potential means of reducing the fitness cost resulting from incorrect repair. One strategy might involve targeting a gene whose loss impairs gametogenesis, such as a ribosomal gene. Increased replication of correctly repaired cells carrying the drive system could potentially result in a wild-type number of gametes, all of which carry the drive system.

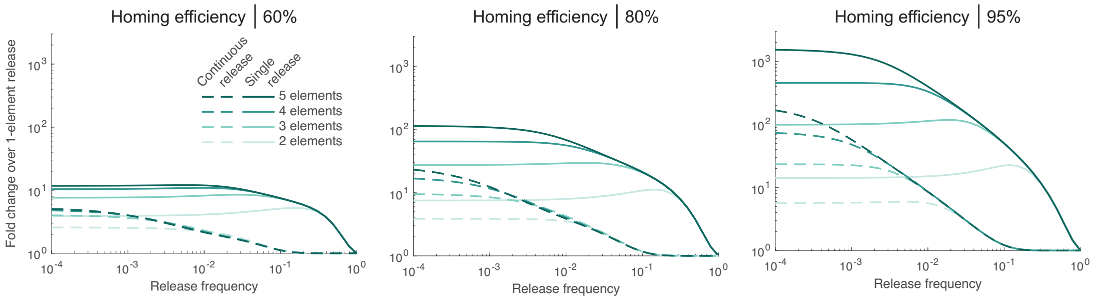


Figure B.3: Fold-change in daisy drive cargo frequency after 20 generations for various daisy chain lengths relative to a release of organisms only containing the cargo. (left) Homing efficiency is assumed to be 60%, and (middle) 80% and (right) 95%. All figures assume 10% cargo fitness cost, 0.01% upstream element cost and neutral resistance. Solid lines correspond to a single release with initial release frequency indicated by the horizontal axes, while dashed lines correspond to continuous releases with frequency indicated by the horizontal axes. See SI Text Section 2.3 for details on our continuous release implementation.

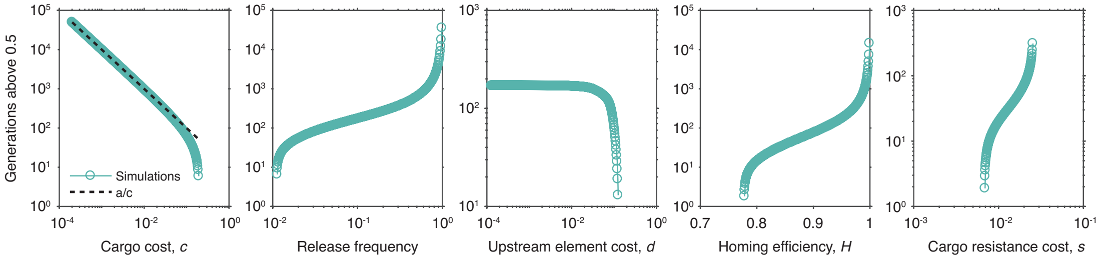


Figure B.4: Analysis of the time that a 3-element (CBA) daisy-chain cargo element remains above 50% frequency in a single population. In each plot, the parameter indicated on the horizontal axis is varied, and the other parameters are fixed. The fixed values in the first four panels are: cargo element (A) fitness cost, c, 5%, release frequency 10%, upstream element (C, B) fitness cost, d, 0.01%, homing efficiency, H, 95%, cargo resistance cost, s, 0. In the far-right panel, all parameters are identical except the release frequency is 1%. In the first panel, we additionally plot a function a/c, fitted with $a = 9.99$, to illustrate the inverse relationship between cargo time above 50% and cargo fitness cost, c, when c is low. The model used throughout is described in SI Text Section 2.2.

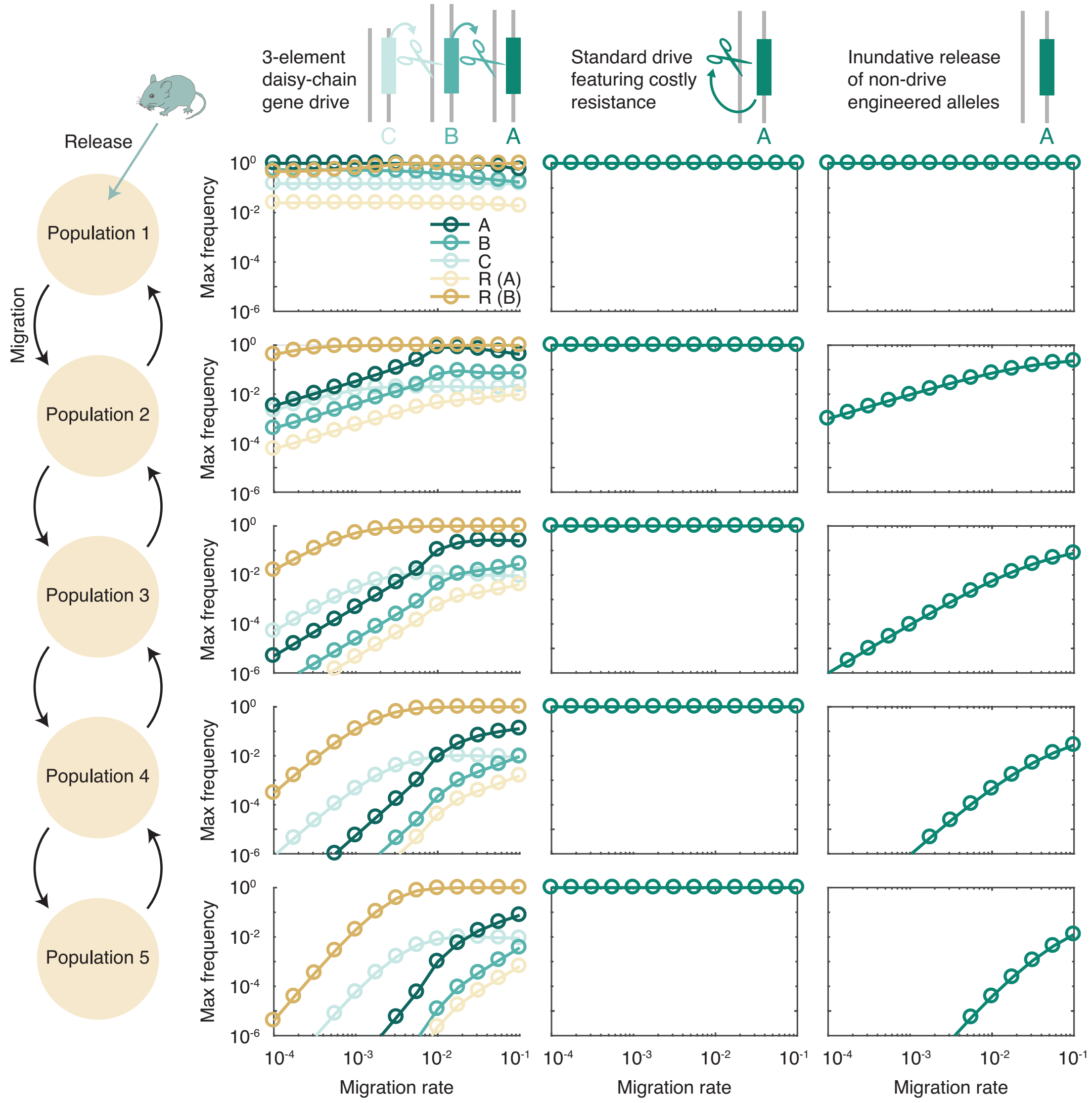


Figure B.5: Further analysis of the 5-population model, shown in Fig. 4 of the main text and described in SI Text Section 5. Three drive systems are considered, as in main text Fig. 4: CBA daisy-chain drive (left), self-propagating drive (middle) and inundative non-drive release (right). We assume 5 equally-sized populations connected in a chain via one constant migration rate, which varies on the horizontal axes from 10^{-4} to 10^{-1}. (e.g., 10^{-2} corresponds to a scenario where each population is connected to its neighbors via a migration rate of 10^{-2} in each direction.) Maximum frequencies of each allele in each population over 500 generations are then plotted as functions of the migration rate.

