Imperial College London

Expanding the flexibility of genome editing approaches for population control of the malaria mosquito

Nace Kranjc

Department of Life Sciences

Imperial College London

Thesis submitted for the degree of

Doctor of Philosophy

2022

Acknowledgements

First and foremost, I want to thank my supervisors Prof Andrea Crisanti and Dr Tony Nolan, for the opportunity they gave me to embark on this wonderful journey and for their scientific guidance through these past four years. Because they always asked the hard questions and pointed me in the right direction, I discovered joy in the work I do, and I will forever be grateful for that.

I am thankful to my progress review panel, Dr John Pinney and Prof Geoff Baldwin, for their feedback and guidance during my PhD.

I want to thank Roberto for all the lunches and coffee breaks that sparked numerous conversations, ideas, possibilities, and, most importantly, a very dear friendship.

Many other colleagues became friends along the way too, and they all helped me become a better person and a better scientist. So Antonios, Federica, Chryssa, Kyros, Xenia, Chris and the rest of the extended Crisanti lab family, thank you, you're the best. A special thank you also goes to Teresa, Ioanna and Thomas for their support that made this thesis better.

I was fortunate to be closely surrounded, despite the distance, by my three best friends, Katja, Matic and Črt, like always - *hvala za mehurčke*.

To my Ana, I could never thank you enough for your love, for your patience and your support. You inspired me and kept me going. *Hvala.*

Hvala tudi vama, ati in Miša, za vajino brezpogojno podporo, ki sem jo vedno deležen na vsakem koraku.

Finally, this thesis is dedicated to my mum, Mateja, who I lost during the final year of my PhD. She was the strongest and most loving woman I know.

Declaration of Originality

I hereby declare that the project presented in this thesis is my own work. It is the result of grouped discussions with my supervisors and my progress review panel.

Laboratory work that included molecular cloning, embryo microinjections, mosquito rearing and DNA extractions for validation of *in silico* results obtained in the thesis was carried out by Crisanti lab members Ioanna Morianou, Teresa Petrucci and Thomas Courty.

Additional sources of information that have been used are duly cited.

Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-NonCommercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Contents

DECLARATION OF ORIGINALITY **4**

COPYRIGHT DECLARATION **5**

CONTENTS **6**

LIST OF FIGURES **9**

LIST OF TABLES **12**

LIST OF ACRONYMS **13**

PUBLICATIONS **14**

ABSTRACT **16**

1 INTRODUCTION **18**

1.1 Malaria 19

1.1.1 Symptoms and complications 20

1.2 Malaria vectors 22

1.2.1 Mosquito biology and life cycle 22

1.2.2 *Anopheles* species 23

1.2.3 Mosquito genomics 24

1.3 Malaria control interventions 28

1.3.1 Antimalarial drugs 28

1.3.2 Vaccines 29

1.4 Vector control strategies 32

1.4.1 Indoor residual spraying (IRS) and insecticide-treated nets (ITN) 32

1.5 Genetic approaches to vector control 35

1.5.1 Self-limiting strategies 36

1.5.1.1 Sterile insect techique - SIT 36

1.5.1.2 Release of insects carrying a dominant lethal gene – RIDL 38

1.5.2 Self-sustaining strategies 39

1.5.2.1 Maternal-effect selfish genetic elements - Medea 39

1.5.2.2 Genetic underdominance 40

1.5.2.3 Meiotic drive and distorting sex ratio 43
1.6 Homing-based gene drives 46
1.7 CRISPR as the endonuclease platform 50
1.7.1 CRISPR-Cas systems 50
1.8 Challenges and limitations 57
1.8.1 Gene drive resistance 57
1.8.2 Error-prone repair of double-strand breaks: NHEJ and MMEJ 58
1.8.3 Regulation of cutting activity 62
1.8.4 Overcoming resistance 63

2 AIMS 66

3 EXPLORING ALTERNATIVE CRISPR-CAS NUCLEASES FOR GENOME EDITING AND THEIR SPECIFICITIES 67

3.1 Identification of PAM preferences of CRISPR-Cas nucleases 68
3.1.1 PAM recognition pipeline 68
3.1.2 PAM identification of uncharacterized Cas9 orthologs 73
3.1.3 Validation 78
3.1.4 Putative PAM sequence identified in a previously uncharacterized Cas9 ortholog 84
3.1.5 *In vitro* confirmation of the specificities of the PAMs 86
3.2 Assessing the specificity of novel CRISPR-Cas nucleases in the *An. gambiae* genome 89
3.3 Discussion 94
3.4 Methods 99
3.4.1 Dataset 99
3.4.2 Target search 100
3.4.3 Extraction and visualization of flanking regions 100
3.4.4 BfCas9 and LrCas9 gRNA design and testing for *in vitro* activity 101
3.4.5 Genome scan 103

4 RATIONAL SELECTION OF GENOMIC TARGET SITES FOR CRISPR-BASED GENE DRIVES 105

4.1 Identification of ultra-conserved genomic regions 106
4.1.1 Conservation score calculation 107
4.1.2 Conservation across the *An. gambiae* genome 110
4.1.3 Conservation at the *doublesex* gene drive target site 112

4.2 Prediction of on-target site mutations 119
4.2.1 Simulating MMEJ deletions 120
4.2.2 Testing the MMEJ outcomes 123
4.3 Target site selection in the genes of interest 132
4.3.1 Guido - *in silico* gRNA designer 135
4.3.2 Gene drive target sites in the female-fertility genes and sex-determination pathway 141
4.3.2.1 AGAP007280 142
4.3.2.2 AGAP005958 147
4.3.2.3 AGAP011377 151
4.3.2.4 AGAP004050 - *doublesex* 155
4.4 Discussion 159
4.5 Methods 165
4.5.1 Identification of ultra-conserved genomic regions 165
4.5.1.1 Genomic data 165
4.5.1.2 Genome alignment 165
4.5.1.3 Sequence identity scans 166
4.5.1.4 Conservation score calculation 166
4.5.1.5 Gene ontology (GO) enrichment analysis 167
4.5.2 Prediction of on-target site mutations 167
4.5.2.1 Molecular cloning 167
4.5.2.2 Mosquito microinjection 170
4.5.2.3 Genetic crosses, gDNA extractions and sequencing 171
4.5.2.4 Data analysis 172
4.5.3 Guido 172
4.5.4 Target sites in genes of interest 173

CONCLUSIONS AND FUTURE DIRECTIONS 174

BIBLIOGRAPHY 176

LICENCES AND PERMISSIONS 195

PUBLICATIONS 196

List of Figures

Figure 1 - Geography, vector status, and molecular phylogeny of 16 anopheline mosquitoes and selected other dipterans....26

Figure 2 - Schematic representation of Medea and Genetic underdominance (GU). Both systems are based on toxin element (red) and antidote element (blue). 42

Figure 3 - Homing endonuclease genes for population suppression and population replacement. ..48

Figure 5 - Overview of endogenous CRISPR-Cas system. 51

Figure 6 - The principal building blocks of CRISPR-Cas system types....53

Figure 7 - The repurposed CRISPR-Cas9 system for genome editing. 54

Figure 4 - MMEJ and NHEJ repair mechanisms. 61

Figure 8 - The PAM recognition pipeline....68

Figure 9 - Comparison in computational pipelines behind Dataset A and Dataset B used in the analysis....74

Figure 10 - Comparison between the number of unique target hits deriving from viral database search by using spacer sequences grouped per bacterial species from dataset A and dataset B....75

Figure 11 - Comparison between the number of unique target hits in viral database and the number of their query spacer sequences from Dataset A. 76

Figure 12 - Comparison between the number of unique target hits in viral database and the number of their query spacer sequences from Dataset B. 77

Figure 13 - Sequence logos from PAM identification pipeline for *Streptococcus pyogenes* Cas9 in Dataset A....80

Figure 14 - Sequence logos from PAM identification pipeline for *Streptococcus pyogenes* Cas9 in Dataset B....82

Figure 15 - A putative PAM sequences identified for Cas9 orthologs from *B. fragilis*, *L. rhamnosus* and *L. rhamnosus* GG strain using *in silico* PAM identification pipeline. 86

Figure 16 - Schematic of the TXTL method. 88

Figure 17 - PAM occurrence density calculated in 100 kb windows for SpCas9, LrCas9 and BfCas9 for each chromosome arm in *An. gambiae* genome AgamP4. 91

Figure 18 - Cumulative distribution function for the distribution of number of off-targets (OT) per target site in the AgamP4 genome for target sites of SpCas9, LrCas9 and BfCas9. 92

Figure 19 - Venn diagrams for exons and genes targetable by SpCas9, LrCas9 and BfCas9. 93

Figure 20 - crRNA and tracrRNA hybridization. 102

Figure 21 - Average sequence identity in *An. gambiae* AgamP4 reference genome. 111

Figure 22 - Sequence identity at the gene drive target site locus in *doublesex*. 113

Figure 23 - SNP density, PhyloP and conservation score (Cs) at the *doublesex* target site. 114

Figure 24 - CDF function for positive conservation score (Cs) values per chromosome arm. 115

Figure 25 - Gene ontology enrichment for the most conserved genes in *An. gambiae* genome. 117

Figure 26 - Graphical representation of *in silico* MMEJ prediction algorithm. 121

Figure 27 - Selection of gRNAs for *in vivo* testing based on their MMEJ properties. 124

Figure 28 - Genetic crosses for *in vivo* MMEJ experiment. 125

Figure 29 - Abundance of NHEJ and MMEJ deletions at the target sites T5, T6 and T7. 126

Figure 30 - EJ deletions observed at the target site T6 and T7. 128

Figure 32 - Predicted MMEJ-associated deletions, their MH scores and observed frequency at the target sites T6 and T7. 129

Figure 31 - Rank correlation between predicted MMEJ score and observed MMEJ deletion frequency for target sites T6 and T7. 130

Figure 33 - Candidate gRNA target sites for *AGAP007280. AGAP007280* gene with 5 highest-ranking gRNAs in yellow and gRNA used by Hammond et al. (2016) in red. 144

Figure 34 - Variation at target site of the highest-ranking gRNA-526 in *AGAP007280*. 146

Figure 35 - Candidate gRNA target sites for *AGAP005958*. *AGAP005958* gene with 5 highest-ranking gRNAs in yellow and gRNA used by Hammond et al. (2016) in red. 148

Figure 36 - Variation at target site of the highest-ranking gRNA-526 in gene *AGAP005958.* 150

Figure 37 - Candidate gRNA target sites for *AGAP011377*. 152

Figure 38 - SNPs found in the sequence targeted by gRNA-64 and gRNA-70 in gene *AGAP011377.* 154

Figure 39 - Candidate gRNA target sites for AGAP004050-RB (*dsxF*). 156

Figure 40 - Variation at target site of the highest-ranking gRNA-1151 and gRNA-1152 used by Kyrou et al. (2018) in exon 5 of gene *AGAP004050 (dsx)* 158

Figure 41 - Schematic of p125-Gypsy plasmid 172

List of Tables

Table 1 - Predicted PAM sequences for *S. pyogenes* Cas9 based on filtering strategies for Dataset A and Dataset B. 81

Table 2 - Potential target sites in *An. gambiae* AgamP4 genome for three different Cas9 orthologs: SpCas9, LrCas9, BfCas9 90

Table 3 - Number of off-target sites predicted *in silico* and detected *in vitro* and *in vivo* at the gene drive target sites in *AGAP004050* and *AGAP007280*. 138

Table 4 - List of criteria used in gRNA ranking. 140

Table 5 - Candidate gRNAs for *AGAP007280*. 145

Table 6 - Candidate gRNAs for *AGAP005958*. 149

Table 7 - Candidate gRNAs for *AGAP011377*. 153

Table 8 - Candidate gRNAs for *AGAP004050*. 157

Table 9 - Oligo pairs used for GoldenGate cloning of T5, T6, T7 and T8 gRNAs. 170

List of Acronyms

- ACTs - Artemisinin-based Combination Therapies
- BfCas9 - *Bacteroides fragilis* Cas9
- bp - base pairs
- Cas - CRISPR-associated protein
- CDS - Coding Sequence
- CRISPR - Clustered Regularly Interspaced Short Palindromic Repeats
- Cs - Conservation score
- DSB - Double-Strand Break
- *dsx* - *doublesex*
- EJ - End-Joining
- gRNA - guide RNA
- HDR - Homology-Directed Repair
- HE - Homing Endonuclease
- HEG - Homing Endonuclease Gene
- ITNs - Insecticide-Treated Nets
- LrCas9 - *Lactobacillus rhamnosus* Cas9
- MH - Microhomology
- MMEJ - Microhomology-Mediated End-Joining
- NHEJ - Non-Homologous End-Joining
- OT - Off-Targets
- PAM - Protospacer Adjacent Motif
- PI domain - PAM-Interacting domain
- PTS - Potential Target Sites
- RIDL - Release of Insects carrying a Dominant Lethal gene
- SpCas9 - *Streptococcus pyogenes* Cas9
- SIT - Sterile Insect Technique
- UTR - Untranslated Region

Publications

Part of the results obtained in this thesis was published in Kranjc et al. (2021). The publication is attached at the end of the thesis.

Additional analyses relevant to this thesis were carried out alongside the main work for the thesis and were published in Hammond et al. (2017), Kyrou et al. (2018), Hammond et al. (2021) and Garrood et al. (2021).

Kranjc, Nace, Andrea Crisanti, Tony Nolan, and Federica Bernardini. 2021. "Anopheles Gambiae Genome Conservation as a Resource for Rational Gene Drive Target Site Selection." *Insects* 12 (2): 97. https://doi.org/10.3390/insects12020097.

Hammond, Andrew, Kyros Kyrou, Marco Bruttini, Ace North, Roberto Galizi, Xenia Karlsson, Nace Kranjc, et al. 2017. "The Creation and Selection of Mutations Resistant to a Gene Drive over Multiple Generations in the Malaria Mosquito." Edited by Nick H. Barton. *PLOS Genetics* 13 (10): e1007039. https://doi.org/10.1371/journal.pgen.1007039.

Kyrou, Kyros, Andrew M Hammond, Roberto Galizi, Nace Kranjc, Austin Burt, Andrea K Beaghton, Tony Nolan, and Andrea Crisanti. 2018. "A CRISPR–Cas9 Gene Drive Targeting Doublesex Causes Complete Population Suppression in Caged Anopheles Gambiae Mosquitoes." *Nature Biotechnology* 36 (11): 1062–66. https://doi.org/10.1038/nbt.4245.

Hammond, Andrew, Xenia Karlsson, Ioanna Morianou, Kyros Kyrou, Andrea Beaghton, Matthew Gribble, Nace Kranjc, et al. 2021. "Regulating the Expression of Gene Drives Is Key to Increasing Their Invasive Potential and the Mitigation of Resistance." *PLOS Genetics* 17 (1): e1009321. https://doi.org/10.1371/journal.pgen.1009321.