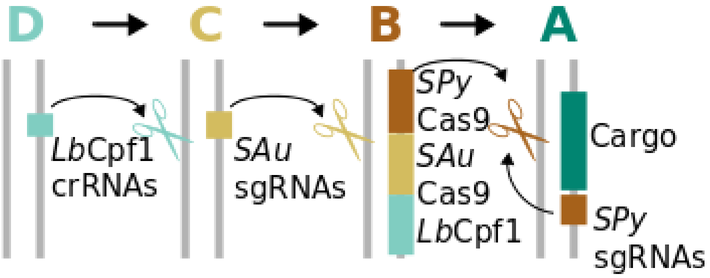


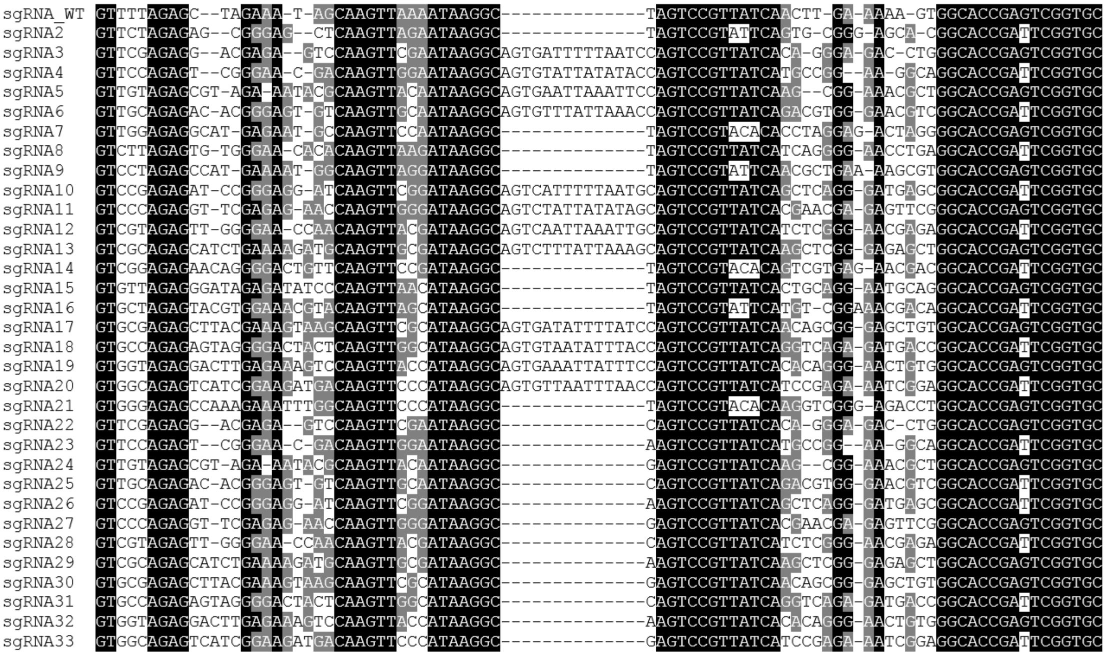
Figure B.6: Daisy drive systems can be constructed using orthogonal Cas9 elements. Such a drive system is resistant to conversion into a daisy necklace, which would require a recombination event that moved the entire Cas9 gene and associated guide RNAs into a subsequent locus in the daisy-chain. Ensuring that all the Cas9 proteins are expressed appropriately without re-using promoters and thereby creating homology between elements could be challenging.

```
template          1 -----------------------------------------N----NNC-AAGTTVVVATAAGGC--------------N
Br_S.pyogenes     1 -----------------TTGTTGGAACCATTCAAAACAGCAT----AGC-AAGTTAAAATAAGGC--------------T
Br_S.dysagalact   1 -----------------TTGTTGGAACCATTCAAAACAACGT----AGC-AAGTTAAAATAAGGC--------------T
Br_S.equi         1 -----------------CCTATGGAACTATTCAATACAGCAT----AGCAAAGTTAAAATAAGGC-------------TT
Br_S.thermophil   1 ----------------CTTACACAGTTACTTAAATCTTGCAG----AAG-CTACAAAGATAAGGC-------------TT
Br_S.salivarius   1 -----------------GGTTTGAAACCATTCGAAACAATAC----AGCAAAGTTAAAATAAGGC-------------TT
Br_S.gallolytic   1 -----------------TTGTTGGAGCTATTCGAAACAACAC----AGC-GAGTTAAAATAAGGC-------------TT
Br_S.lutetiensi   1 -----------------TTGTTGGAACTATTCGAAACAACAC----AGT-GAGTTAAAATAAGGC-------------TT
Br_S.anginosisB   1 -----------------ATGTTGGAATCATTCGAAACAACAC----AGC-AAGTTAAAATAAGGC-------------TT
Br_S.mitis        1 -----------------TCGTTGGAACTATTCGAAACAACAC----AGCAAAGTTAAAATAAGGC-------------TT
Br_S.sanguinis    1 -----------------TTGTTGGAACTATTCGAAACAACAC----AGC-AAGTTAAAATAAGGC-------------TT
Br_S.oralis       1 -----------------TTGTTGGAACTATTCGAAACAACAC----AGC-AAGTTAAAATAAGGC-------------TT
Br_S.mutans       1 ----------------CTTACACAGTTACTTAAATCTTGCAG----AAG-CTACAAAGATAAGGC-------------TT
Br_S.intermediu   1 ----------------CTTACACAGTTACTTAAATCTTGCAG----AAG-CTACAAAGATAAGGC-------------TT
Br_S.anginosusA   1 ----------------CTTACACAGTTACTTAAATCTTGCAG----AAG-CTACAAAGATAAGGC-------------TT
Br_S.thermophil   1 ----------------CTTACACAGTTACTTAAATCTTGCAG----AAG-CTACAAAGATAAGGC-------------TT
Br_S.vestibular   1 ----------------CTTACACAGTTACTTAAATCTTGCAG----AAG-CTACAAAGATAAGGC-------------TT
Br_S.gordonii     1 ----------------CTTACACAGTTACTTAAATCTTGCAG----AAG-CTACAAAGATAAGGC-------------TT
Br_S.parasangui   1 ----------------CTTACACAGTTACTTAAATCTTGCAG----AAG-CTACAAAGATAAGGC-------------TT
Br_S.orisratti    1 ----------------CTTGCACAGTTACTTAAATCTTGCAG----AGC-CTACAAAGATAAGGC-------------TT
Br_S.henryi       1 ----------------CTTGCACAGTTACTTAAATCTTGCTG----AGC-CTACAAAGATAAGGC-------------TT
Br_S.infantariu   1 ----------------CTTGCACGGTTACTTAAATCTTGCAG----AGC-CTACAAAGATAAGGC-------------TT
Ch_C.jejuni       1 ------------------------------AAGAAATTTAAA----AAG-GGACTAAAATAAAGAGTT--TGCGGGACTC
Ch_F.novicida     1 ----------------------------ATCTAAAATTATAA--ATGTA-CCAAATAATTAATGC-------------TC
Ch_S.thermophil   1 ------TTG-------TGGTTTGAAACCATTCGAAACAACAC----AGC-GAGTTAAAATAAGGC-------------TT
Ch_M.mobile       1 TGTATTTCGAAATACAGATGTACAGTTAAGAATACATAAGAATGATACA-TCACTAAAAAAAGGC-------------TT
Ch_L.innocua      1 -------------------ATTGTTAGTATTCAAAATAACAT----AGC-AAGTTAAAATAAGGC-------------TT
Ch_S.pyogenes     1 -------------------GTTGGAACCATTCAAAACAGCAT----AGC-AAGTTAAAATAAGGC--------------T
Ch_S.mutans       1 -------------------GTTGGAATCATTCGAAACAACAC----AGC-AAGTTAAAATAAGGCAGTGATTTTTAATCC
Ch_S.thermophil   1 ------TTG-------TGGTTTGAAACCATTCGAAACAACAC----AGC-GAGTTAAAATAAGGC-------------TT
Ch_N.meningitid   1 ---------------ACATATTGTCGCACTGCGAAATGAGAA----CCG-TTGCTACAATAAGGCC--------------
Ch_P.multocida    1 ---------------GCATATTGTTGCACTGCGAAATGAGAG----ACG-TTGCTACAATAAGGC---------------

template         21 AGTCCGTYHYCANNNNGRRA--NNNNG-GCACCGAKTCGGTGC-------------------------------------
Br_S.pyogenes    45 AGTCCGTTATCAACTTGAAA--AAGTG-GCACCGAGTCGGTGCTTTTTTT------------------------------
Br_S.dysagalact  45 AGTCCGTTATCAACTTGAAA--AAGTG-GCACCGAGTCGGTGCTTTTT--------------------------------
Br_S.equi        47 TGTCCGTAATCAACCTGAAA--AGGGGAGCACCGAATCGGTGCTTTTTT-------------------------------
Br_S.thermophil  47 AGTCCGTACTCAACTTGAAA--AGGTG-GCACCGATTCGGAAGGC-------------TTCATGCCGAAATCAACACCCT
Br_S.salivarius  47 AGTCCGTATTCAACTTGAGA--AAGTG-GCACCGATTCGGTGCTTTTTT-------------------------------
Br_S.gallolytic  46 TGTCCGTACACAACTTGTAA--AAGTGGCACCCGATTCGGGTGCGTTTTTTT----------------------------
Br_S.lutetiensi  46 TGTCCGTACACAACTTATAA--AAGTGGCACCCGATTCGGATGCATTTTTT-----------------------------
Br_S.anginosisB  46 TGTCCGTACTCAACTT-AAA--AAGTGCGCACCGATTCGGTGCTTTTTT-------------------------------
Br_S.mitis       47 TGTCCGTACACAACTTGAAA--AAGTGCGCACCGATTCGGTGCTTTTTT-------------------------------
Br_S.sanguinis   46 TGTCCGTACACAACTTGAAA--AAGTGCGCACCGATTCGGTGCTTTTTT-------------------------------
Br_S.oralis      46 TGTCCGTACACAACTTGAAA--AAGTGCGCACCGATTCGGTGCTTTTT--------------------------------
Br_S.mutans      47 CATGCCGAAATCAACACCCT--ATCTATTATAAGATAGGGTGTTTT----------------------------------
Br_S.intermediu  47 CATGCCGAAATCAACACCCT--GTC-----TATGACGGGGTGTTTT----------------------------------
Br_S.anginosusA  47 CATGCCGAAATCAACACCCT--GTC-----TATGACGGGGTGTTTT----------------------------------
Br_S.thermophil  47 AGTCCGTACTCAACTTGAAA--AGGTG-GCACCGATTCGGAAGGC-------------TTCATGCCGAAATCAACACCCT
Br_S.vestibular  47 CATGCCGAAATCAACACCCT--GTCA-TTTTATGGCAGGGTGTTTT----------------------------------
Br_S.gordonii    47 CATGCCGAATTCAACACCCT--GTCA--TTTATGGCGGGGTGTTTT----------------------------------
Br_S.parasangui  47 CATGCCGAATTCAACACCCT--GTCA--TTTATGGCGGGGTGTTTT----------------------------------
Br_S.orisratti   47 TATGCCGAAATCAAGCACCC--C-----GTTTATACGAGGTGCTTTT---------------------------------
Br_S.henryi      47 CATGCCGAAATCAAGCACCC--CCGT-TTTTAACGAGGGGTGCTTTT---------------------------------
Br_S.infantariu  47 CATGCCGAATTCAAGCACCC--CA---TGTTTACATGGGGTGCTTTT---------------------------------
Ch_C.jejuni      44 TGCGGGGTTACAATCCCCTA--AAAC--------------CGCTTTT---------------------------------
Ch_F.novicida    37 TGTAATCATTTAAAAGTATTTTGAACGGACCTCTGTTTGACACGTCTGAATAACTAAAAA--------------------
Ch_S.thermophil  50 CATGCCGAAATCAACACCCT--GTCA-TTTTATGGCAGGGAAGGC-------------TTAGTCCGTACTCAACTTGAAA
Ch_M.mobile      67 TATGCCGTAACTACTACTTA-------TTTTCAAAATAAGTAGTTTTTTTT-----------------------------
Ch_L.innocua     44 TGTCCGTTATCAACTTTTAATTAAGTA-GCGCTGTTTCGGCGCTTTTTTT------------------------------
Ch_S.pyogenes    43 AGTCCGTTATCAACTTGAAA--AAGTG-GCACCGAGTCGGTGCTTTTTTT------------------------------
Ch_S.mutans      57 AGTCCGTACACAACTTGAAA--AAGTGCGCACCGATTCGGTGCTTTTTTATTT---------------------------
Ch_S.thermophil  50 CATGCCGAAATCAACACCCT--GTCA-TTTTATGGCAGGGAAGGC-------------TTAGTCCGTACTCAACTTGAAA
Ch_N.meningitid  47 ------------GTCTGAAA--AGATGTGCCGCAACGCTCTGCCCCTTAAAGCTTCTGCTTTAAGGGG---------CA-
Ch_P.multocida   46 ------------TTCTGAAA--AGAATGACCGTAACGCTCTGCCCCTTGTGATTCTTAATTGCAAGGGGCATCGTTTTT-

template            ---------------------------------------------------------------------
Br_S.pyogenes       ---------------------------------------------------------------------
Br_S.dysagalact     ---------------------------------------------------------------------
Br_S.equi           ---------------------------------------------------------------------
Br_S.thermophil 111 --GTCA-TTTTATGGCAGGGGTTTTTTTT------------------------------TGTTTT----
Br_S.salivarius     ---------------------------------------------------------------------
Br_S.gallolytic     ---------------------------------------------------------------------
Br_S.lutetiensi     ---------------------------------------------------------------------
Br_S.anginosisB     ---------------------------------------------------------------------
Br_S.mitis          ---------------------------------------------------------------------
Br_S.sanguinis      ---------------------------------------------------------------------
Br_S.oralis         ---------------------------------------------------------------------
Br_S.mutans         ---------------------------------------------------------------------
Br_S.intermediu     ---------------------------------------------------------------------
Br_S.anginosusA     ---------------------------------------------------------------------
Br_S.thermophil 111 --GTCA-TTTTATGGCAGGGGTTTTTTTT------------------------------TGTTTT----
Br_S.vestibular     ---------------------------------------------------------------------
Br_S.gordonii       ---------------------------------------------------------------------
```

Figure B.7: Multiple sequence alignment of existing tracrRNA sequences from closely related Cas9 systems with the tracrRNA component of our sgRNA template (i.e., the template from Fig. 3.5C, GTNNNAGAGNNN–GRRA–NNNCAAGTTVVVATAAGGCNAGTCCGTYHYCANNNN-GRR-A-NNNNGGCACCGAKTCGGTGC). The sequences with names beginning "Br_" are taken from Fig. S2 of Briner et al., *Mol. Cell* (2014) (Ref. 103), and the sequences with names beginning "Ch_" are taken from Fig. 4 of Chylinski et al., *RNA Biol.*, (2013) (Ref. 111). Alignments were performed via the MAFFT program with default parameters (Refs. 113, 112).



```
sgRNA_WT GTTTTAGAGC--TAGAAA-T-AGCAAGTTAAAATAAGGC--------------TAGTCCGTTATCAACTT-GA-AAAA-GTGGCACCGAGTCGGTGC
sgRNA2   GTTCTAGAGAG--CGGGAG--CTCAAGTTAGAATAAGGC--------------TAGTCCGTATTCAGTG-CGGG-AGCA-CGGCACCGATTCGGTGC
sgRNA3   GTTCGAGAGG--ACGAGA--GTCCAAGTTCGAATAAGGCAGTGATTTTTAATCCAGTCCGTTATCACA-GGGA-GAC-CTGGGCACCGAGTCGGTGC
sgRNA4   GTTCCAGAGT--CGGGAA-C-GACAAGTTGGAATAAGGCAGTGTATTATATACCAGTCCGTTATCATGCCGG--AA-GGCAGGCACCGATTCGGTGC
sgRNA5   GTTGTAGAGCGT-AGA-AATACGCAAGTTACAATAAGGCAGTGAATTAAATTCCAGTCCGTTATCAAG--CGG-AAACGCTGGCACCGAGTCGGTGC
sgRNA6   GTTGCAGAGAC-ACGGGAGT-GTCAAGTTGCAATAAGGCAGTGTTTATTAAACCAGTCCGTTATCAGACGTGG-GAACGTCGGCACCGATTCGGTGC
sgRNA7   GTTGGAGAGGCAT-GAGAAT-GCCAAGTTCCAATAAGGC--------------TAGTCCGTACACACCTAGGAG-ACTAGGGGCACCGAGTCGGTGC
sgRNA8   GTCTTAGAGTG-TGGGAA-CACACAAGTTAAGATAAGGC--------------TAGTCCGTTATCATCAGGGG-AACCTGAGGCACCGATTCGGTGC
sgRNA9   GTCCTAGAGCCAT-GAAAAT-GGCAAGTTAGGATAAGGC--------------TAGTCCGTATTCAACGCTGAA-AAGCGTGGCACCGAGTCGGTGC
sgRNA10  GTCCGAGAGAT-CCGGGAGG-ATCAAGTTCGGATAAGGCAGTCATTTTTAATGCAGTCCGTTATCAGCTCAGG-GATGAGCGGCACCGATTCGGTGC
sgRNA11  GTCCCAGAGGT-TCGAGAG-AACCAAGTTGGGATAAGGCAGTCTATTATATAGCAGTCCGTTATCACGAACGA-GAGTTCGGGCACCGAGTCGGTGC
sgRNA12  GTCGTAGAGTT-GGGGAA-CCAACAAGTTACGATAAGGCAGTCAATTAAATTGCAGTCCGTTATCATCTCGGG-AACGAGAGGCACCGATTCGGTGC
sgRNA13  GTCGCAGAGCATCTGAAAAGATGCAAGTTGCGATAAGGCAGTCTTTATTAAAGCAGTCCGTTATCAAGCTCGG-GAGAGCTGGCACCGAGTCGGTGC
sgRNA14  GTCGGAGAGAACAGGGGACTGTTCAAGTTCCGATAAGGC--------------TAGTCCGTACACAGTCGTGAG-AACGACGGCACCGATTCGGTGC
sgRNA15  GTGTTAGAGGGATAGAGATATCCCAAGTTAACATAAGGC--------------TAGTCCGTTATCACTGCAGG-AATGCAGGGCACCGAGTCGGTGC
sgRNA16  GTGCTAGAGTACGTGGAAACGTACAAGTTAGCATAAGGC--------------TAGTCCGTATTCATGT-CGGAAACGACAGGCACCGATTCGGTGC
sgRNA17  GTGCGAGAGCTTACGAAAGTAAGCAAGTTCGCATAAGGCAGTGATATTTTATCCAGTCCGTTATCAACAGCGG-GAGCTGTGGCACCGAGTCGGTGC
sgRNA18  GTGCCAGAGAGTAGGGGACTACTCAAGTTGGCATAAGGCAGTGTAATATTTACCAGTCCGTTATCAGGTCAGA-GATGACCGGCACCGATTCGGTGC
sgRNA19  GTGGTAGAGGACTTGAGAAAGTCCAAGTTACCATAAGGCAGTGAAATTATTTCCAGTCCGTTATCACACAGGG-AACTGTGGGCACCGAGTCGGTGC
sgRNA20  GTGGCAGAGTCATCGGAAGATGACAAGTTCCCATAAGGCAGTGTTAATTTAACCAGTCCGTTATCATCCGAGA-AATCGGAGGCACCGATTCGGTGC
sgRNA21  GTGGGAGAGCCAAAGAAATTTGGCAAGTTCCCATAAGGC--------------TAGTCCGTACACAAGGTCGGG-AGACCTGGCACCGAGTCGGTGC
sgRNA22  GTTCGAGAGG--ACGAGA--GTCCAAGTTCGAATAAGGC--------------CAGTCCGTTATCACA-GGGA-GAC-CTGGGCACCGAGTCGGTGC
sgRNA23  GTTCCAGAGT--CGGGAA-C-GACAAGTTGGAATAAGGC--------------AAGTCCGTTATCATGCCGG--AA-GGCAGGCACCGATTCGGTGC
sgRNA24  GTTGTAGAGCGT-AGA-AATACGCAAGTTACAATAAGGC--------------GAGTCCGTTATCAAG--CGG-AAACGCTGGCACCGAGTCGGTGC
sgRNA25  GTTGCAGAGAC-ACGGGAGT-GTCAAGTTGCAATAAGGC--------------CAGTCCGTTATCAGACGTGG-GAACGTCGGCACCGATTCGGTGC
sgRNA26  GTCCGAGAGAT-CCGGGAGG-ATCAAGTTCGGATAAGGC--------------AAGTCCGTTATCAGCTCAGG-GATGAGCGGCACCGATTCGGTGC
sgRNA27  GTCCCAGAGGT-TCGAGAG-AACCAAGTTGGGATAAGGC--------------GAGTCCGTTATCACGAACGA-GAGTTCGGGCACCGAGTCGGTGC
sgRNA28  GTCGTAGAGTT-GGGGAA-CCAACAAGTTACGATAAGGC--------------CAGTCCGTTATCATCTCGGG-AACGAGAGGCACCGATTCGGTGC
sgRNA29  GTCGCAGAGCATCTGAAAAGATGCAAGTTGCGATAAGGC--------------AAGTCCGTTATCAAGCTCGG-GAGAGCTGGCACCGAGTCGGTGC
sgRNA30  GTGCGAGAGCTTACGAAAGTAAGCAAGTTCGCATAAGGC--------------GAGTCCGTTATCAACAGCGG-GAGCTGTGGCACCGAGTCGGTGC
sgRNA31  GTGCCAGAGAGTAGGGGACTACTCAAGTTGGCATAAGGC--------------CAGTCCGTTATCAGGTCAGA-GATGACCGGCACCGATTCGGTGC
sgRNA32  GTGGTAGAGGACTTGAGAAAGTCCAAGTTACCATAAGGC--------------AAGTCCGTTATCACACAGGG-AACTGTGGGCACCGAGTCGGTGC
sgRNA33  GTGGCAGAGTCATCGGAAGATGACAAGTTCCCATAAGGC--------------GAGTCCGTTATCATCCGAGA-AATCGGAGGCACCGATTCGGTGC
```