Garrood, William T., Nace Kranjc, Karl Petri, Daniel Y. Kim, Jimmy A. Guo, Andrew M. Hammond, Ioanna Morianou, et al. 2021. "Analysis of Off-Target Effects in CRISPR-Based Gene Drives in the Human Malaria Mosquito." *Proceedings of the National Academy of Sciences* 118 (22). https://doi.org/10.1073/pnas.2004838117.

Abstract

Discovery and adaptation of CRISPR-Cas systems for genome editing have allowed us to gain an efficient and yet simple tool for genetic manipulation in various fields of molecular biology and biotechnology. One of the most promising applications is the use of CRISPR-Cas9 endonuclease for gene drive systems as a population control strategy for various insect pests of medical and agricultural importance. Use of CRISPR-Cas9 endonuclease in gene drive applications has shown great promise in the laboratory, particularly for the control of *Anopheles gambiae*, the major vector of malaria. However, the performance of such gene drives can be limited by the range of available target sequences and by a propensity of existing endonuclease formulations to generate resistant mutations that hinder the gene drive's efficiency. To expand the flexibility of gene drive systems, computational analysis was performed to identify additional Cas9 orthologs and their specificities that could usefully augment the targeting range of endonuclease-based gene drives. Two alternative variants of CRISPR-Cas endonucleases found in the bacterial species *Lactobacillus rhamnosus* and *Bacteroides fragilis* were assessed for their potential to expand the targeting space in the genome *Anopheles gambiae*. In addition, a computational tool was developed that evaluates neighbouring sequences to the target site to measure both its likely functional constraint and its likely propensity for DNA repair that could generate in-frame alleles. Using this approach we were able to generate a prioritized list of *Anopheles gambiae* target sites for gene drive applications that are less likely to be compromised by resistant alleles.

1 Introduction

1.1 Malaria

Malaria is a life-threatening, infectious disease responsible for more than 400,000 deaths annually. It is estimated that there were more than 229 million cases recorded in 2019 only, a number that has been increasing again since 2015. The age group most affected by the disease are children, accounting for 247,000 deaths in 2019. Other groups with a high risk of contracting malaria and developing the severe disease are pregnant women, HIV/AIDS patients, non-immune immigrants, mobile populations and travellers (WHO, 2021).

The heaviest burden of malaria is in sub-Saharan Africa, where only six countries accounted for more than half of all deaths globally. Almost a quarter of all malaria cases were recorded in Nigeria (23%), 11% in the Democratic Republic of the Congo, 5% in the United Republic of Tanzania, followed by Burkina Faso, Mozambique and Niger with 4% of cases each (WHO, 2020).

Malaria is caused by the parasites of the genus *Plasmodium* and transmitted through bites of infected female mosquitos from the *Anopheles* genus (WHO, 2021). Only five species of *Plasmodium* can infect humans and cause illness: *P. falciparum, P. malariae, P. vivax, P. ovale* and *P. knowlesi*. The majority of malaria infections are caused by two species of parasite, *P. falciparum* and *P. vivax*. The infection cycle begins with an infected female *Anopheles* mosquito bite. During the mosquito bloodmeal, *Plasmodium* sporozoites are injected from the mosquito salivary glands into the skin of the mammalian host. *Plasmodium* sporozoites then migrate through the skin cells and in the bloodstream until they reach the liver, where they invade the hepatocytes.

The liver stage of infection is an obligatory step in the life cycle. In the hepatocytes, they form a parasitophorous vacuole where they undergo the first round of asexual replication. The liver stage is concluded when multinucleated exo-erythrocytic schizonts (or meront), filled with *Plasmodium*, are released back to the bloodstream as clustered vesicles called merosomes (Sturm et al., 2006). Merosomes eventually burst, and the parasites, now called merozoites, infect the erythrocytes. The second asexual replication of merozoites undergoes in the red blood cells, which leads to their destruction.

A small proportion of asexual parasites enter the sexual cycle and produce male and female gametocytes, ingested by a female *Anopheles* mosquito during a blood meal. Once the infected erythrocyte enters the mosquito midgut, male and female gametocytes exit the red blood cell and fuse together into a zygote. The zygotes transform into motile and elongated ookinetes, which allows them to penetrate the midgut epithelium, after which they form into an oocyst upon contact with the basal lamina.

During the third asexual replication, the oocyst grows and ruptures, releasing thousands of sporozoites into the haemolymph. Sporozoites then make their way to the salivary glands of the mosquito, where they wait until the mosquito bites another host and the cycle starts again (Cowman et al., 2012; Venugopal et al., 2020).

1.1.1 Symptoms and complications

The first symptoms of malaria, regardless of different *Plasmodium* species, are often flu-like. Fever and shivers are distinguishing characteristics of any malaria infection. Upon the first symptoms, the infection is confirmed with a positive parasitological test (microscopy or rapid diagnostic test). Cases where the infection is confirmed

but no other significant symptoms are present are classified as uncomplicated malaria cases. The more severe form of the disease can develop in non-immune subjects. It can include symptoms like impaired consciousness or unrousable coma, inability to walk, failure to feed, multiple convulsions, respiratory distress, circulatory collapse or shock, vital organ dysfunction, haemoglobinuria, abnormal spontaneous bleeding and pulmonary oedema. Cases with one or more such symptoms are classified as severe malaria (Bartoloni and Zammarchi, 2012).

1.2 Malaria vectors

1.2.1 Mosquito biology and life cycle

There are approximately 3,500 species of mosquitoes (family Culicidae, order Diptera), divided into two subfamilies: Anophelinae and Culicinae. The main vectors of malaria, *Anopheles* species, are part of the Anophelinae subfamily. In contrast, Culicinae subfamily contains other major human disease vectors, like *Aedes aegypti*, responsible for Dengue fever, Zika fever, Chikungunya virus and yellow fever. Mosquitoes have distinctive developmental stages, starting from larval, pupal to the adult stage. *Anopheles* species deposit their fertilised eggs on the water surface, and after two days, they hatch into the larval stage. Larvae undergo four instar larval stages in approximately 6 to 7 days before pupating. The complete metamorphosis happens during two days of the pupal stage, upon which adult mosquitoes emerge from the water and rest on the surface for a couple of hours until their outer cuticle hardens. Mating happens typically at dusk when an adult female mosquito enters a swarm made up of adult males. Females usually mate only once in their life, and after copulation, females leave the swarm, with sperm stored in spermatheca for the rest of their lives. Unlike males who feed exclusively on nectar, female mosquitoes also need a blood meal, a source of proteins and energy for egg production. *Anopheles gambiae* females will blood feed almost exclusively on human blood, irrespective of their mating status. After two to three days after the blood meal, the female will search for a body of calm water to lay the eggs. During their lifespan, females will go through three or four gonotrophic cycles (blood meal, egg maturation, oviposition) and can lay up to 200 eggs in each cycle.

1.2.2 *Anopheles* species

More than 400 species of *Anopheles* mosquito have been identified, out of which around 30 are malaria vectors (WHO, 2021). *Anopheles gambiae sensu stricto*, *Anopheles coluzzii*, *Anopheles arabiensis* and *Anopheles funestus* are the five species that are the world's most important vectors of malaria, widespread across the sub-Saharan region of Africa where the burden of malaria is the highest. *An. gambiae sensu stricto*, *An. coluzzii* and *An. arabiensis* belong to the *Anopheles gambiae* complex (*Anopheles gambiae sensu lato*) (Coetzee et al., 2013; Davidson, 1962) whereas *An. funestus* is a part of *An. funestus* species group (Derua et al., 2015).

Additional to *An. gambiae s.s.*, *Anopheles coluzzii*, *An. arabiensis* and *An. funestus*, the *Anopheles gambiae* complex consists of 6 more species: *An. bwambae, An. melas, An. merus, An. quadriannulatus, An. amharicus, An. comorensis.*

The species of the complex have different ecological properties, vectorial capacity, host feeding preferences and can be found in other geographic locations across the continent. Three of the species from the complex, *An. melas, An. merus* and *An. bwambae* have the capacity to lay their eggs in saltwater as opposed to freshwater prefered by the rest. *An melas* and *An. merus* can be found mainly in the coastal areas, whereas *An. bwambae* is known only from the vicinity of hot springs in Bwamba County in the west of Uganda (White, 1985). Their efficiency to transmit malaria is considerably lower compared to *An. gambiae, An. coluzzii* or *An. arabiensis.* Becouse of the zoophilic nature of *An. quadriannulatus* and *An. amharicus* and geographic confinement of *An. comorensis* to the island of Grande Comore, these species are not considered as major vectors of human malaria (Coetzee et al., 2013; Sinka et al., 2012). Besides the four major malaria vectors in *Anopheles gambiae* complex,

anthropophilic *Anopheles funestus* is widespread across a large part of sub-Saharan Africa and poses a great danger to the public health (Sinka et al., 2012).

It is estimated that species of *Anopheles gambiae* complex diverged approximately 0.5 MYA (Thawornwattana et al., 2018). Autosomal introgression was shown to be extensive between the species of the complex, while the X chromosome is largely resistant to introgression (Coluzzi et al., 1979; Fontaine et al., 2015; Slotman et al., 2005). Reproductive barriers that limit gene flow between divergent species can act pre-zygotically or post-zygotically. Pre-zygotical reproductive barriers are evident in different breeding habitats (freshwater and saltwater) and in observed assortative mating, where species found in the same area or even in the same mating swarms show low hybridisation despite mixed swarming (Diabaté et al., 2006). Some evidence suggest that wing beat frequencies may play a role in recognising members of the same species (Charlwood and Jones, 1979). Post-zygotic isolation mechanisms include F1 hybrid sterility between all species of the complex, except between *An. gambiae* and *An. coluzzii,* which are the youngest species to diverge and were previously known as two cytological forms named S and M (Savanna and Mopti, respectively) (Bernardini et al., 2017; Coetzee et al., 2013; Davidson, 1962).

1.2.3 Mosquito genomics

Advancements in genome sequencing methods in the late 1990s and early 2000s offered the first genomic sequences of several model organisms in the span of a few years. In 2000 the genome of the fruit fly, *Drosophila melanogaster* (Adams et al., 2000) was published, paving the way for the genome of *An. gambiae*. The first mosquito genome to be published in 2002 (Holt et al., 2002) was the PEST strain, a hybrid population of *An. gambiae s.s.* and *An. coluzzii* (Coetzee et al., 2013).

An updated version of the assembly (AgamP3) was later proposed. Major scaffolds were assigned to chromosome arms X, 2R, 2L, 3R and 3L based on physically mapped BAC clones and overlaps from scaffold-to-scaffold sequence comparisons (Sharakhova et al., 2007). Additionally, the mitochondrial genome was added to the assembly (AgamP4), adding up the size of the genome to approximately 264 Mb of non-redundant sequences. The genome assembly covers over 90% of the genome and holds around 14,000 annotated genes. None of the available scaffolds from the Y chromosomes has been physically mapped, and the Y chromosome still is unassembled due to the high density of repetitive sequences.

Genome assemblies of other *Anopheles* mosquitoes followed the publication of the reference genome of *An. gambiae*. The genome of *An. darlingi* offered insights into the ~100 million years of evolutionary divergence between African and South American malaria vectors (Marinotti et al., 2013). Assembly of the genome of the key urban vector of malaria, *An. stephensi*, showed that its genome landscape has relatively low repeat content compared to *An. gambiae* (Jiang et al., 2014).

Sequencing of 16 anopheline genomes (Figure 1) from various geographic locations resulted in novel reference genomes for *Anopheles* species (Neafsey et al., 2015).

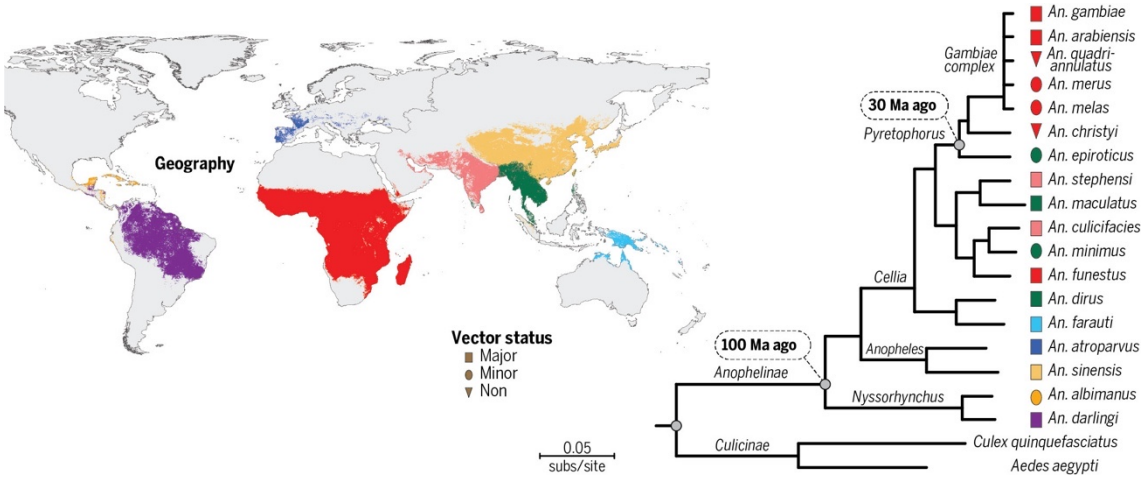


Figure 1 - Geography, vector status, and molecular phylogeny of 16 anopheline mosquitoes and selected other dipterans. From Neafsey et al., 2015. Reprinted with permission from AAAS.

The results from the 16 Genomes Consortium (Neafsey et al., 2015) provided evidence of functional variation among the *Anopheles* species that determine vectorial capacity.

When *An. gambiae* genes implicated in the immunity response to parasite infection were compared to other background genes it was shown that they have, on average, higher evolutionary rates and are more often duplicated. Genes responsible for parasite recognition, as well as effector enzymes, exhibit relatively low levels of sequence divergence and show high nucleotide sequence conservation. Genes involved in signal transduction are rarely duplicated and could be found across all species and cascade modulation genes are lineage-specific with higher rates of sequence divergence and are more often duplicated. Furthermore, a significant difference was seen between rearrangement rates on the autosomes compared to the X chromosome. Especially low sequence alignability was observed in centromeres and on the X chromosome when compared to *An. gambiae* reference genome. The gene shuffling was attributed to multiple families of transposons and retroelements found in the assembled genomes.

Perhaps most importantly, the results supplied a robust framework for comparative analysis across the genus and expedited the research of different *Anopheles* species (Fontaine et al., 2015; Kyrou et al., 2018).