Figure B.8: Complete list of sequence-divergent guide RNAs generated and assayed using the transcriptional activation reporter.

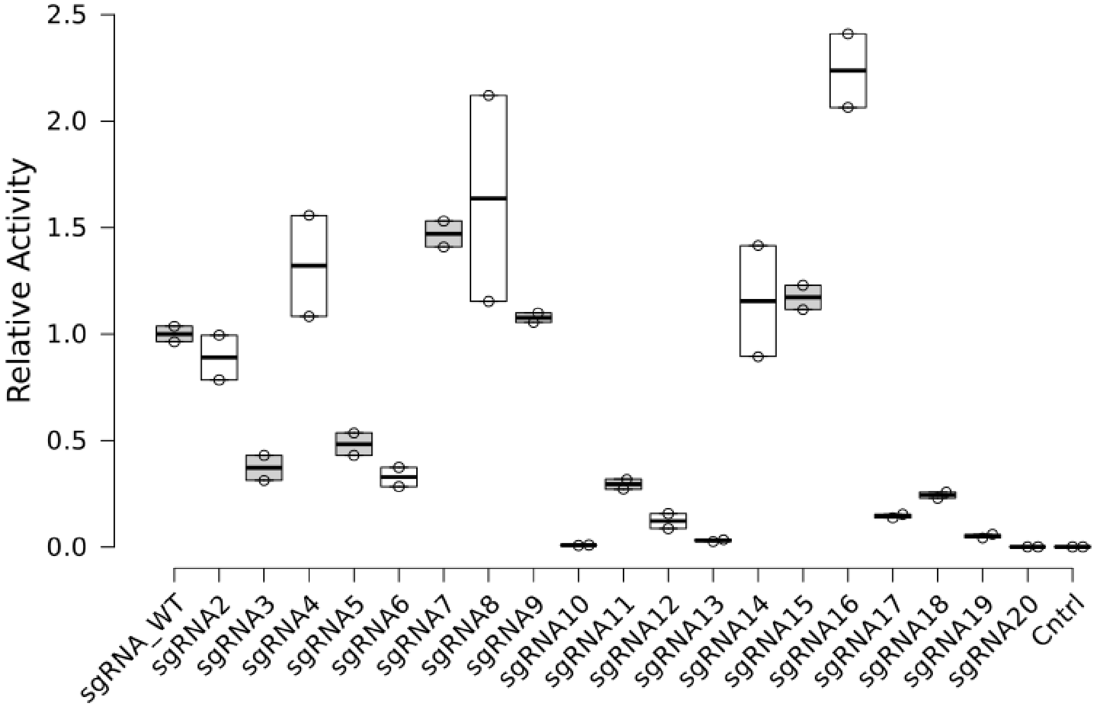


Figure B.9: Results of the pilot screen of the first set of designed sgRNA sequences. 3-6, 10-13, and 17-20 all carried the extra insert; the latter 8 displayed markedly lower activity and were not further considered. The cause of the difference is unclear, although it is worth noting that these all had longer stem-loops than did 3-6, all of which were closer to the activity of the standard or 'wild-type' sgRNA.

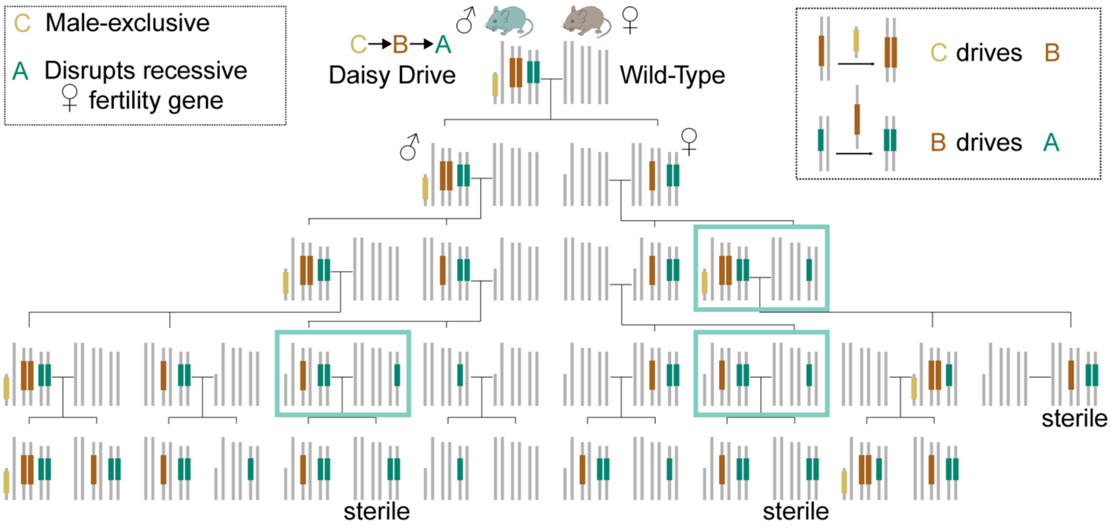


Figure B.10: Potential family tree of a C→B→A genetic load daisy drive for which the cargo in the A element disrupts a female fertility gene. The C element is male-linked, ensuring that it does not suffer a fitness cost from the loss of female fertility. Mating events between two parents carrying the A element (boxed) can produce sterile female offspring that will suppress the population. Males do not suffer a fitness cost due to disruption of female-specific fertility genes. Genome illustrations depict germline cells after drive has occurred. Females are placed on the right side in each pair of individuals.

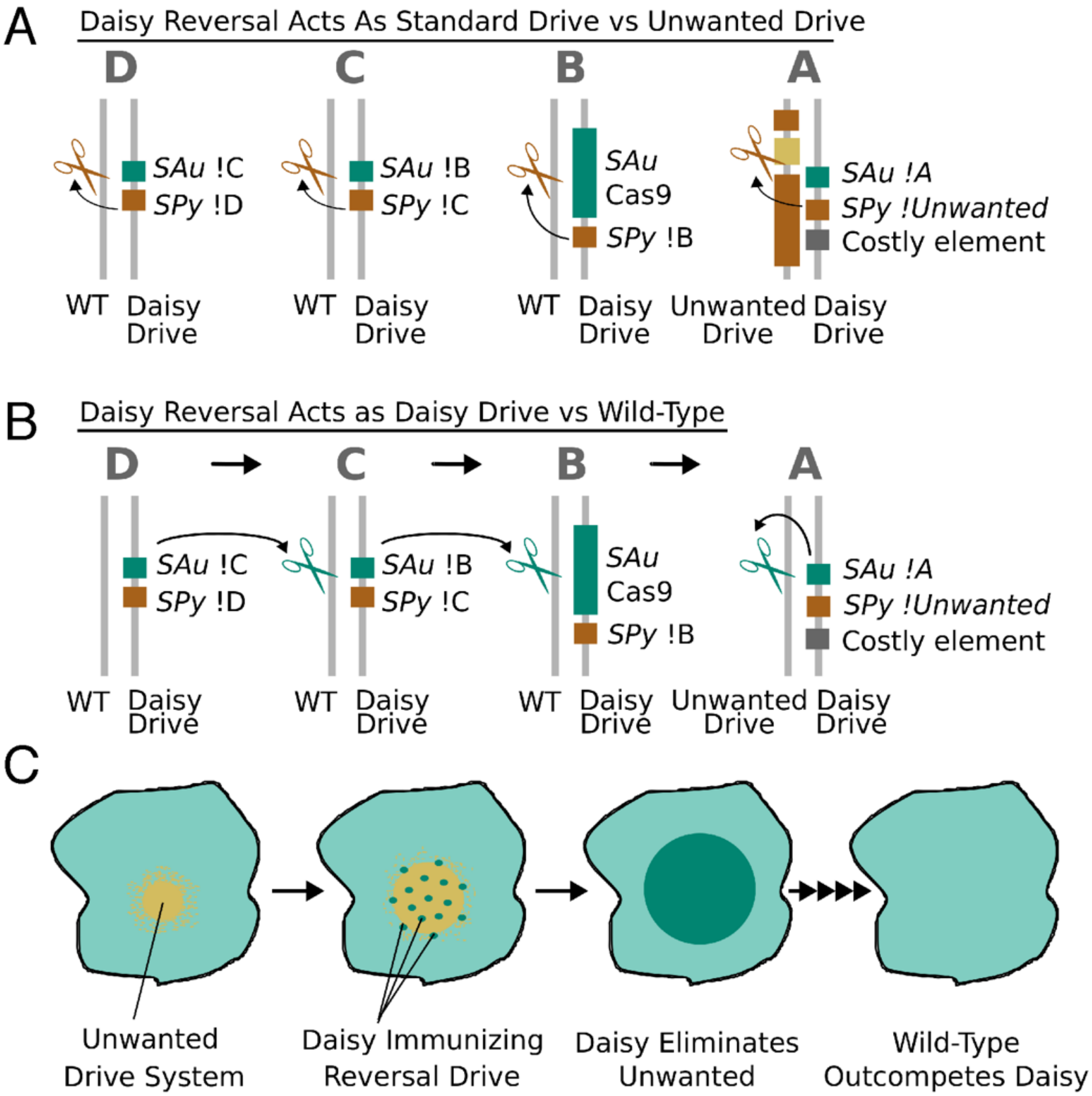


Figure B.11: Utility of a costly daisy reversal drive with orthogonal Cas9 elements in achieving complete genetic reversal of an unwanted drive system to wild-type. Suppose an unwanted drive system has spread a harmful cargo (yellow) through the target locus A via the commonly used Cas9 protein from *Streptococcus pyogenes*. (A) A daisy reversal drive system uses guide RNAs for *S. pyogenes* Cas9 to copy all elements while overwriting the unwanted drive system and its cargo. (B) The same daisy reversal drive system spreads as a normal daisy drive using its own orthogonal CRISPR system (e.g. S. aureus Cas9) on encountering wild-type sequences. (C) An unwanted drive system is countered by releasing the daisy reversal system at multiple sites. The daisy drive system efficiently overwrites the unwanted drive system throughout its range, spreading into and through the wild-type sequences at the edges of that range to ensure that it reaches and eliminates every copy. This immediately eliminates the harmful cargo. Because the A element of the daisy drive system is costly and the other elements are always co-resident with it due to the daisy drive effect, all elements of the daisy drive will be outcompeted and eliminated by wild-type alleles over time, potentially leading to complete genetic reversal.