1.3 Malaria control interventions

The vision of WHO and the global malaria community is a world free of malaria by 2030 (WHO and Global Malaria Programme, 2015). The ambitious yet feasible goals are to reduce malaria mortality rates and to reduce malaria case incidence globally. Additionally, the WHO strategy aims to eliminate malaria from at least 35 countries in which malaria was transmitted in 2015 and prevent re-establishment of malaria in all countries that are malaria-free in 2030.

The strategy is supported by three pillars, which guide global efforts to move closer to malaria elimination: to ensure universal access to malaria prevention, diagnosis, and treatment; to accelerate efforts towards elimination and attainment of malaria-free status, and to transform malaria surveillance into a core intervention. Two crucial aspects of malaria prevention and treatment are chemoprevention and vector control. The former includes preventive treatment strategies like antimalarial drugs and vaccines, and the latter aims to reduce the vectorial capacity of vectors transmitting malaria (WHO and Global Malaria Programme, 2015).

1.3.1 Antimalarial drugs

The main purposes of antimalarial drugs are to eliminate the infection from the patient's body and to prevent the further spread of the parasite to others. Most of the drugs target asexual erythrocytic stages of the parasite.

The first antimalarial drug discovered and extracted in the 17th century from the cinchona tree was quinine, which proved to be one of the most effective treatments for malaria (Achan et al., 2011). In the 1940s, synthetic drug chloroquine (CQ) was discovered and showed great promise and fewer side effects than quinine. Thus, it

became the most widely used antimalarial drug. However, the parasite started to show resistance after the extensive use, which imposed great selection pressure on the parasite (D et al., 1996). Due to its decreased efficiency, it was eventually removed as a first-line antimalarial drug. Several other compounds were subsequently developed in the 20th century that showed great potential, including sulfadoxine/pyrimethamine (SP), amodiaquine, mefloquine and atovaquone but parasite resistance followed shortly thereafter and spread throughout Southeast Asia and further to central Africa (Beshir et al., 2010; Cowman et al., 1994; Mita et al., 2007; Vaidya and Mather, 2000).

The current first line of defence is Artemisinin-based combination therapies (ACTs) in all countries where malaria is present. Unfortunately, like CQ- and SP-resistant parasites, artemisinin resistance was reported in Southeast Asia (Noedl et al., 2008). The resistance was conferred by a single point mutation in the »propeller« domain of the *Pfkelch13* gene (Ariey et al., 2014; Ashley et al., 2014). Parasites with the resistance-conferring mutation showed reduced haemoglobin endocytosis, which diminishes ACT susceptibility, resulting in resistance (Birnbaum et al., 2020). The resistance to ACTs is not yet prevalent in Africa and ARTs continue to offer good defence against the parasite despite the evidence of mutations in the *Pfkelch13* gene (Kamau et al. 2015; WWARN Artemisinin-based Combination Therapy (ACT) Africa Baseline Study Group 2015). However, recent reports of *de novo* mutations and the emergence of ACT resistance, so far localised in Rwanda, pose danger to the success of antimalarial chemotherapy in Africa (Uwimana et al., 2020).

1.3.2 Vaccines

Malaria vaccine development was propelled by successful studies on protective immunity in mice by vaccinations with irradiated sporozites (Nussenzweig et al.,

1967). Modern vaccine development focuses on targeting parasite antigens in different stages of the malaria parasite life-cycle. Transmission-blocking vaccines (TBV) target the parasite's sexual stages, intending to prevent their transmission to mosquitoes. Blood-stage vaccines (BSV) target asexual replication of merozites in the blood stages of the parasite life cycle. The third class of malaria vaccines acting against the pre-erythrocytic stages (PEV) have shown the most success so far (Duffy and Patrick Gorres, 2020).

The most extensively tested malaria vaccine candidate against *P. falciparum* malaria is a PEV named RTS,S/AS01. The antigenic target of recombinant RTS,S vaccine is the circumsporozoite protein (CSP) on the surface of sporozoites. RTS,S vaccine has been the only malaria vaccine that has undergone the large-scale phase 3 of clinical trials in Africa. Unfortunately, over the time the trials have shown low efficacy of only around 30% against severe malaria in children (RTS,S Clinical Trials Partnership, 2015). The low efficacy among vaccinated infants and the rapid decrease in immunity provide little promise for this vaccine to be the single tool in the malaria eradication effort (Mahmoudi and Keshavarz, 2017).

Despite disappointing results after 20 years of development and trials, RTS,S vaccine was deemed safe and efficacious enough for reducing clinical malaria in African children and will be reviewed in 2022 by WHO for broader use in Africa. Newer PEV and TBV candidates are currently progressing in the pipeline of clinical trials and will play an essential role in malaria eradication together with other malaria control tools (Duffy and Patrick Gorres, 2020). It has been reported recently that a novel PEV vaccine named R21 was developed and reached the WHO-specified efficacy goal of 75% in the target population of African children from areas of highly seasonal malaria transmission. The R21 vaccine is a virus-like particle based on the C-terminal portion of the CSP from *P. falciparum* fused to the N-terminus of Hepatitis

B surface antigen (HBsAg). It is similar to RTS,S but lacks the excess HBsAg found in RTS,S and provides higher density of circumsporozoite epitope on the particle surface, resulting in high levels of malaria-specific antibodies (Datoo et al., 2021). The development and obtained results of R21 have reinvigorated the hopes of possessing an effective vaccine and is seen as an important step forward in malaria vaccine development. Nonetheless, the study by Datoo et al. (2021) is only the first step in a long process of deploying the vaccine for broader use.

1.4 Vector control strategies

Malaria as a vector-borne disease presents an opportunity for transmission control not only through controlling the parasite but also through controlling its vector. The definition of vector control includes any kind of measures directed against the vector with the intention to limit or prevent its ability to transmit the disease. Historically, the biggest leaps in malaria eradication were achieved through malaria vector control, which in effect removes or suppresses the mosquito populations in the human-populated areas. These strategies of vector control are summarised below.

1.4.1 Indoor residual spraying (IRS) and insecticide-treated nets (ITN)

The discovery of dichloro-diphenyl-trichloroethane (DDT) in 1939 and the recognition of its insecticidal properties in the 1940s led to its widespread use for indoor residual spraying (IRS). DDT proved to be very effective for IRS due to its contact toxicity and as a repellant against mosquitoes when applied to the walls of indoor dwellings where people sleep. Following the success of the malaria reduction by IRS in the 1940s, the Global Malaria Eradication Campaign was introduced by WHO. Its activities directly contributed to malaria eradication in the countries with lower transmission intensity, mainly in Europe and the Americas. By 1978 malaria was eradicated from 37 of the 143 countries where malaria was endemic at the time when the Global Malaria Eradication Campaign started (Raghavendra et al., 2011).

DDT spraying was subsequently replaced by other insecticides, which are less toxic in humans but retain strong insecticide properties. Only 4 directly lethal insecticide

classes are recommended for control of adult malaria vectors with ITNs or IRS: pyrethroids, organochlorines, carbamates and organophosphates (WHO, 2012).

Pyrethroids, the most used insecticides for malaria vector control, are applied in various settings, including IRS, but mostly serve as the active ingredient for impregnating bed nets (insecticide-treated nets - ITNs). ITNs and long-lasting insecticide-treated bed nets (LLINs) quickly became the most widely adopted and most effective preventive measure in high-burden countries (WHO, 2020). It was estimated that ITNs alone were responsible for 68% reduction of cases between 2000 and 2015, which together with IRS effectively prevented around 500 million cases in those 15 years (Bhatt et al., 2015).

More than 250 million ITNs were delivered to malaria endemic countries in 2019 and the percentage of homes in sub-Saharan Africa with at least one ITN per two people in the household rose to 68%. However, the availability of ITNs, which are exclusively treated with pyrethroids, has also increased selection pressure on the anthropophilic mosquitos and by 2019 a total of 72 countries reported resistance to at least one insecticide in at least one malaria vector. The most prevalent in 2019 was the resistance to pyrethroids (69.9% of the sampling sites for which data were available), followed by organochlorines (63.4%), carbamates (31.7%) and organophosphates (24.9%) (WHO, 2020).

Multiple complex mechanisms of resistance to insecticides have been identified in mosquitoes. Resistant mosquitoes showed increased metabolic detoxification of insecticides and decreased sensitivity of the insecticide target sites (Liu, 2015).

Decreased sensitivity to the insecticides is conferred by mutations in the voltage-gated sodium channel in neurons, which is the target of both DDT and pyrethroid insecticides. They bind to the voltage-gated sodium channels and keep them open

for longer time than usual, prolonging the action potential and eventually knocking down the insects. So called knock-down resistance (*kdr* genotype) arises from point-mutations in the gene, which in consequence decrease the binding affinity of DDT or pyrethroids. Resistant mutations effectively enable insects exposed to pyrethroids to withstand the delays in the closing of the sodium channel (Davies et al., 2007; Liu, 2015; Ranson et al., 2011).

Metabolic resistance in mosquitoes in mainly enabled by overexpression of genes from cytochrome P450 family (*Cyp* genes), primary enzyme family responsible for pyrethroid metabolism in insects. The overexpression of *Cyp* genes was linked to amplification of *Cyp* genes and their copy number variation (CNV) can serve as a marker for CYP-mediated pyrethroid resistance (Lucas et al., 2019; The Anopheles gambiae 1000 Genomes Consortium, 2020; Weetman et al., 2018). Introduction of novel insecticides and the surveillance of resistance, through single nucleotide polymorphism (SNP) markers in pyrethroid-targets and CNV markers for metabolic resistance will play instrumental role in modifying the approaches that are currently used for vector control in the malaria endemic countries.

1.5 Genetic approaches to vector control

Genetic control strategies rely on vertical transmission of a desired heritable element (or a genetic modification) through sexual reproduction of a modified mosquito. Different approaches have been proposed in which these heritable elements could decrease mosquito's vectorial capacity by either introducing mosquitoes incapable to carry parasite (population replacement) or by reducing the number of competent mosquitoes (population suppression) and in result decrease the transmission of parasite (Alphey, 2014).

The methods developed for creating the first transgenic *Anopheles* mosquito (Catteruccia et al., 2000) and the genomic landscape uncovered by the first *Anopheles* genome assembly (Holt et al., 2002) opened a new frontier in malaria vector control. The prospect of controling mosquito populations solely by the use of ITNs has been hampered by increasing problem of insecticide resistance and the development of novel genetic interventions to complement malaria vector control has received a lot of interest since.

Both strategies, population supression and replacement, can have different degree of invasiveness and persistence in the field after the release. The introduced modification in 'self-limiting' strategy is eventually lost from the target population unless new modified mosquitoes are introduced. A more invasive approach is 'self-sustaining' strategy where modified mosquitoes tend to increase in frequency in the population after the release.

Population replacement refers to the strategies that target vector competence with the aim to reduce the ability of mosquitoes to transmit a given pathogen. Such strategy could be used for introducing mosquitoes in the wild that prevent the

parasite from developing in the mosquito (Curtis, 1968; Isaacs et al., 2012; Serebrovskii, 1940).

A necessary step in this approach is to identify mosquito genes that permit *Plasmodium* infection (specific receptors used by parasite) or to introduce anti-parasite effector genes. Several genetic promoters in the mosquito genome were elucidated to express such anti-parasite effector molecules in specific adult female tissues where parasite is present during its mosquito stage (reviewed in Carballar-Lejarazú and James 2017).

Population suppression can be achieved by interfering with the survival or reproduction of female mosquitoes or for example with biasing sex ratio towards males. Different approaches for population suppression are discussed more in detail below.

1.5.1 Self-limiting strategies

1.5.1.1 Sterile insect techique - SIT

Sterile insect technique (SIT) is a self-limiting population suppression strategy for vector control. Population suppression in this strategy is achieved through the continuous release of large numbers of reproductively sterile males into the wild population (Knipling, 1955). Sterility is induced by exposing larvae or pupae to ionising radiation or by chemosterilization. Like many other genetic approaches to vector control, SIT is highly species-specific as it relies on sexual reproduction of male and female individuals. When male mosquitoes are released and mate with wild-type females no progeny is produced. This decreases the reproductive potential of females and leads to the local suppression of the mosquito population (Alphey et al., 2009). However, it was shown that males sterilised by gamma-

irradiation obtained significantly fewer matings than their wild-type competitors (Andreasen and Curtis, 2005). To reduce reproductive potential of females in the population and to overcome the reduced fitness of sterile males, the ratio between sterile and wild-type males needs to be sufficiently high. So called "over-flooding" sterile male ratio is needed to impact the reproductive potential of females in the laboratory settings when competitiveness of sterile males is around 50% (Maïga et al., 2014). The sterile male ratio depends on the biology of the target species and the vigor of the sterile lines. Theoretically, the population crashes in 5 generations if the ratio is increased to 9:1 and the initial release size of sterile males is maintained throughout generations (Dyck et al., 2021; Knipling, 1955).

Many SIT interventions have been successfully implemented for pest control across the world in various pest species. A release of 8.5 million sterile males of *Glossina austeni* (tsetse fly) led to a successfully eradication and prevented later re-invasion in Zanzibar (Vreysen et al., 2000). Furthermore, successful eradication of *Cochliomyia hominivorax*, the New World screwworm, was achieved in the USA, Mexico, all of Central America, and Libya. The screwworm larvae attack the healthy tissue of warm-blooded animals and are highly damaging to the livestock industry. The employment of SIT for the screwworm control is considered to be a big success in the pest eradication programs (Scott et al., 2017; Wyss, 2000). Successful containment programs based on SIT have also been implemented for the Mediterranean fruit fly, *Ceratitis capitata*, one of the most destructive fruit pests in the world (Enkerlin et al., 2015) and the codling moth, *Cydia pomonella* (Thistlewood and Judd, 2019).

Despite the clear successes of SIT for different pest species, the progress of developing a comprehensive SIT program for malaria vector control stays slow. The implementation is in part hindered by rudimentary sex separation methods

available as the methods are based on naturally occurring differences between male and female mosquitoes and currently cannot scale to fulfil the needs of a mass-release in a region as vast as Africa (Papathanos et al., 2018). Additionally, the nonself-sustaining nature of SIT and the scale of rearing and mass-releases needed to cover the geographical range of malaria-transmitting *Anopheles* species are the main issues that SIT cannot overcome.

1.5.1.2 Release of insects carrying a dominant lethal gene – RIDL

A more advanced approach that circumvents the issues of SIT was developed and named release of insects carrying a dominant lethal gene – RIDL. RIDL is in some ways similar to SIT – it is a species-specific, self-limiting approach that effectively leads to population suppression through the reduction of female reproductive potential. However, the most important difference is the implementation of a transgene system to induce lethality, unlike SIT where dominant lethal phenotype is achieved by gamma-radiation. RIDL is based on a repressible dominant lethal gene product expressed under a female-specific promoter (female-specific RIDL) or it can be expressed in both females and males (bi-sexual RIDL). Any individual with RIDL element, expressing the toxic effector, will not develop unless its expression is repressed by an antidote. Introducing the antidote (a chemical repressor) as a food additive in the laboratory allows mass rearing of the organism and achieves the desired sex-specific lethality in the offspring when released in the wild population where the repressor cannot be found (Phuc et al., 2007; Thomas et al., 2000).

The reduction of female reproductive potential is reached by continuous releases of transgenic homozygous RIDL males into the wild population. When the released males carrying RIDL construct mate with the wild-type females, their female progeny is expected to die during the larval or pupal stage, depending on the specific promotor that expresses the toxic genetic product. The choice of different

promoters also gives the method extra flexibility to achieve sex-specific, temporal and/or spatial expression of the lethal gene.

RIDL has been widely used to suppress the local population of *Aedes aegypti*, a mosquito vector responsible for yellow fever (Carvalho et al., 2015; Harris et al., 2012; Lacroix et al., 2012; Phuc et al., 2007), and other pest species, namely Med fly *Ceratitis capitata* (Gong et al., 2005) and bollworm *Pectinoptera gossypiella* (Morrison et al., 2012).

1.5.2 Self-sustaining strategies

There are numerous examples of selfish genetic elements in nature that have inspired synthetic designs for population control. Each of these have different characteristics - they can vary in rates of invasion, feasibility of design and, importantly, in threshold frequency. The threshold frequency of a particular self-sustaining system defines how many individuals need to be released into a population to achieve a suppressive effect. Lower threshold frequencies of self-sustaining strategies compared to others also present potential for these strategies to have impact in much larger geographical regions.

1.5.2.1 Maternal-effect selfish genetic elements - Medea

Maternal-effect selfish genetic elements were first described in red flour beetle, *Tribolium castaneum*, which showed unusual property where the progeny of a carrier mother survive only if they inherit a copy of the gene from either parent (Figure 2). This instance of maternal lethality to hatchlings was named *M*aternal-*E*ffect *D*ominant *E*mbryonic *A*rrest - *Medea,* also a sorceress from Greek mythology who killed her progeny fathered by Jason (Beeman et al., 1992; Beeman and Friesen, 1999).

The selfish genetic element consists of two genes, the maternally expressed toxin deposited in the embryo and the antidote, expressed in the zygote carrying Medea element. The offspring that does not inherit the Medea element (from maternal or paternal genome) is killed by the maternally deposited toxin, making the Medea element the only protection to its own lethal maternal effect. This dynamic makes this genetic element selfish as it is inherently increasing in frequency in the population due to the fitness advantage of chromosomes bearing Medea element compared to those that do not.

A synthetic Medea element was inserted in the *Drosophila melanogaster* genome and was able to drive itself into a population, increasing in frequency through generations. For the toxin component the researchers used the miRNA-mediated silencing of a maternally expressed gene essential for embryogenesis. Zygotic expression of a rescuing transgene served as the antidote (Chen et al., 2007). The system was in theory proposed to be used for population replacement in mosquitoes where Medea-like element would be used to drive an effector that prevents parasite infection, however more focus has been given to implementation of other drive systems in mosquitoes, due to the limited understanding how to apply effective RNAi-mediated silencing in the germline of mosquitoes (Champer, Buchman, and Akbari 2016).

1.5.2.2 Genetic underdominance

The principle of genetic underdominance (GU), also sometimes called heterozygous inferiority, is based on lower fitness of heterozygous individuals compared to their homozygous parents. A method with balanced reciprocal chromosomal translocations to impose reduced viability of heterozygotes was theoretically proposed for population control. If a translocation strain was developed that had a disease-refractory gene tightly linked to the translocation break point, disease-

resistance would spread into that population as the translocation fixes (Curtis, 1968; Serebrovskii, 1940). A system where chromosomal translocation imposed 50% sterility in heterozygotes, whilst keeping homozygotes fully fertile, was implemented to suppress a population of *Culex pipiens* (Laven et al., 1972). The strategy gathered a lot of attention but was later abandoned due to the inability to anchor the payload gene to the translocation breakpoints and due to relatively high fitness cost in individuals that had translocations induced by X-rays.

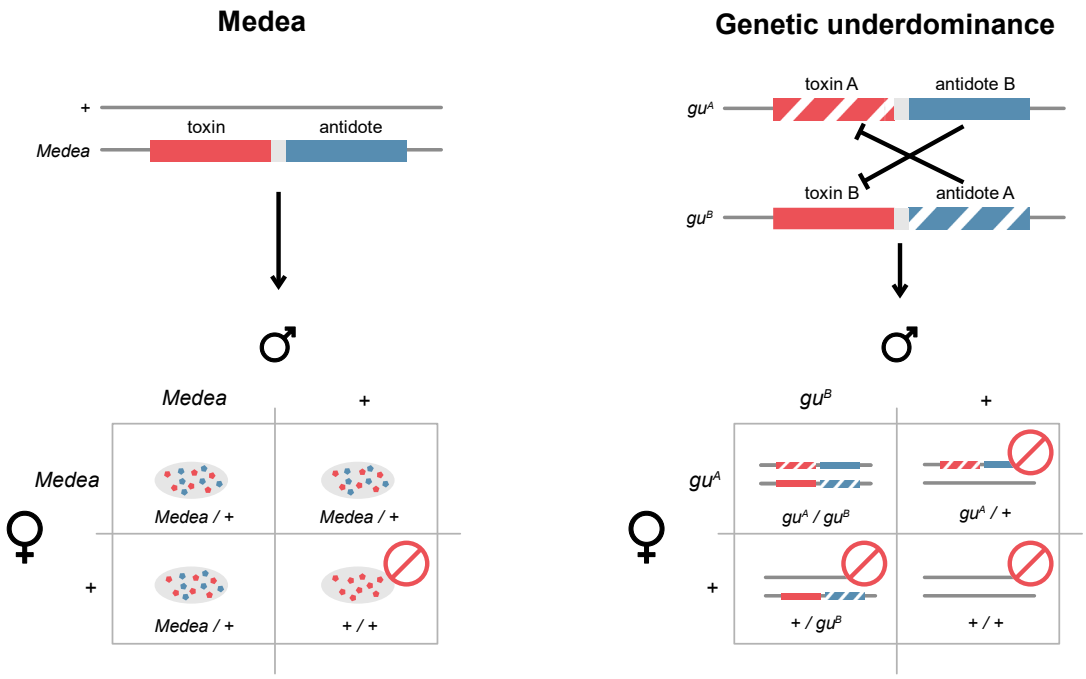


Figure 2 - Schematic representation of Medea and Genetic underdominance (GU). Both systems are based on toxin element (red) and antidote element (blue). In Medea systems a female-expressed toxin is passed to all the progeny by Medea-carrying mother. If the embryo inherited Medea element from any of the parents, the antidote expressed from inherited Medea embryo neutralises the maternally deposited toxin and allows the embryo to survive, unlike wild-type embryos (+/+) that do not carry Medea element and the antidote. In GU systems two different toxins (toxin A and B) and two different antidotes are expressed (antidote A and B). Only embryos that carry both GU elements with both antidotes gu^A and gu^B are able to survive.

Synthetic GU elements carrying toxin and antidote effectors were designed and used for driving the transgenic construct in a population of caged caged fruitfly (*Drosophila melanogaster*). The system uses two maternally expressed toxins from two separated chromosomes and two zygotically expressed antidotes also separated on two chromosomes (Figure 2). Only homozygous individuals with both loci can suppress the maternally deposited toxins (Akbari et al., 2013). The advancement of synthetic biology bring possibility for researchers to generate

translocations at almost any genomic location without irradiation by using RNA-guided endonucleases to induce sequence-specific translocations (Champer et al., 2016; Egli et al., 2004; Jinek et al., 2012).

The main advantage of genetic underdominance over RIDL and SIT systems is that, due to the »selfishness« of engineered underdominance, the construct increases in frequency over generations and allows smaller release thresholds. Additionally, modelling suggests that the transgene can be removed from the population by dilution with wild-type individuals, making the technology reversible (Akbari et al., 2013).

1.5.2.3 Meiotic drive and distorting sex ratio

Another approach was suggested to achieve population suppression for pest control when Y chromosome meiotic drivers have been identified in *Aedes* and *Culex* mosquitoes (Hamilton, 1967; Hickey and Craig, 1966; Sweeny and Barr, 1978). Meiotic drive elements rely on impairing the transmission of specific gametes during meiosis by removing the gametes that do not carry the drive element. When such drive elements are positioned on the sex chromosomes, for example, the Y chromosome meiotic driver, they bias their own inheritance and result in a progeny with a strong male bias. The imbalance in the male to female ratio interferes with reproductive capacity of a population and can lead to population suppression.

A similar strategy was proposed for the control of malaria vectors: a synthetic element positioned on the male-specific Y chromosome that would actively disrupt X-bearing gametes during spermatogenesis, preventing them from contributing to the next generation, and favour its own inheritance (Burt, 2003). A release of transgenic male mosquitoes that produce only Y-bearing sperm was modelled and

showed that their mating with wild-type females could propagate a strong male bias throughout a population and cause its crash. (Deredec et al., 2008)

In the proposed system, the transgenic element selectively cuts the X chromosome during the meiotic stages of spermatogenesis. A naturally occurring endonuclease I-PpoI was found to specifically recognise X-linked ribosomal DNA (rDNA) repeats and was deemed a suitable endonuclease for the incapacitation of X-carrying spermatozoa in *Anopheles gambiae*. The first experiments showed that I-PpoI expressed during spermatogenesis resulted in the bias of Y-bearing spermatozoa in mosquito testes. However, the endonuclease was paternally deposited in the embryo and cleaved maternally inherited X chromosome, causing early mortality of embryos (Windbichler et al., 2008, 2007). A mutated, more labile version of I-PpoI was developed and limited its activity only to spermatogenesis and prevented cleavage of the X chromosome in the zygote. The use of unstable I-PpoI reduced the mortality of embryos and imposed a severe male bias (up to 97%) in the progeny. Cage experiments showed that a release of transgenic males that carry the sex distorter leads to a successful population suppression when released in 3-fold excess (Galizi et al., 2014). The technology was also proven to work with CRISPR-Cas9 endonuclease (Galizi et al., 2016).

To create a self-sustaining and effective sex distorter, suitable for vector control, the transgene with I-PpoI would need to be transmitted in tight linkage to the Y chromosome, such that not only would the offspring be nearly all male, but all of these would also receive a copy of the element. This would allow for a release of transgenic males in much lower frequency and enable the modified Y chromosome to invade the population (Burt, 2003; Deredec et al., 2008). Currently, no such system has been implemented in *Anopheles gambiae* yet. The main obstacle in the way of creating Y-driving sex distorter system is the meiotic sex chromosome

inactivation (MSCI), a process in meiosis that represses almost all gene expression from the sex chromosomes (Hall et al., 2016; Maggert and Golic, 2002; Turner, 2007). A recent study of gene expression during spermatogenesis in *Anopheles gambiae* suggests that there are some genes that evade the silencing caused by MSCI (Taxiarchi et al., 2019). Further characterisation of these genes could offer better insights into mechanism of MSCI evasion and could potentially help to achieve the expression from the Y chromosome for the purpose of developing Y-linked sex-distorter.

1.6 Homing-based gene drives

Gene drive is a genetic phenomenon that describes some genetic elements that can defy the normal Mendelian laws of heredity and in result increases the prevalence of a genetic element (natural or synthetic) in a population, even in the presence of some fitness cost (Burt, 2003). Homing endonuclease genes (HEGs) are naturally occurring genetic elements that can bias their inheritance by copying their own genetic sequence from one chromosome to the other.

In a diploid organism this means that they switch from heterozygosity to homozygosity in a process called 'homing'. Homing occurs through a mechanism in which a homing endonuclease (HE), encoded by HEG, recognises and cleaves a specific 15 to 30 bp long DNA sequence at the position of the wild-type chromosome where the nuclease gene is inserted. When the HEG cleaves the target site, the double-strand break (DSB) induced by cleavage is often repaired by homology directed repair (HDR). (Figure 3)

HDR is one of the cellular mechanisms used to repair a double-stranded break of a chromosome. The homologous chromosome serves as a template for the repair of the induced break. HEGs use this process to their own advantage and copy themselves over to the cleaved wild-type chromosome, effectively converting a heterozygote HEG-carrier into a homozygote.

The homing happens only in heterozygosity when one chromosome carries the HEG allele, and the homologous chromosome has the target site intact. In multicellular organisms, when the HEG is active in the germline it will be able to be present in nearly 100% of the produced gametes instead of the 50% expected under Mendelian

inheritance. This allows the HEG to spread rapidly in the population, increasing in frequency each generation.

Homing gene drives offer an elegant solution to the issue of genetic control of insect population because of their self-sustaining nature and invasive potential. The self-sustaining dynamics of gene drive make it a very promising system for both strategies of malaria vector control, population supression and population replacement. For the latter, a synthetic construct carrying an effector, which for example, interferes with *Plasmodium* life cycle in the mosquito, coupled with a homing endonuclease driving the transgene into the population, could substantially reduce the vectorial capacity of the mosquito population.