References

[1] Organization, W. H. *World malaria report 2015* (World Health Organization, 2016).

[2] World Health Organization & World Health Organization. Global Malaria Programme. *Global technical strategy for malaria, 2016-2030*.

[3] Walker, P. G., Griffin, J. T., Ferguson, N. M. & Ghani, A. C. Estimating the most efficient allocation of interventions to achieve reductions in plasmodium falciparum malaria burden and transmission in africa: a modelling study. *The Lancet Global Health* 4, e474–e484 (2016).

[4] Burt, A., Coulibaly, M., Crisanti, A., Diabate, A. & Kayondo, J. K. Gene drive to reduce malaria transmission in sub-saharan africa. *Journal of Responsible Innovation* 5, S66–S80 (2018).

[5] Gantz, V. M. *et al.* Highly efficient Cas9-mediated gene drive for population modification of the malaria vector mosquito Anopheles stephensi. *Proceedings of the National Academy of Sciences* 201521077 (2015). URL http://www.pnas.org/content/early/2015/11/18/1521077112.abstract.

[6] Fuchs, S., Nolan, T. & Crisanti, A. Mosquito transgenic technologies to reduce plasmodium transmission. In *Malaria*, 601–622 (Springer, 2012).

[7] Wang, S. & Jacobs-Lorena, M. Genetic approaches to interfere with malaria transmission by vector mosquitoes. *Trends in biotechnology* 31, 185–193 (2013).

[8] Corby-Harris, V. *et al.* Activation of akt signaling reduces the prevalence and intensity of malaria parasite infection and lifespan in anopheles stephensi mosquitoes. *PLoS pathogens* 6, e1001003 (2010).

[9] Sumitani, M. *et al.* Reduction of malaria transmission by transgenic mosquitoes expressing an antisporozoite antibody in their salivary glands. *Insect molecular biology* 22, 41–51 (2013).

[10] Franz, A. W., Balaraman, V. & Fraser Jr, M. J. Disruption of dengue virus transmission by mosquitoes. *Current opinion in insect science* 8, 88–96 (2015).

[11] Franz, A. W. *et al.* Fitness impact and stability of a transgene conferring resistance to dengue-2 virus following introgression into a genetically diverse aedes aegypti strain. *PLoS neglected tropical diseases* 8, e2833 (2014).

[12] Champer, J., Buchman, A. & Akbari, O. S. Cheating evolution: engineering gene drives to manipulate the fate of wild populations. *Nature Reviews Genetics* 17, 146 (2016).

[13] Hammond, A. *et al.* A CRISPR-Cas9 gene drive system targeting female reproduction in the malaria mosquito vector Anopheles gambiae. *Nature biotechnology* (2015). URL `http://dx.doi.org/10.1038/nbt.3439`.

[14] Galizi, R. *et al.* A synthetic sex ratio distortion system for the control of the human malaria mosquito. *Nature communications* 5, 3977 (2014).

[15] Najjar, D. A., Normandin, A. M., Strait, E. A. & Esvelt, K. M. Driving towards ecotechnologies. *Pathogens and global health* 1–11 (2018).

[16] Data and Statistics | Lyme Disease | CDC. URL `https://www.cdc.gov/lyme/stats/index.html`.

[17] Esvelt, K. M. & Gemmell, N. J. Conservation demands safe gene drive. *PLoS Biology* 15, e2003850 (2017).

[18] Goldson, S. *et al.* New zealand pest management: current and future challenges. *Journal of the Royal Society of New Zealand* 45, 31–58 (2015).

[19] Grafton-Cardwell, E. E., Stelinski, L. L. & Stansly, P. A. Biology and management of asian citrus psyllid, vector of the huanglongbing pathogens. *Annual Review of Entomology* 58, 413–432 (2013).

[20] Buchman, A., Marshall, J. M., Ostrovski, D., Yang, T. & Akbari, O. S. Synthetically engineered medea gene drive system in the worldwide crop pest drosophila suzukii. *Proceedings of the National Academy of Sciences* 115, 4725–4730 (2018).

[21] Walsh, D. B. *et al.* Drosophila suzukii (diptera: Drosophilidae): invasive pest of ripening soft fruit expanding its geographic range and damage potential. *Journal of Integrated Pest Management* 2, G1–G7 (2011).

[22] Hamilton, W. D. Extraordinary sex ratios. *Science* 156, 477–488 (1967).

[23] Curtis, C. Possible use of translocations to fix desirable genes in insect pest populations. *Nature* 218, 368 (1968).

[24] Akbari, O. S. *et al.* A synthetic gene drive system for local, reversible modification and suppression of insect populations. *Current biology* 23, 671–7 (2013). URL http://www.ncbi.nlm.nih.gov/pubmed/23541732.

[25] Reeves, R. G., Bryk, J., Altrock, P. M., Denton, J. A. & Reed, F. A. First Steps towards Underdominant Genetic Transformation of Insect Populations. *PLoS ONE* 9, e97557 (2014). URL http://dx.plos.org/10.1371/journal.pone.0097557.

[26] Chen, C.-H. *et al.* A Synthetic Maternal-Effect Selfish Genetic Element Drives Population Replacement in Drosophila. *Science* 316, 597–600 (2007). URL http://www.sciencemag.org/content/316/5824/597.short.

[27] Akbari, O. S. *et al.* Novel synthetic medea selfish genetic elements drive population replacement in drosophila; a theoretical exploration of medea-dependent population suppression. *ACS synthetic biology* 3, 915–928 (2014).

[28] Burt, A. Site-specific selfish genes as tools for the control and genetic engineering of natural populations. *Proceedings. Biological sciences / The Royal Society* 270, 921–928 (2003).

[29] Windbichler, N. *et al.* A synthetic homing endonuclease-based gene drive system in the human malaria mosquito. *Nature* 473, 212–215 (2011). URL http://dx.doi.org/10.1038/nature09937.

[30] Windbichler, N. *et al.* Homing endonuclease mediated gene targeting in Anopheles gambiae cells and embryos. *Nucleic acids research* 35, 5922–33 (2007). URL http://nar.oxfordjournals.org/content/35/17/5922.short.

[31] Chan, Y.-S., Huen, D. S., Glauert, R., Whiteway, E. & Russell, S. Optimising homing endonuclease gene drive performance in a semi-refractory species: the drosophila melanogaster experience. *PloS one* 8, e54130 (2013).

[32] Chan, Y.-S., Naujoks, D. A., Huen, D. S. & Russell, S. Insect population control by homing endonuclease-based gene drive: an evaluation in drosophila melanogaster. *Genetics* 188, 33–44 (2011).

[33] Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 337, 816–21 (2012). URL http://science.sciencemag.org/content/337/6096/816.abstract.

[34] Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* 339, 823–6 (2013). URL http://www.sciencemag.org/content/339/6121/823.short.

[35] Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* 339, 819–23 (2013). URL http://science.sciencemag.org/content/339/6121/819.abstract.

[36] Doudna, J. A. & Charpentier, E. The new frontier of genome engineering with CRISPR-Cas9. *Science* 346, 1258096–1258096 (2014). URL http://science.sciencemag.org/content/346/6213/1258096.abstract.

[37] DiCarlo, J. E., Chavez, A., Dietz, S. L., Esvelt, K. M. & Church, G. M. Safeguarding CRISPR-Cas9 gene drives in yeast. *Nature Biotechnology* (2015). URL http://dx.doi.org/10.1038/nbt.3412.

[38] Gantz, V. M. & Bier, E. The mutagenic chain reaction: A method for converting heterozygous to homozygous mutations. *Science* 348, 442–444 (2015). URL http://www.sciencemag.org/content/early/2015/03/18/science.aaa5945.abstract.

[39] Noble, C., Olejarz, J., Esvelt, K. M., Church, G. M. & Nowak, M. A. Evolutionary dynamics of CRISPR gene drives. *Science Advances* 3, e1601964 (2017). URL http://advances.sciencemag.org/lookup/doi/10.1126/sciadv.1601964.

[40] Noble, C., Adlam, B., Church, G. M., Esvelt, K. M. & Nowak, M. A. Current crispr gene drive systems are likely to be highly invasive in wild populations. *eLife* 7, e33423 (2018). URL https://doi.org/10.7554/eLife.33423.