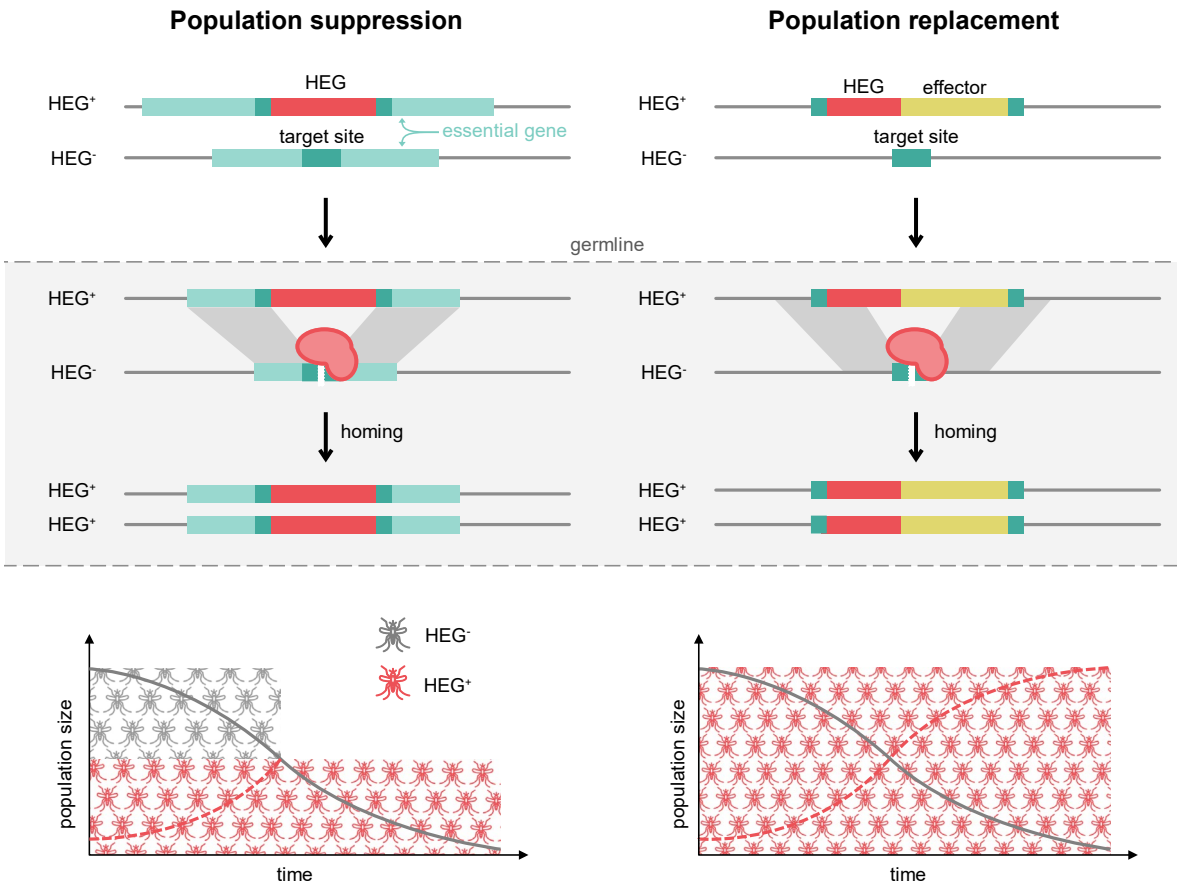


Figure 3 - Homing endonuclease genes for population suppression and population replacement. The HEG is inserted in its own target site (dark green) and, when active in the germline, cuts the wild-type (HEG^-) chromosome. Upon the chromosome cleavage the cell activates homology-directed repair (HDR) mechanism, which uses the intact homologous chromosome carrying HEG (HEG^+) as a template to fix the double-stranded break. In this process called homing the HEG goes from heterozygosity to homozygosity. When homing is active in the germline, the heterozygous individual will produce more HEG-bearing gametes than expected 50% under normal Mendelian inheritance. This enables the HEG to rapidly increase in frequency. If the HEG is inserted into a gene involved in, for example, female fertility, the homing will disrupt its function and spread in the population and effectively reduce its reproductive capacity. For the HEG to effectively spread in the population and lead to its suppression (left) the gene must be functional in heterozygosity. Alternatively, the HEG can be used to drive an effector element (right), that makes the individual refractory to malaria transmission.

Alternatively, to achieve the population suppression, a synthetic construct that presents a recessive lethal phenotype would need to be driven into the population.

A population suppression gene drive should preferably target genes that are involved in fertility, viability, sex differentiation, or other processes that reduce reproductive capacity when disrupted. Importantly, the target gene needs to be required in somatic cells, not required in the germline and fully functional in heterozygosity (haplosufficient) so it can be spread to the next generation without causing sterility or mortality in 'carrier' adults.

The goal of population suppression gene drive systems is to impose high fitness cost in homozygous somatic cells but retain full fitness in heterozygosity. The perfect gene drive will be very efficient in homing in the germline (high homing rate) and leave the heterozygous individuals' fitness high, so they are unobstructed to transmit the gene drive allele to 100% of their progeny when they mate with wild-type individuals.

Initial experiments used the naturally occurring *I-SceI* HEG for the synthetic, homing-based gene drive in *An. gambiae* (Windbichler et al., 2011) and *Drosophila melanogaster* (Chan et al., 2011) showed that mosquito germline chromosomes are amenable to homing. The gene drive alleles successfully homed and invaded the caged populations, increasing in frequency throughout generations. However, the HEGs needed to be re-engineered to be able to target endogenous sites in the relevant target genes. This process was shown to be challenging as the observed homing frequency with engineered HEGs was lower and the process of identifying target sequence specificities of novel HEG variants was cumbersome (Chan et al., 2013; Thyme et al., 2014). More easily programmable and modular options, such as transcription activator-like effector nucleases (TALENs) and zinc finger nucleases (ZFNs) were a good alternative to native HEGs. They were engineered for gene drive in *Drosophila melanogaster* and demonstrated transmission rates up to 70% but failed to remain effective in subsequent rounds of homing. (Simoni et al., 2014)

1.7 CRISPR as the endonuclease platform

In 2014, the emergence of repurposed CRISPR-Cas system in genome editing sparked the interest in using the system for gene drive. Following the essential idea of Burt (2003), Cas9 was proposed as an alternative endonuclease that would drive synthetic homing-based constructs into a mosquito population (Esvelt et al., 2014). The principle was soon after demonstrated in yeast, fruitfly and mosquito (DiCarlo et al., 2015; Gantz et al., 2015; Gantz and Bier, 2015; Hammond et al., 2016). The CRISPR-based gene drives in *An. gambiae* immediately superseeded the transmission rate previously achieved by gene drives using HEGs, TALENs and ZFNs and showed the transmission rates to be nearly 100% (Champer et al., 2017; Hammond et al., 2017).

1.7.1 CRISPR-Cas systems

CRISPR-Cas systems are part of a relatively recently discovered adaptive immunity system found in Bacteria and Archaea (Barrangou et al., 2007). Clustered regularly interspaced palindromic repeats (CRISPPRs) consist of an array of repeats and unique spacers that are acquired from segments of foreign viral or plasmid DNA (Bolotin et al., 2005). The CRISPR array is accompanied by a diverse set of Cas proteins that participate in this immunity response which culminates in the destruction of foreign DNA. The CRISPR-Cas response consists of three stages: adaptation, expression and interference (van der Oost et al., 2014).

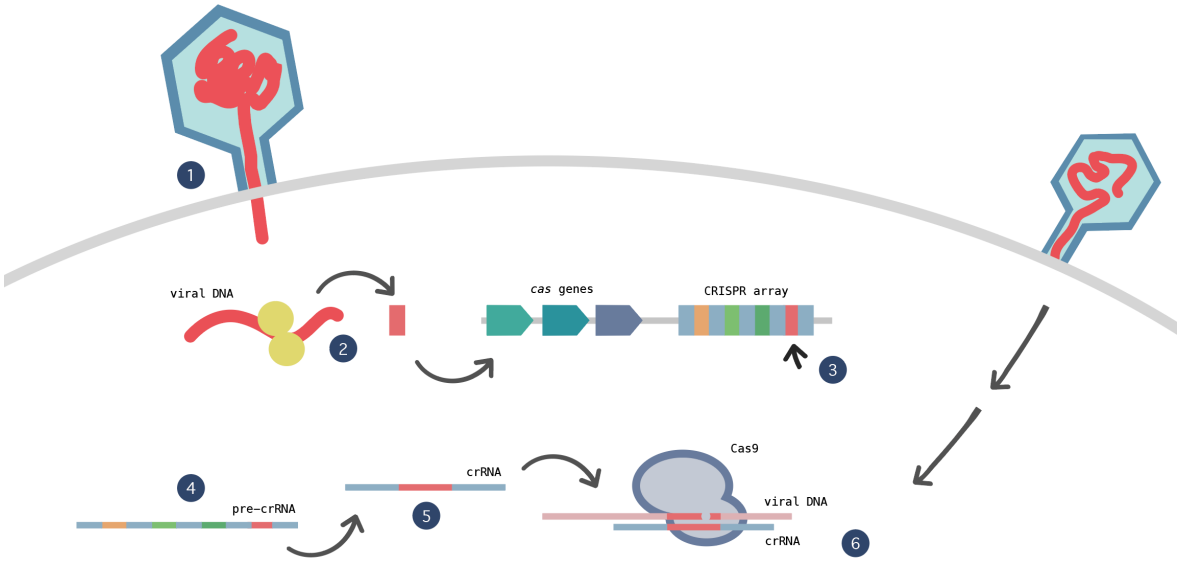


Figure 4 - Overview of endogenous CRISPR-Cas system. CRISPR array is accompanied by *cas* genes that participate in three stages of CRISPR response: acquisition, expression and interference. In the acquisition stage the invading (viral) DNA (1) is recognized by a complex of Cas proteins which then excises a DNA fragment (2). The DNA fragment is inserted as a new protospacer sequence into the CRISPR array (3) between two sets of repeats. In the expression stage CRISPR array is transcribed to pre-crRNA (4), which gets processed into crRNA that contains only one spacer sequence (5). In the interference stage an effector complex consisting of processed crRNA and Cas proteins (Cas9 in Class II CRISPR systems) scans DNA and specifically degrades invading genetic elements that contain a sequence match for the initial DNA fragment (6).

During the adaptation stage, the Cas1-Cas2 complex excises a segment of foreign DNA (known as protospacer) (Yosef et al., 2012). The CRISPR–Cas system selects only protospacers by the detection of a specific protospacer adjacent motif (PAM). The excised DNA fragment is then processed to a specific length and incorporated into the leader-end of the CRISPR array, but without the presence of PAM (Díez-Villaseñor et al., 2013; Swarts et al., 2012). Such protospacers provide the sequence memory for a targeted defense against viruses or plasmids that contain a matching protospacer. In the expression stage, the CRISPR array is transcribed into precursor CRISPR RNA (pre-crRNA) which is subsequently processed by Cas proteins into mature crRNA. In the interference stage, CRISPR RNA (crRNA) guides the effector

module (Cas9 in Class II CRISPR systems) through homologous base pairing with the target DNA. After the hybridization with complementary DNA of the invading genetic element the effector module degrades the DNA through cleavage (Barrangou et al., 2007; Marraffini, 2015).

For a successful target sequence recognition and specific cleavage of foreign genetic elements a short sequence known as a protospacer adjacent motif (PAM) needs to be present just outside the recognition sequence (Deveau et al., 2008). This step is crucial for successful target recognition and cleavage. Even in a case when a target sequence perfectly matches to crRNA sequence, which guides the effector module, the hybridisation can not happen if the proper PAM is missing at the target site. The presence of PAM at the target site is thus essential for CRISPR-Cas system to efficiently differentiate 'self' from 'non-self' sequences to avoid autoimmune activity, since the spacer sequences in the CRISPR array do not contain PAM (Mojica, 2009; Sternberg et al., 2015).

Although this immune response mechanism is consistent across different species of Bacteria and Archaea, operons encoding Cas proteins are extraordinarily diverse in their architecture. Considering the conservation of Cas proteins involved in immune response, we can distinguish conserved ones in the spacer acquisition step (Cas1, Cas2) from the different effector proteins (RNA-guided DNA endonucleases) that are facilitating the degradation of target sequences (Cas9, Cas12a). A two-step classification was proposed that separates distinct gene architectures of CRISPR-Cas operons into two main classes and several types and subtypes (Makarova et al., 2015, 2020; Makarova and Koonin, 2015; S. Shmakov et al., 2017).

CRISPR-Cas systems that use multiple smaller subunits in a complex as effector modules in the interference stage are classified as Class I that are divided into types

I, III and IV, with signature genes *cas3*, *cas10* and *dinG* (or *recD* family helicase in subtype IV-D), respectively (Figure 6). Class II is divided into types II, V and VI with signature effector proteins Cas9, Cas12 and Cas13, respectively (Makarova et al., 2020; Shmakov et al., 2015; S. Shmakov et al., 2017).

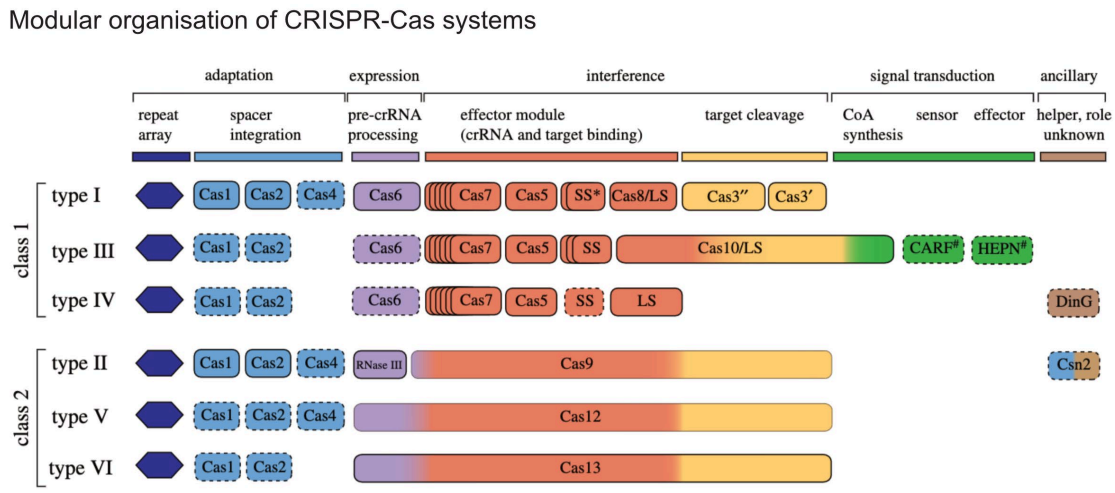


Figure 5 - The principal building blocks of CRISPR-Cas system types. An asterisk indicates the putative small subunit (SS) that might be fused to the large subunit in several type I subtypes. The # next to the CARF and HEPN domain labels indicates that other unknown sensor and effector domains can be involved in the signalling pathway. Dispensable genes are indicated by a dashed outline. Figure adapted from © 2019 Koonin and Makarova. Published by the Royal Society under the terms of the Creative Commons Attribution License

Class II CRISPR-Cas systems were selected to be repurposed for genome editing due to their effector being a single molecule that is able to perform specific DNA binding and cleavage. This molecular architecture makes this system much more portable and desirable for genome editing in various organisms as it can be packaged into two-component system. Particularly effector proteins Cas12a and Cas9, have proven their value as the tool of choice for genome editing (Cho et al., 2013; Hsu et al., 2013; Hwang et al., 2013; Mali et al., 2013; Zetsche et al., 2015). Since guide RNAs

(gRNAs) can be easily reprogrammed to hybridize to any DNA sequence by altering crRNA component and fusing it to trans-activating crRNA (tracrRNA), the PAM sequences becomes the major constrain that limits the available sequence space for genome editing applications (Siksnys and Gasiunas, 2016).

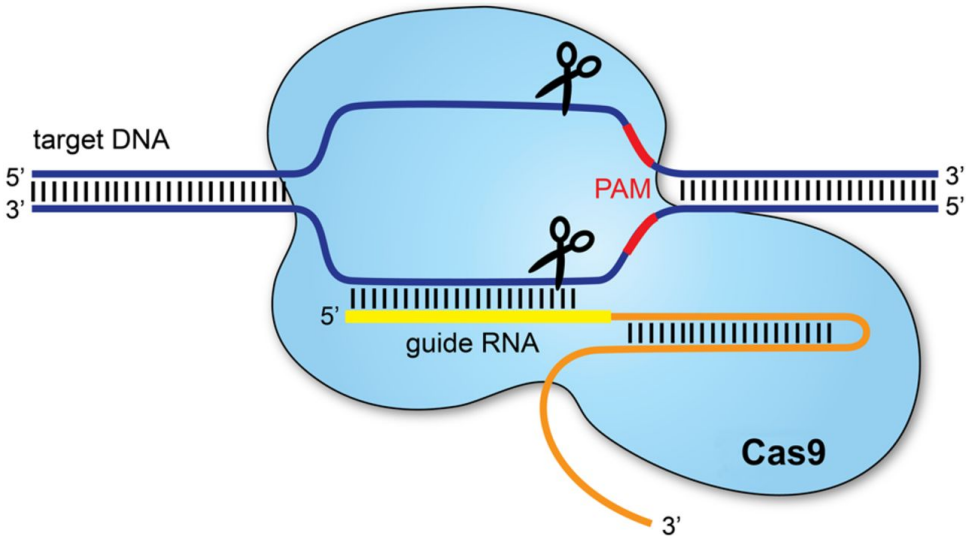


Figure 6 - The repurposed CRISPR-Cas9 system for genome editing. Cas9 endonuclease is directed to its target by a guide RNA (gRNA) synthetized based on crRNA and tracrRNA to recognise an 18-20 bp long sequence. In order to cut, a specific sequence called protospacer adjacent motif (PAM) must lie at the 3' end of the guide RNA. Image reproduced with permission of the rights holder, Redman et al. 2016, © 2016, BMJ Publishing Group Ltd.

The C-terminal end of Cas9 was shown to interact with PAM at the target site, hence named PAM interacting (PI) domain (Anders et al., 2014). The amino acid residues in PI domain were shown to be the main determinant of different PAM specificities of different Cas9 orthologs (Nishimasu et al., 2014). This finding opened the possibility to engineer Cas9 to achieve different PAM specificities. PAM specificity of *S. pyogenes* Cas9 (Kleinstiver et al., 2015b) and *S. aureus* Cas9 (Kleinstiver et al., 2015a) were improved by altering the amino acid sequence of the PI domain, demonstrating the feasibility of engineering Cas9 nucleases with useful new properties for the purpose of genome editing. By extending the PAM sequence,

recognized by a modified SpCas9, from canonical 5'-NGG-3' PAM to 5'-NGAG-3' or 5'-NGCG-3', a higher targeting specificity was achieved. Several other successful attempts to change or improve specificity of different Cas9 orthologs followed (Casini et al., 2018; Chen et al., 2017; Hirano et al., 2016; Hu et al., 2018; Miller et al., 2020; Slaymaker et al., 2016). An SpCas9 variant called SpRY was engineered to exhibit very low PAM recognition specificity. SpRY variant recongises 5'-NRN-3' PAM, which means that recognition is limited only by either A or G in the second position in the PAM. For that reason the variant was termed to be a near PAMless Cas9 (Walton et al., 2020). The SpRY variant represents the greatest progress in Cas9 engineering to achieve the PAM-free variant and allowing boundless sequence recognition in genome editing (Collias and Beisel, 2021).

In addition to engineering existing Cas9 proteins to demonstrate different PAM specificities, exploring the wide range of Cas9 orthologs could identify new PAM preferences. Tapping into the natural diversity of CRISPR-Cas systems in Bacteria to explore more Cas9 orthologs might provide insights into the nuances of varying sensitivity, specificity, kinetic requirements and propensity to generate off-target cleavage in the context of the engineering application and, perhaps most importantly, further widen the range of potential sequences that can be targeted (Braff et al., 2016).

Several different approaches used to address the novel PAM elucidation were developed and can be divided into three categories: *in silico, in vitro* and *in vivo*. *In silico* approaches for decoding PAM for a Cas9 protein use spacer sequences stored in CRISPR array to find hits in viral and plasmid databases. By analysing sequences flanking immediately upstream or downstream of the target in invading DNA it is possible to identify the consensus PAM (Mendoza and Trinh, 2018). The principle of negative selection is implemented in plasmid depletion experiments in *in vivo*

methods. Cells bearing plasmids encoding the Cas9 and protospacer-targeting guide RNA are co-transformed with a plasmid library containing an antibiotic resistance gene, and a protospacer sequence flanked by a randomized PAM sequence. Plasmids containing functional PAMs are cleaved by the Cas9 leading to cell death. Deep-sequencing of the Cas9 cleavage resistant plasmid pool isolated from the surviving cells displays a set of depleted plasmids that contain functional cleavage-permissive PAMs (Karvelis et al., 2017). *In vitro* methods are also based on similar depletion principles, focusing on the cleavage of PAM libraries with selected Cas9 endonuclease. Only plasmids with target sequences and an adjacent, cleavage-permissive PAM that is recognized will be cleaved and depleted from the starting library. The remaining non-cleaved sequences can be amplified and sequenced and compared to the starting library to determine which functional PAM sequences were depleted during cleavage assay. (Marshall et al., 2018)

By using a combination of these approaches, additional biochemically diverse CRISPR-Cas9 orthologs were identified and their PAM specificities characterized using a cell-free *in vitro* assay. The 79 phylogenetically-distinct Cas9 orthologs identified in this way substantially expand the genome editing toolbox (Gasiunas et al., 2020).

1.8 Challenges and limitations

The first CRISPR-based gene drive in *An. gambiae* targeted three female-fertility genes, namely *AGAP005958*, *AGAP011377* and *AGAP007280*. The three selected genes were identified as putative candidates that would confer a recessive female-sterility phenotype upon disruption, hence being suitable target genes for population suppression as the disruption would reduce the female reproduction capacity. Following the model prediction, the gene drive allele increased in frequency up to 80% transgenic ratio but subsequently failed to increase to fixation or to remain at that level. The gene drive failed to spread due to high selection pressure for the alleles that were resistant to the gene drive target site recognition (Champer et al., 2017; Hammond et al., 2017).

1.8.1 Gene drive resistance

Observance of resistance to homing endonuclease gene drives was not a phenomenon that was not expected. Expectation of resistance and the ways to tackle the problem was discussed already at the beginning, when homing endonuclease gene drives were first proposed (Burt, 2003). The suppressive nature of gene drives is likely to select for resistance, therefore the mitigation strategies play an essential role. There are many potential mechanisms of resistance, however the most likely for resistance to occur is at the target site. Source of resistance there could be standing variation in the population, which is unlikely in a laboratory setting due to limited genetic diversity, or it could be nuclease-generated.

It was shown that, following the chromosome cleavage, HDR is not always the only repair mechanism that cells employ. The two alternative repair pathways to HDR are non-homologous end joining (NHEJ) and microhomolgy-mediated end joining

(MMEJ) and both are generally considered as error-prone. NHEJ and MMEJ do not use the sister chromatid as the template to repair DSB but ligate the loose ends through two different processes that can introduce insertions, deletions and substitutions in the nucleotide sequence at the target site (Lieber, 2010; Moynahan and Jasin, 2010; Wang and Xu, 2017).

Mutations at the target site that restore some function to the gene targeted by the gene drives yet prevent endonuclease activity are expected to be strongly selected in the face of the gene drive. Such mutations that restore the gene function were termed restorative resistant alleles. The selection pressure for the restorative resistant alleles (or functional resistant alleles) in the gene drive targeting *AGAP007280* halted the spread of the gene drive allele, preventing the population crash in the cage. Mutations replaced the drive-susceptible allele in the caged population of mosquitoes were generated by MMEJ and NHEJ. The most abundant resistant alleles were functional in-frame 6 bp long deletions (Hammond et al., 2017).

Resistance allele formation was also observed in *Drosophila melanogaster* at high rates both prior to fertilization in the germline and post-fertilization in the embryo due to maternally deposited Cas9 (Carrami et al., 2018; Champer et al., 2018, 2017; Oberhofer et al., 2018). The first gene drive experiments in mice also showed prevalence of drive-resistant alleles (Grunwald et al., 2019).

1.8.2 Error-prone repair of double-strand breaks: NHEJ and MMEJ

NHEJ is active in all phases of the cell cycle and can induce small sequence alterations at junction or it can be high-fidelity mechanism without producing any

mutations. The repair process is initiated by abundant nuclear Ku heterodimer, consisting of Ku70 and Ku80, which binds to the exposed DNA ends in sequence-independent manner (Figure 4). Ku heterodimer binding is essential for facilitation of DSB repair through NHEJ as it recruits other core factors needed to repair the DSB (Downs and Jackson, 2004). In addition to the recognition and recruitment role Ku also performs 5′-RP/AP lyase activity, playing a catalytic role in the repair mechanism by modifying damaged bases proximal to the exposed DNA ends (Weinfeld et al., 2011). Ku recruits DNA-PKc (DNA-dependent protein kinase) that has 3'-DNA phosphatase and 5'-DNA kinase activities, which allow it to remove non-ligatable DNA-end groups at DSB termini (Weinfeld et al., 2011). Ku70 and Ku80 orthologs were found in *Drosophila* and in mosquitos, however DNA-PKc was lost in *Drosophila* and other species from Schizophora, Coleoptera, and some other insects, though it is still found in mosquitoes (Sekelsky, 2017). Other proteins were suggested to also take part in the processing of the ends, namely: endonuclease Artemis, the X family DNA polymerases Polλ, Polμ, or terminal deoxynucleotidyl transferase (TdT). However, most insects lack Polλ, Polμ, and TdT orthologs. Despite strong evidence for end processing in NHEJ in *Drosophila*, there is no evidence what type of nucleases and polymerases perform the processing (Bozas et al., 2009). After the end processing the exposed ends are ligated together by a complex that consists of ligase 4 (Lig4), XRCC4 and XLF (Radhakrishnan et al., 2014; Williams et al., 2014). It was shown that the mutation of Lig4 in *Drosophila* resulted in imperfect NHEJ with large deletions, reduced progeny and higher frequency of HDR (Wei and Rong, 2007). XLF orthologs differing in C-terminus were found in *D. melanogaster* species complex, suggesting a tandem duplication of the gene (Sekelsky, 2017). Despite missing XRCC4 in *An. aegypti*, NHEJ is functional and was observed in the generation of genetically modified mosquitoes, which suggests an alternative pathway of NHEJ

compared to *D. melanogaster* (Aryan et al., 2013; Chang et al., 2017; Mota et al., 2019).

Alternatively, DSB can be repaired by MMEJ, a genetically disctict repair mechanism that is a Ku- and Lig4-independent when compared to the factors involved in NHEJ. It is also independent of RAD51 and BRCA2, genes that are linked to HDR (Wei and Rong, 2007). Similarly to NHEJ, MMEJ too ligates the exposed DNA end together without using the homologous chromosome (Figure 4). However, MMEJ is always associated with deletions flanking the break site due to its mechanism that utilises michrohomologous sequences internal to the broken strand ends, a step which also defines MMEJ repair. MMEJ process can be divided into three steps: three general steps: pre-annealing, annealing, and post-annealing of the flanking microhomologies (MHs). The process starts with end resection, which exposes the flanking MHs. The process of resection is common to HDR in which the MRN complex together with CtIP creates a nick in the proximity of DSB and degrades the DNA strand using 3'-5' exonuclease. Larger 5'-3' resection is generated by Bloom syndrome RecQ-like helicase (BLM)/exonuclease 1 (EXO1) (Takeda et al., 2007; Truong et al., 2013). Both ends of DNA proximal to DSB have ssDNA overhangs exposed after resection. MH annealing could be repressed by the presence of replication protein A (RPA) complex that binds to the ssDNA overhangs and promotes HDR instead of MMEJ, pointing to its important function in the DSB repair and regulation of MMEJ (Deng et al., 2014). In the MMEJ process, to establish the DNA synapse between the exposed ssDNA overhangs, Polθ plays a crucial role. Polθ also promotes and regulates annealing of 2-20 bp long MH and catalyses the overhang extension (Kent et al., 2015). Even in the early studies Polθ was shown to promotes I-sceI-induced MMEJ in *Drosophila* (Chan et al., 2010). MMEJ-specific role of Polθ was further highlighted by evidence that MMEJ is fully dependent on its presence in human cells (Kent et al., 2015). The ligation step in higher eukaryotic

cells is performed by Ligase I (Lig1), Ligase III (Lig3) and Polθ (Kent et al., 2015; Oh et al., 2014). Notably, Lig3 has been lost in mosquitoes, suggesting that Lig1 and Polθ may be critical for MMEJ in these species (Overcash et al., 2015). Non-homologous tails that remain after MH annealing and ligation are removed by XPF/ERCC1 nuclease. This allows DNA ends to undergo priming repair synthesis and triggering the deletion of the sequence in between MHs (Ahmad et al., 2008).