[41] Esvelt, K. M., Smidler, A. L., Catteruccia, F. & Church, G. M. Concerning RNA-guided gene drives for the alteration of wild populations. *eLife* 3, e03401 (2014). URL http://elifesciences.org/content/3/e03401.abstract.

[42] Akbari, B. O. S. *et al.* Safeguarding gene drive experiments in the laboratory. *Science* 349, 927–9 (2015). URL http://www.sciencemag.org/content/early/2015/07/29/science.aac7932.full.

[43] Oye, K. A. *et al.* Regulating gene drives. *Science* 345, 626–8 (2014). URL http://www.sciencemag.org/content/345/6197/626.short.

[44] Sinkins, S. P. & Gould, F. Gene drive systems for insect disease vectors. *Nature reviews. Genetics* 7, 427–435 (2006).

[45] Alphey, L. Genetic control of mosquitoes. *Annual review of entomology* 59, 205–24 (2014). URL http://www.annualreviews.org/doi/abs/10.1146/annurev-ento-011613-162002.

[46] Charlesworth, B. & Langley, C. H. The population genetics of Drosophila transposable elements. *Annual review of genetics* 23, 251–287 (1989).

[47] Ward, C. M. *et al.* Medea selfish genetic elements as tools for altering traits of wild populations: a theoretical analysis. *Evolution* 65, 1149–62 (2011).

[48] Lyttle, T. W. Segregation distorters. *Annual review of genetics* 25, 511–557 (1991).

[49] Charlesworth, B. & Hartl, D. L. Population dynamics of the segregation distorter polymorphism of drosophila melanogaster. *Genetics* 89, 171–192 (1978). URL http://www.genetics.org/content/89/1/171.short.

[50] Tao, Y., Hartl, D. L. & Laurie, C. C. Sex-ratio segregation distortion associated with reproductive isolation in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America* 98, 13183–8 (2001). URL http://www.pnas.org/content/98/23/13183.short.

[51] Henikoff, S., Ahmad, K. & Malik, H. S. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* 293, 1098–102 (2001). URL http://www.sciencemag.org/content/293/5532/1098.short.

[52] Deredec, A., Godfray, H. C. J. & Burt, A. Requirements for effective malaria control with homing endonuclease genes. *Proceedings of the National Academy of Sciences of the United States of America* 108, E874–80 (2011). URL http://www.pnas.org/content/108/43/E874.short.

[53] Deredec, A., Burt, A. & Godfray, H. C. J. The population genetics of using homing endonuclease genes in vector and pest management. *Genetics* 179, 2013–2026 (2008).

[54] Mali, P., Esvelt, K. M. & Church, G. M. Cas9 as a versatile tool for engineering biology. *Nature methods* 10, 957–63 (2013). URL http://dx.doi.org/10.1038/nmeth.2649.

[55] Lajoie, M. J. *et al.* Probing the limits of genetic recoding in essential genes. *Science* 342, 361–3 (2013). URL http://science.sciencemag.org/content/342/6156/361.abstract.

[56] Lajoie, M. J. *et al.* Genomically recoded organisms expand biological functions. *Science* 342, 357–60 (2013). URL http://www.sciencemag.org/content/342/6156/357.abstract.

[57] Ostrov, N. *et al.* Design, synthesis, and testing toward a 57-codon genome. *Science* 353 (2016).

[58] Wang, K. *et al.* Defining synonymous codon compression schemes by genome recoding. *Nature* 539, 59–64 (2016). URL http://www.nature.com/doifinder/10.1038/nature20124.

[59] Mackay, T. F. C. *et al.* The Drosophila melanogaster Genetic Reference Panel. *Nature* 482, 173–8 (2012).

[60] Nei, M. & Li, W.-H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America* 76, 5269–5273 (1979).

[61] Nowak, M. A. *Evolutionary Dynamics* (Harvard University Press, 2006).

[62] Unckless, R. L., Clark, A. G. & Messer, P. W. Evolution of Resistance Against CRISPR/Cas9 Gene Drive. *Genetics* 205, 827–841 (2017).

[63] Bull, J. J. Lethal gene drive selects inbreeding. *Evolution, Medicine, and Public Health* eow030 (2016). URL http://emph.oxfordjournals.org/lookup/doi/10.1093/emph/eow030.

[64] Rasgon, J. L. Multi-Locus Assortment (MLA) for Transgene Dispersal and Elimination in Mosquito Populations. *PLoS ONE* 4, e5833 (2009). URL http://dx.plos.org/10.1371/journal.pone.0005833.

[65] Noble, C. *et al.* Daisy-chain gene drives for the alteration of local populations. *bioRxiv* (2016).

[66] National Academies of Sciences, Engineering, and Medicine. *Gene Drives on the Horizon: Advancing Science, Navigating Uncertainty, and Aligning Research with Public Values* (National Academies Press, 2016).

[67] Champer, J. *et al.* Novel CRISPR/Cas9 gene drive constructs reveal insights into mechanisms of resistance allele formation and drive efficiency in genetically diverse populations. *PLOS Genetics* 13, e1006796 (2017). URL http://dx.plos.org/10.1371/journal.pgen.1006796.

[68] Drury, D. W., Dapper, A. L., Siniard, D. J., Zentner, G. E. & Wade, M. J. CRISPR/Cas9 gene drives in genetically variable and nonrandomly mating wild populations. *Science Advances* 3 (2017). URL http://advances.sciencemag.org/content/3/5/e1601910.

[69] Marshall, J. M., Buchman, A., Sánchez C., H. M. & Akbari, O. S. Overcoming evolved resistance to population-suppressing homing-based gene drives. *Scientific Reports* 7, 3776 (2017). URL http://dx.doi.org/10.1038/s41598-017-02744-7.

[70] Hesman Saey, T. Gene drives' fatal flaw has an upside. *Science News* 190, 13 (2016).

[71] Callaway, E. Gene drives thwarted by emergence of resistant organisms. *Nature* 542 (2017).

[72] Unckless, R. L., Messer, P. W., Connallon, T. & Clark, A. G. Modeling the manipulation of natural populations by the mutagenic Chain reaction. *Genetics* 201, 425–431 (2015).

[73] Wright, S. Evolution in Mendelian populations. *Genetics* 16, 97–159 (1931).

[74] Fisher, R. *The genetical theory of natural selection* (The Clarendon Press, Oxford, 1930).

[75] Haldane, J. B. S. A mathematical theory of natural and artificial selection. Part V. Selection and mutation. *Proc. Camb. Phil. Soc* 23, 838–844 (1927).

[76] Marshall, J. M. The effect of gene drive on containment of transgenic mosquitoes. *Journal of Theoretical Biology* 258, 250–265 (2009).

[77] Noble, C. 2018. drive-invasiveness. d294fdo (GitHub). URL https://github.com/charlestonnoble/drive-invasiveness.

[78] Yaro, A. S. *et al.* Reproductive Output of Female Anopheles gambiae (Diptera: Culicidae): Comparison of Molecular Forms. *Journal of Medical Entomology* 43, 833–839 (2006).

[79] Hill, W. G. Effective size of populations with overlapping generations. *Theoretical population biology* 3, 278–289 (1972).

[80] Tanaka, H., Stone, H. A. & Nelson, D. R. Spatial gene drives and pushed genetic waves. *Proceedings of the National Academy of Sciences* 201705868 (2017).

[81] Traulsen, A. & Reed, F. A. From genes to games: Cooperation and cyclic dominance in meiotic drive. *Journal of Theoretical Biology* 299, 120–125 (2012).

[82] Funk, C. & Rainie, L. Public and scientists' views on science and society. *Pew Research Center* 29 (2015).

[83] Couzin, J. & Kaiser, J. Gene therapy. As Gelsinger case ends, gene therapy suffers another blow. *Science (New York, N.Y.)* 307, 1028 (2005). URL http://www.ncbi.nlm.nih.gov/pubmed/15718439.

[84] Magori, K. & Gould, F. Genetically engineered underdominance for manipulation of pest populations: a deterministic model. *Genetics* 172, 2613–20 (2006).

[85] Gould, F., Huang, Y., Legros, M. & Lloyd, A. L. A killer-rescue system for self-limiting gene drive of anti-pathogen constructs. *Proceedings. Biological sciences / The Royal Society* 275, 2823–9 (2008). URL http://rspb.royalsocietypublishing.org/content/275/1653/2823.short.