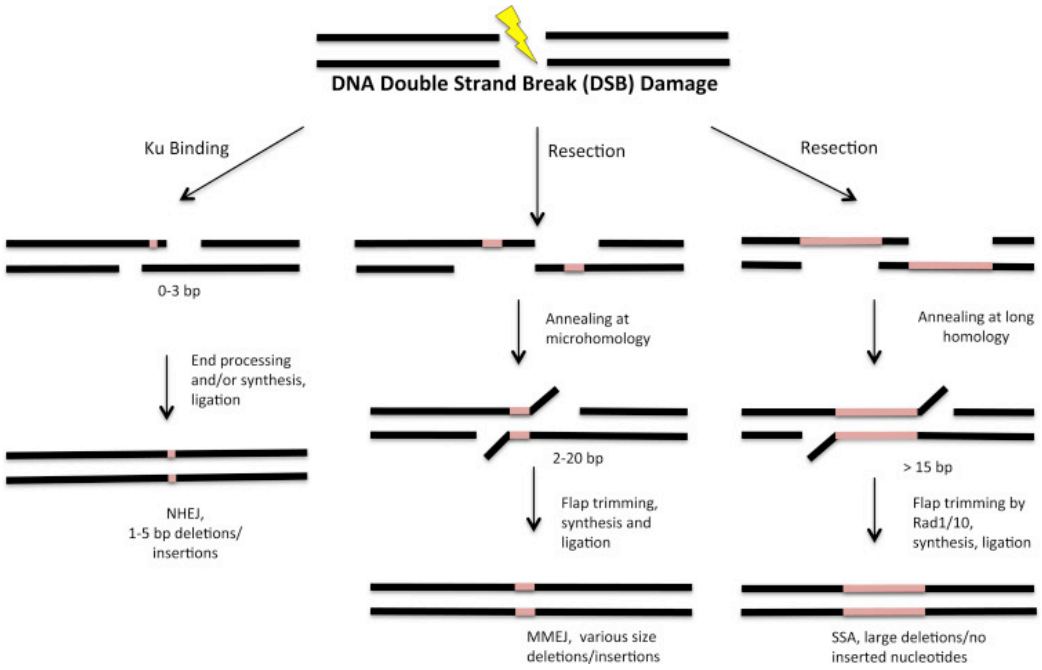


Figure 7 - MMEJ and NHEJ repair mechanisms. DSB can be repaired by NHEJ pathway, where Ku proteins bind to each end of the broken DNA molecule and recruit enzymes that process ends and ligate the broken DNA together. The NHEJ pathway results in short deletions and insertions at the DBS site. In the MMEJ pathway both ends of DNA anneal together through the short microhomology patterns. Flaps that are generated after the annealing get trimmed and the gaps are filled and ends get ligated. Result of the MMEJ pathway is variable-size deletion. Reprinted from Risky business: Microhomology-mediated end joining, 24/11, Sinha Supriya et al. © 2016 with permission from Elsevier

Different pathways of DSB repair are more prevalent in certain contexts than others due to their outcomes. HDR is favoured to be more prevalent after DNA replication

when the cell contains identical sister chromatids required for HDR. NHEJ is active in all phases of the cell cycle but is favoured in G1 cells (Chan et al., 2011). Experiments in *Drosophila* and *Caenorhabditis elegans* also showed that HDR is favoured in the germline compared to NHEJ, which was more abundant in somatic cells (O'Brochta et al., 1991; Gloor et al., 2000; Clejan et al., 2006).

1.8.3 Regulation of cutting activity

Regulation of cutting activity through endonuclease expression was also shown to play a major role in generating problematic end-joining mutations in current gene drive designs (Champer et al., 2018; Gantz et al., 2015; Hammond et al., 2017). High levels of end-joining repair were observed following strong maternal deposition and cleavage of the nuclease in the embryo. The undesired cleavage of the endonuclease was attributed to a *vas2* promoter controlling the germline expression that was active through oocyte development and deposited in the embryo as endonuclease mRNA and/or protein. The same promoter caused a strong unintended fitness cost in heterozygous females. The nuclease expression in somatic cells caused unintended homing, partially converting gene drive allele to homozygosity, hence disrupting function of the somatically required gene. Furthermore, somatic leakiness contributed to high rate of mutations in the embryo. This reduced the fecundity of females and reduced bias in gene drive inheritance, as well as increased selection pressure for resistant mutations. Selection of different germline promoter, called *zpg*, improved the speed of invasion, increased suppressive effect and reduced the resistant mutations at the target site, mainly due to less propensity to induce end-joining repair in the germline and due to improved fecundity of females that contain a single copy of the gene drive and transmit it in a super-Mendelian fashion to their offspring (Hammond et al., 2021a).

1.8.4 Overcoming resistance

Different strategies were suggested to either reduce the impact of resistant alleles or to prevent the resistant alleles to arise in the first place. A combined release that targets more independent loci rather than just one with gene drive in a population would help to reduce the impact of resistant alleles (Burt, 2003; Marshall et al., 2017). Alternatively to the combined release, a single target locus could be used but targeted with multiple gRNAs. In this case resistance would need to evolve at all sites to successfully prevent subsequent drive conversion. In the opposite case if resistance at only some of the target sites allowed cleavage then the process would result in successful drive insertion by HDR and the resistant allele would be removed from the population. This approach was called multiplexing and it was successfully used to reduce the resistance rates in *Drosophila melanogaster* (Champer et al., 2018).

Another important mitigation strategy against resistance is rational target site selection. To reduce the likelihood of resistant restorative alleles arising, a target site could be chosen for which microhomology on both sites of the cut site would ensure that the most likely end-joining repair event would be an MMEJ-derived out-of-frame deletion. A non-restorative allele created in this manner would be less likely to be selected for (Hammond et al., 2017; Hammond and Galizi, 2017; Nolan, 2021).

Furthermore, selecting a conserved target site indicated by low genetic variation in wild *An. gambiae* populations and across multiple *Anopheles* species would imply functional constraint (Miles et al., 2017; Neafsey et al., 2015; The Anopheles gambiae 1000 Genomes Consortium, 2020). Resistant mutations at such a site are less likely to restore function of the gene.

With this in mind, a gene drive was designed to target the gene *doublesex* which is involved in the sex determination pathway in *An. gambiae*. *Doublesex* (*dsx* or *AGAP004050*) acts as the final binary switch between male and female sex development and was selected as a gene drive target site by Kyrou et al. (2018). It produces two alternatively spliced isoforms, male-specific *dsxM* (*AGAP004050-RA*) and female-specific *dsxF* (*AGAP004050-RB*), which in turn encode transcription factors that control sexual dimorphism between the sexes. On the 5'-end the isoforms share CDS that spans over exon 1, 2, 3 and 4 and codes for two DNA binding domains, namely OD1 and OD2. The female-specific segment corresponds to exon 5, and the male-specific part is encoded in exon 6 but transcribed as 3'-UTR in females. Both male and female isoforms also share exon 7 as 3'-UTR (Scali et al., 2005). The gene drive target site was selected in a highly conserved and female-specific exon 5. Exon 5 is alternatively spliced and present only in the female transcript, making the target site female specific. The *dsx* target site has been shown to be very conserved across *An. gambiae* wild populations and across different *Anopheles* species, suggesting tight functional constraint at this region of the *dsx* gene. Females with homozygous *dsxF* mutation showed abnormal intersex phenotype manifested in male and female morphological features. Importantly, females homozygous for the disrupted allele were completely sterile, while males with heterozygous or homozygous mutation were not affected. This allowed the spread of the gene drive allele through a caged population and consequently caused this population to collapse. Contrary to previous gene drive experiments, in which resistant alleles prevented population suppression, a study of *dsx* target site mutations revealed a very low frequency of indels and no restorative mutations that could be selected for and cause resistance to the gene drive (Kyrou et al., 2018).

Recently, a system based on *dsx* gene drive coupled with sex distorter gene was applied in laboratory settings and successfully crashed a caged mosquito population

by gene drive in *dsx* locus and creating a strong male bias in the progeny (Simoni et al., 2020). Following the successful laboratory experiments, the suppressing activity of *dsx* gene drive was tested in large indoor cages, which more closely mimicked conditions and mosquito population in the wild. The gene drive was able to spread rapidly from a single release at low frequency and achieved full population suppression within a year's time, without being stopped by gene drive resistance (Hammond et al., 2021b).

2 Aims

My research project aimed to identify and characterise novel CRISPR-Cas endonucleases that possess different PAM specificities for extending the targeting space in genome engineering. Furthermore I aimed to analyse and predict the consequences of their activity and find new targets for their use in genome editing and gene drive applications in the malaria vector *Anopheles gambiae*.

The aims were addressed in two main chapters of the thesis.

Chapter 3 describes a computational approach to identify CRISPR-Cas9 orthologs in bacterial genomes. The PAM specificities of identified Cas9 orthologs were further elucidated using *in silico* methods. Additionally, the targeting space in the *Anopheles gambiae* genome was analysed for three selected Cas9 orthologs.

Chapter 4 addresses the consequences of genome editing activity in the context of the *Anopheles gambiae* genome and how they affect the gene drive target site selection. A computational model was implemented to predict outcomes of MMEJ and compared to observed end-joining mutations *in vivo*. Furthermore, highly conserved genomic regions were identified in the genome as a proxy for genomic regions under functional constraint. These findings were integrated into a computational tool that is capable of informative selection of gene drive target sites. The resulting computational tool was used in search of suitable alternative gene drive target sites in four candidate genes.

3 Exploring alternative CRISPR-Cas nucleases for genome editing and their specificities

3.1 Identification of PAM preferences of CRISPR-Cas nucleases

3.1.1 PAM recognition pipeline

The proposed *in silico* PAM recognition pipeline (Figure 8) consists of the following steps:

1. CRISPR array detection in bacterial genomes
2. Extraction of spacer sequences from the CRISPR array
3. Search for homologous sequences in the database of viral sequences
4. Extraction of the flanking sequences from the spacer hits in viral nucleotide database
5. Stacking of the flanking sequences to identify conserved nucleotide positions that indicate PAM

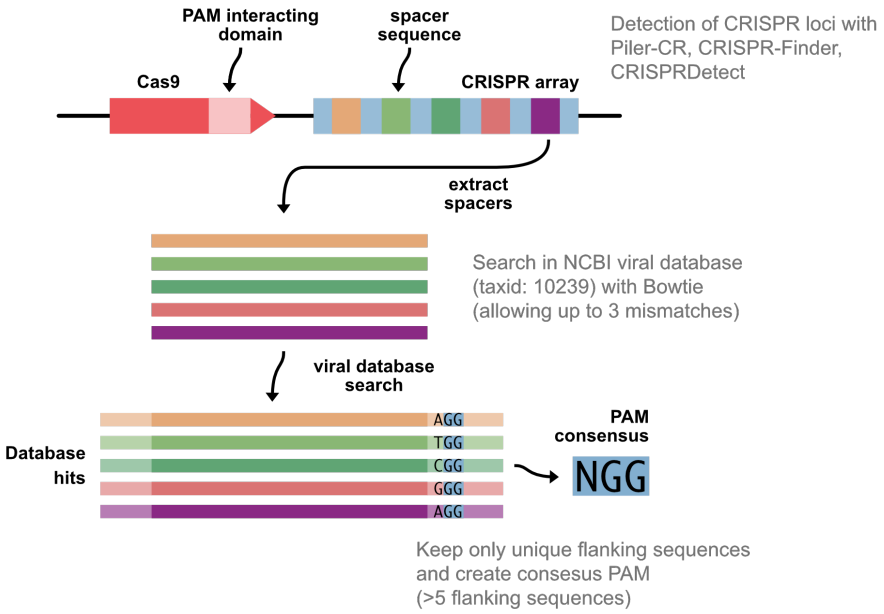


Figure 8 - The PAM recognition pipeline. Spacers are extracted from CRISPR array and used as a query sequence for search in viral database. Flanking sequences are extracted from target sites where nucleotide conservation indicates PAM sequence.

Similar *in silico* PAM elucidation pipelines have been proposed, all of them in principle reversing the steps of spacer adaptation in the CRISPR locus, first obtaining the stored viral sequence in CRISPR array, and then tracking its origin in the viral database. Despite the similarities, there are some notable differences that separate the methodologies between each other.

For the Step 1 two approaches were used to obtain the starting dataset of annotated CRISPR loci and spacers. The first approach was to obtain bacterial genomes with annotated *cas9* genes and then conduct the CRISPR array detection and annotation in them by using CRISPRdetect (Biswas et al., 2016). The second approach was to use already annotated CRISPR loci and spacer sequences that were gathered from the whole NCBI nucleotide database, in contrast to the first approach that used only a subset of genomes to conduct the search (S. A. Shmakov et al., 2017; Shmakov et al., 2018).

The search for homologous sequences in Step 3 can be performed using different search algorithms in different sets of databases. The aim of this step is, firstly, to obtain a homologous match for each extracted 30-35 bp long spacer, and secondly, minimise the rate of false positive hits. Such false positive hits could be, for example, viral sequences that are matching the spacer, but are not the real source of the spacer and therefore do not have the PAM in their flank sequence. Increasing the precision in the search step decreases the number of noisy flanking sequences extracted in the Step 4.

Upon obtaining the flanking sequences for a particular Cas9 ortholog the PAM needs to be discerned. The conserved nucleotide indicating PAM in the flanking regions can be visually identified by using Weblogo (Crooks et al., 2004). It has proven to be

a useful method for visualizing PAM sequences in the past (Horvath et al., 2008; Karvelis et al., 2017; Mojica, 2009; Shah et al., 2013).

A common measure of sequence conservation is Shannon entropy (H), which defines uncertanty or, equivalently, information on each position in the sequence alignment. It is defined as: $H = -\sum_{n=1}^{N} p_n \, log_2 \, p_n$; where p_n stands for the observed frequency of the nucleotide n and N represents the 4 possible nucleotides. The measure is expressed in bits.

This translates into sequence conservation through observed nucleotides on each position. For example, if there are 10 sequences in the alignment and all 10 sequences have G at the first position, the resulting Shannon entropy of that position (or uncertanty) equals zero.

When analysed sequence is DNA only 4 distinct nucleotides are expected to occur for each position, effectively setting the maximum Shannon entropy (or conservation) for each position to $log_2 \, 4 = 2$, also known as 2 bits (Crooks et al., 2004).

The conservation in the logo is indicated by the height of each column/position, whilst the height of a nucleotide within the stack represents the relative frequency of that nucleotide on that position. The sequence conservation on each position (the height of the column) is defined as difference between the maximum possible Shannon entropy and the Shannon entropy of observed nucleotide distribution (eq 3.1).

$$R_i = H_{max} - H_{obs} \qquad 3.1$$

$$R_i = log_2 N - \left(- \sum_{n=1}^{N} p_n \, log_2 \, p_n \right) \qquad 3.2$$

In eq. 3.2 p_n stands for the observed frequency of the nucleotide n at the position i in the sequence and the N is the number of distinct nucleotides that are expected to be observed.

For example, following the eq. 3.2, R_i is going to be zero, meaning no conservation, if we observe all 4 nucleotides in equal frequency (0.25) at a given position i. In contrast, in case of maximum conservation we would observe only one nucleotide at the position i in all sequences (p_n=1), which would mean that R_i is going to be 1.

The authors of CASPERpam, a tool developed for PAM identification, suggest defining the significant positions that contribute to PAM as positions in the flanking sequence that have R_i higher than a half of standard deviation above the mean of all the positions. As additional constraint, the R_i values for positions further than five nucleotides away from the spacer are required to be full standard deviation above the mean to be deemed significant. This constraint was set to further filter out noise from positions unlikely to contribute to the PAM sequence.

The next step after determining the significant positions that constitute PAM is to ascertain the consensus nucleotides on these sites based on their relative frequency. If the relative frequency of a nucleotide is greater than 0.5 it is considered to be the consensus nucleotide on that position. If no nucleotide was higher than 0.5, the nucleotides that had relative frequency greater than 0.25 were considered to be part of the PAM (Mendoza and Trinh, 2018).