[86] Esvelt, K. M. & Gemmell, N. J. Conservation demands safe gene drive. *PLOS Biology* 15, e2003850 (2017). URL http://dx.plos.org/10.1371/journal.pbio.2003850.

[87] Curtis, C. F. Possible Use of Translocations to fix Desirable Genes in Insect Pest Populations. *Nature* 218, 368–369 (1968). URL http://www.nature.com/doifinder/10.1038/218368a0.

[88] Champer, J. *et al.* Reducing resistance allele formation in CRISPR gene drives. *bioRxiv* 150276 (2017). URL https://www.biorxiv.org/content/early/2017/06/14/150276.

[89] Marrelli, M. T., Moreira, C. K., Kelly, D., Alphey, L. & Jacobs-Lorena, M. Mosquito transgenesis: what is the fitness cost? *Trends in parasitology* 22, 197–202 (2006). URL http://www.ncbi.nlm.nih.gov/pubmed/16564223.

[90] Harvey-Samuel, T., Ant, T., Gong, H., Morrison, N. I. & Alphey, L. Population-level effects of fitness costs associated with repressible female-lethal transgene insertions in two pest insects. *Evolutionary Applications* 7, 597–606 (2014). URL http://doi.wiley.com/10.1111/eva.12159.

[91] Slaymaker, I. M. *et al.* Rationally engineered Cas9 nucleases with improved specificity. *Science* 351, 84–88 (2015). URL http://science.sciencemag.org/content/351/6268/84.abstract.

[92] Kleinstiver, B. P. *et al.* High-fidelity CRISPR−Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* 529, 490–495 (2016). URL http://www.nature.com/doifinder/10.1038/nature16526.

[93] Wyss, J. H. Screwworm Eradication in the Americas. *Annals of the New York Academy of Sciences* 916, 186–193 (2006). URL http://doi.wiley.com/10.1111/j.1749-6632.2000.tb05289.x.

[94] Port, F., Chen, H.-M., Lee, T. & Bullock, S. L. Optimized CRISPR/Cas tools for efficient germline and somatic genome engineering in Drosophila. *Proceedings of the National Academy of Sciences* 111, E2967–E2976 (2014). URL http://www.pnas.org/cgi/doi/10.1073/pnas.1405500111.

[95] Ranganathan, V., Wahlin, K., Maruotti, J. & Zack, D. J. Expansion of the CRISPR−Cas9 genome targeting space through the use of H1 promoter-expressed guide RNAs. *Nature Communications* 5 (2014). URL http://www.nature.com/doifinder/10.1038/ncomms5516.

[96] Mefferd, A. L., Kornepati, A. V., Bogerd, H. P., Kennedy, E. M. & Cullen, B. R. Expression of CRISPR/Cas single guide RNAs using small tRNA promoters. *RNA* 21, 1683–1689 (2015). URL http://rnajournal.cshlp.org/lookup/doi/10.1261/rna.051631.115.

[97] Xie, K., Minkenberg, B. & Yang, Y. Boosting CRISPR/Cas9 multiplex editing capability with the endogenous tRNA-processing system. *Proceedings of the National Academy of Sciences* 112, 3570–3575 (2015). URL http://www.pnas.org/lookup/doi/10.1073/pnas.1420294112.

[98] Port, F. & Bullock, S. L. Expansion of the CRISPR toolbox in an animal with tRNA-flanked Cas9 and Cpf1 gRNAs. Tech. Rep. (2016). URL http://biorxiv.org/lookup/doi/10.1101/046417.

[99] Yan, Q. *et al.* Multiplex CRISPR/Cas9-based genome engineering enhanced by Drosha-mediated sgRNA-shRNA structure. *Scientific Reports* 6, 38970 (2016). URL http://www.nature.com/articles/srep38970.

[100] Wang, J. *et al.* The gRNA-miRNA-gRNA Ternary Cassette Combining CRISPR/Cas9 with RNAi Approach Strongly Inhibits Hepatitis B Virus Replication. *Theranostics* 7, 3090–3105 (2017). URL http://www.thno.org/v07p3090.htm.

[101] Xie, C. *et al.* SgRNA Expression of CRIPSR-Cas9 System Based on MiRNA Polycistrons as a Versatile Tool to Manipulate Multiple and Tissue-Specific Genome Editing. *Scientific Reports* 7, 5795 (2017). URL http://www.nature.com/articles/s41598-017-06216-w.

[102] Esvelt, K. M. *et al.* Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nature Methods* 10, 1116–1121 (2013). URL http://www.nature.com/doifinder/10.1038/nmeth.2681.

[103] Briner, A. E. *et al.* Guide rna functional modules direct cas9 activity and orthogonality. *Molecular cell* 56, 333–339 (2014).

[104] Nishimasu, H. *et al.* Crystal Structure of Cas9 in Complex with Guide RNA and Target DNA. *Cell* 156, 935–949 (2014). URL http://dx.doi.org/10.1016/j.cell.2014.02.001.

[105] Dang, Y. *et al.* Optimizing sgRNA structure to improve CRISPR-Cas9 knockout efficiency. *Genome Biology* 16, 280 (2015). URL http://genomebiology.com/2015/16/1/280.

[106] Simoni, A. *et al.* Development of synthetic selfish elements based on modular nucleases in Drosophila melanogaster. *Nucleic Acids Research* 42, 7461–7472 (2014).

[107] Min, J., Noble, C., Najjar, D. & Esvelt, K. Daisy quorum drives for the genetic restoration of wild populations. *bioRxiv* (2017). URL http://www.biorxiv.org/content/early/2017/03/21/115618.

[108] Burt, A. & Deredec, A. Self-limiting population genetic control with sex-linked genome editors. *bioRxiv* 236489 (2017). URL https://www.biorxiv.org/content/early/2017/12/19/236489.

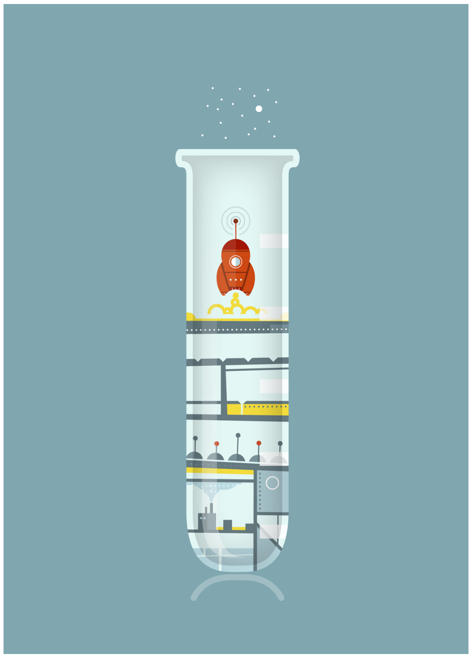
[109] Wu, B., Luo, L. & Gao, X. J. Cas9-triggered chain ablation of cas9 as a gene drive brake. *Nature Biotechnology* 34, 137–138 (2016). URL http://www.nature.com/doifinder/10.1038/nbt.3444.

[110] Dhole, S., Vella, M. R., Lloyd, A. L. & Gould, F. Invasion and migration of spatially self-limiting gene drives: A comparative analysis. *Evolutionary Applications* 0, 1–15 (2018). URL https://onlinelibrary.wiley.com/doi/abs/10.1111/eva.12583.

[111] Chylinski, K., Le Rhun, A. & Charpentier, E. The tracrrna and cas9 families of type ii crispr-cas immunity systems. *RNA biology* 10, 726–737 (2013).

[112] Katoh, K. & Toh, H. Recent developments in the mafft multiple sequence alignment program. *Briefings in bioinformatics* 9, 286–298 (2008).

[113] Katoh, K., Rozewicki, J. & Yamada, K. D. Mafft online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics* bbx108 (2017). URL http://dx.doi.org/10.1093/bib/bbx108.



This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. The above illustration, *Science Experiment 02*, was created by Ben Schlitter and released under CC BY-NC-ND 3.0. A template that can be used to format a PhD dissertation with this look & feel has been released under the permissive AGPL license, and can be found online at github.com/suchow/Dissertate or from its lead author, Jordan Suchow, at suchow@post.harvard.edu.