For the comparison between obtained PAMs two other metrics were used in my work, both deriving from Shannon entropy calculated for each positions. Mendoza

and Trinh (2018) proposed eq. 3.3 to be used for describing the average deviation in entropy (R_{avgdev}) on the significant PAM positions.

$$R_{avgdev} = \frac{\sum H_{i,sig} - H_{avg}}{n_{sig}} \quad 3.3$$

In eq. 3.3, $H_{i,sig}$ denotes Shannon entropy of a significant position; H_{avg} stands for the average Shannon entropy across all positions in the flanking sequence and n_{sig} is the number of significant positions in the flanking sequence.

Alternatively, I calculated Jensen-Shannon divergence (JSD) and its square root ($\sqrt{JSD}$) as Jensen-Shannon distance was used as a metric to describe convergence in total variation between two probability distributions. It quantifies how "distinguishable" two or more distributions are from each other. JSD is defined as following:

$$JSD(P||Q) = H\left(\frac{P+Q}{2}\right) - \frac{H(P)+H(Q)}{2} \quad 3.4$$

In eq. 3.4 P is the distribution of nucleotide frequency of a significant position and Q is the distribution of nucleotide frequency of non-significant positions in the flanking sequence. (Fuglede and Topsoe, 2004) I used both $\sqrt{JSD}$ and R_{avgdev} to quantitatively compare PAMs between each other and to quantify the contribution of positions in the flanking regions to PAM.

For example, $\sqrt{JSD}$ and R_{avgdev} values for a correctly predicted and well conserved *Streptococcus pyogenes* Cas9 PAM were 0.73 and 1.45, respectively, while $\sqrt{JSD}$ and R_{avgdev} values under 0.25 were never associated to any correctly predicted PAMs.

3.1.2 PAM identification of uncharacterized Cas9 orthologs

A computational pipeline was constructed in my study to elucidate PAM sequences for previously uncharacterized Cas9 orthologs. Spacers were extracted from CRISPR arrays found given bacterial species and used as a query sequence for search in viral database. Flanking sequences were extracted from target sites and analysed for nucleotide conservation that would indicate PAM sequence. Only PAMs supported by 5 sequences or more were visualized to ensure better resolution. Obtained sequence logos were evaluated with previously described metrics, as well as qualitatively in order to discern PAMs for each taxonomic group. In many cases, especially for taxa with fewer stacked flanking sequences, no clear motif could be detected, very likely due to predominantly false positive hits in the viral database.

The analysis included two datasets – Dataset A consisted only of CRISPR arrays and spacers from type II CRISPR-Cas systems that were found in the bacterial genomes with annotated Cas9 protein. Dataset B derived from two published datasets of spacers and annotated CRISPR loci from all publicly available bacterial sequences (S. A. Shmakov et al. 2017, 2018). Spacer sequences were linked to their originating CRISPR arrays from extensively annotated CRISPR loci (Figure 9). For this analysis only annotated CRISPR loci of CRISPR type II and their spacer sequences were used.

Comparing the two datasets, it was evident that the new dataset (Dataset B) captures CRISPR-Cas loci in more species and provides a larger pool of spacers and consequently target hits that could be used in PAM identification. Dataset A captured 5863 CRISPR arrays in 248 species, while dataset B captured 7282 arrays labelled as Class II in 1442 different species. For most of the detected CRISPR arrays in Dataset A no CRISPR type was determined.

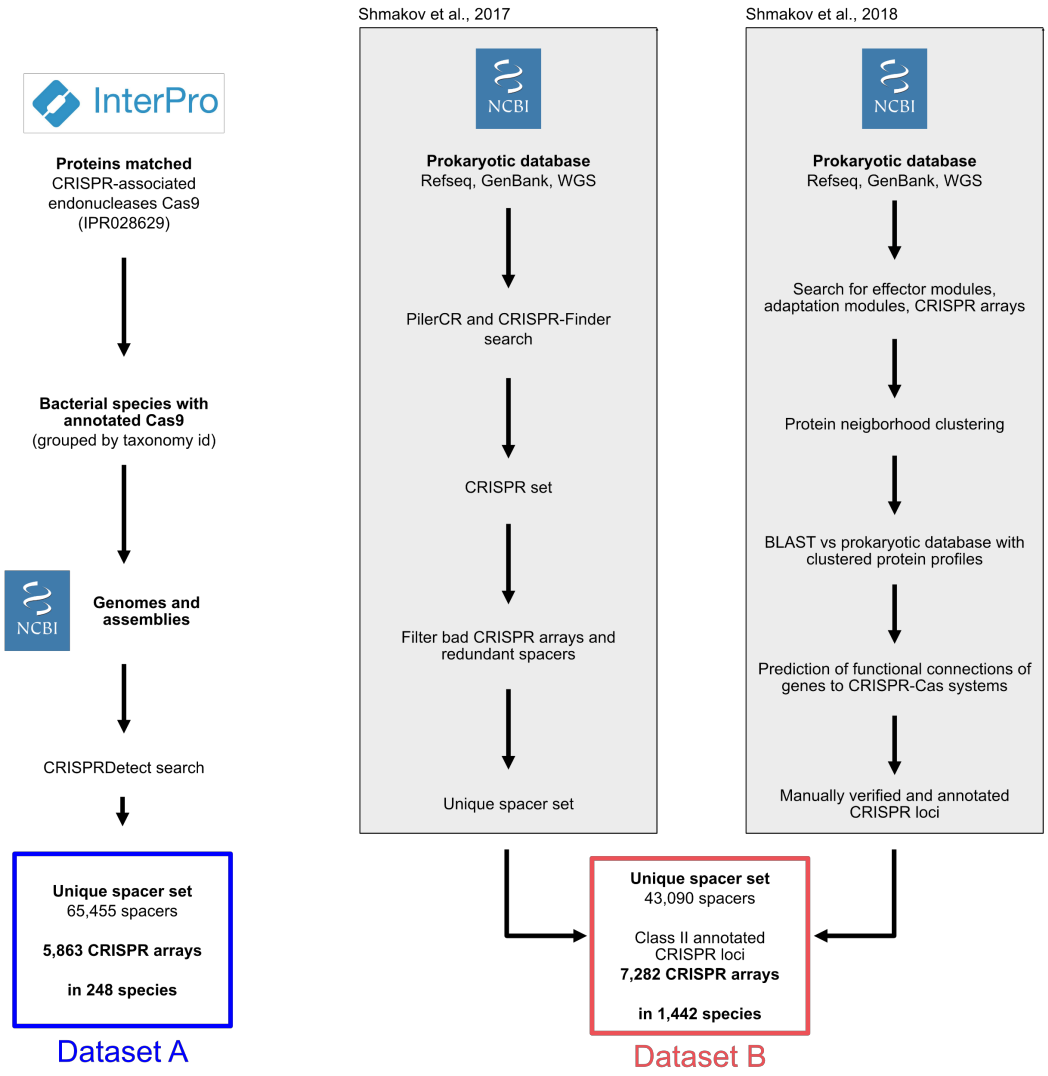


Figure 9 - Comparison in computational pipelines behind Dataset A and Dataset B used in the analysis. Dataset A consists only of CRISPR arrays and spacers from genomes that have Cas9, belonging to CRISPR type II annotated, whilst dataset B consists of annotated CRISPR loci and spacers found in all prokaryotic sequence databases from NCBI (S. A. Shmakov et al. 2017, 2018).

Spacer sequences extracted from CRISPR arrays were used as a query to search the viral database to obtain their viral origin. The database was collated from all sequences that were in NCBI nucleotide databases (nuccore, RefSeq, Genome) and annotated as viral origin (taxid: 10239) at the time of the analysis. The database consisted of over 230,000 viral sequences.

Several differences were observed between datasets A and B when comparing the number of target hits for each species (Figure 10) with a class II CRISPR-Cas system. Significantly more target hits came from spacers in dataset B, including some for bacterial species for which no Cas9 orthologs have been characterised.

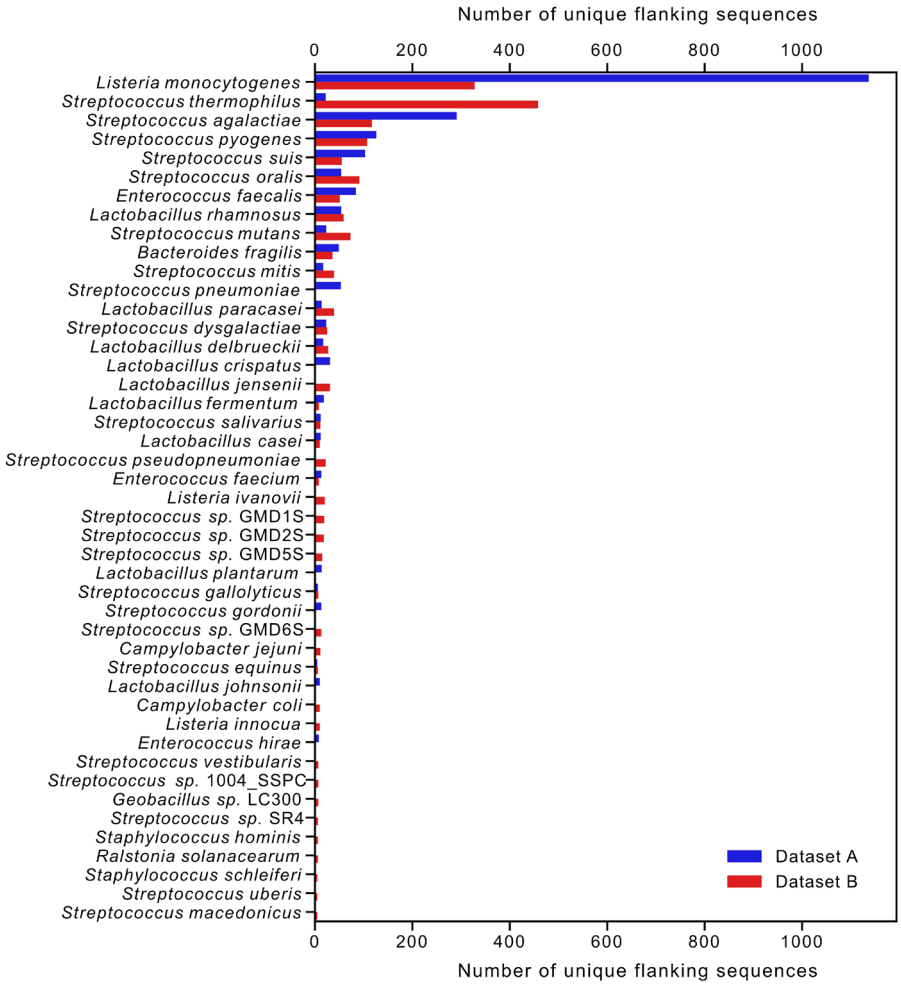


Figure 10 - Comparison between the number of unique target hits deriving from viral database search by using spacer sequences grouped per bacterial species from dataset A and dataset B. The grouping was performed based on NCBI species taxonomy id.

The number of target hits was much lower than the actual number of unique spacers obtained for each species both in Dataset A (Figure 11) and in Dataset B (Figure 12). In dataset A only 26 species had 5 or more unique target hits. On average, only 5% of spacer sequences had at least one unique target hit in the viral databases with up to 3 mismatches per sequence in Dataset A. Potential reasons for this under-representation are addressed later in the discussion (Chapter 3.4). This makes PAM identification for a wider range of Cas9 orthologs very difficult, especially when considering that the conservation in flanking regions of many targets is very low.

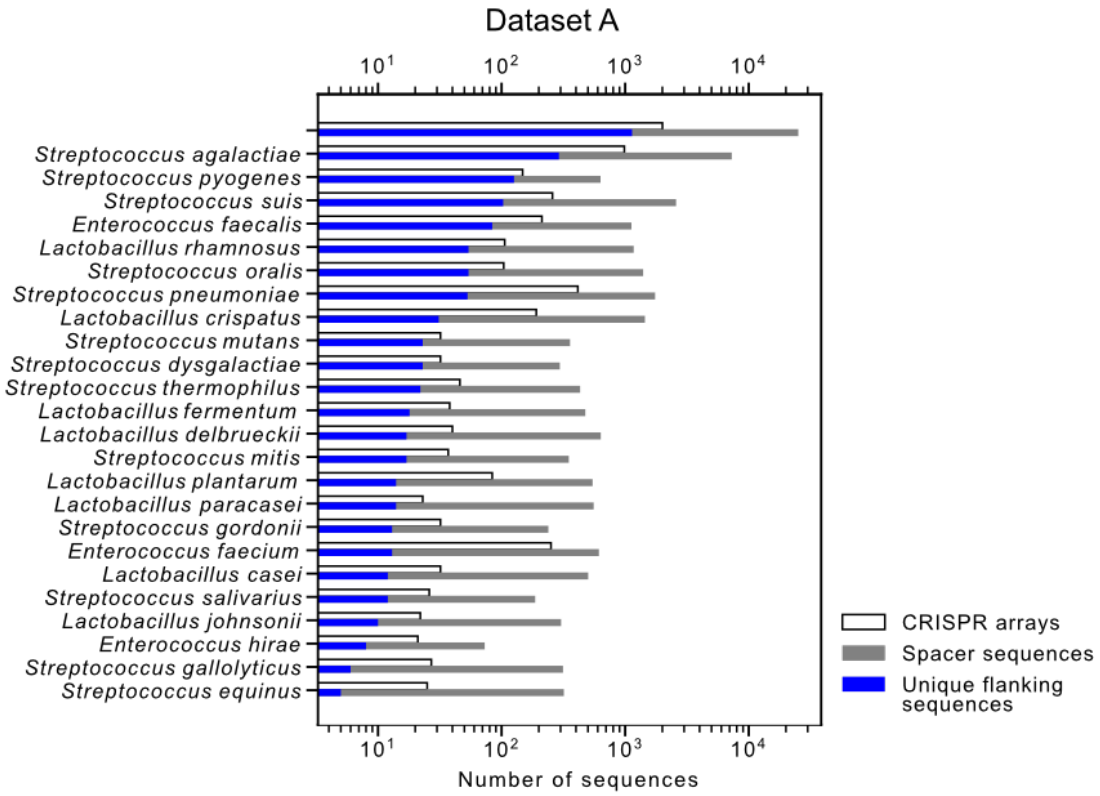


Figure 11 - Comparison between the number of unique target hits in viral database and the number of their query spacer sequences from Dataset A. Unique target hits in viral database are marked with blue and query spacer sequences in grey. Each column represents a bacterial species which spacer sequences were extracted from.

Species with enough target hits in the viral database were predominantly from *Streptococcus* and *Lactobacillus* genus, this means that these had the highest signal, giving the most power to detect PAMs for their Cas9 orthologs using our approach.

For only 39 species in Dataset B enough flanking sequences were obtained to proceed with PAM identification (Figure 12). In those species on average 21% of spacer sequences yielded a unique flanking sequence obtained from a viral target hit with 3 mismatches or less, much higher success rate than in Dataset A.

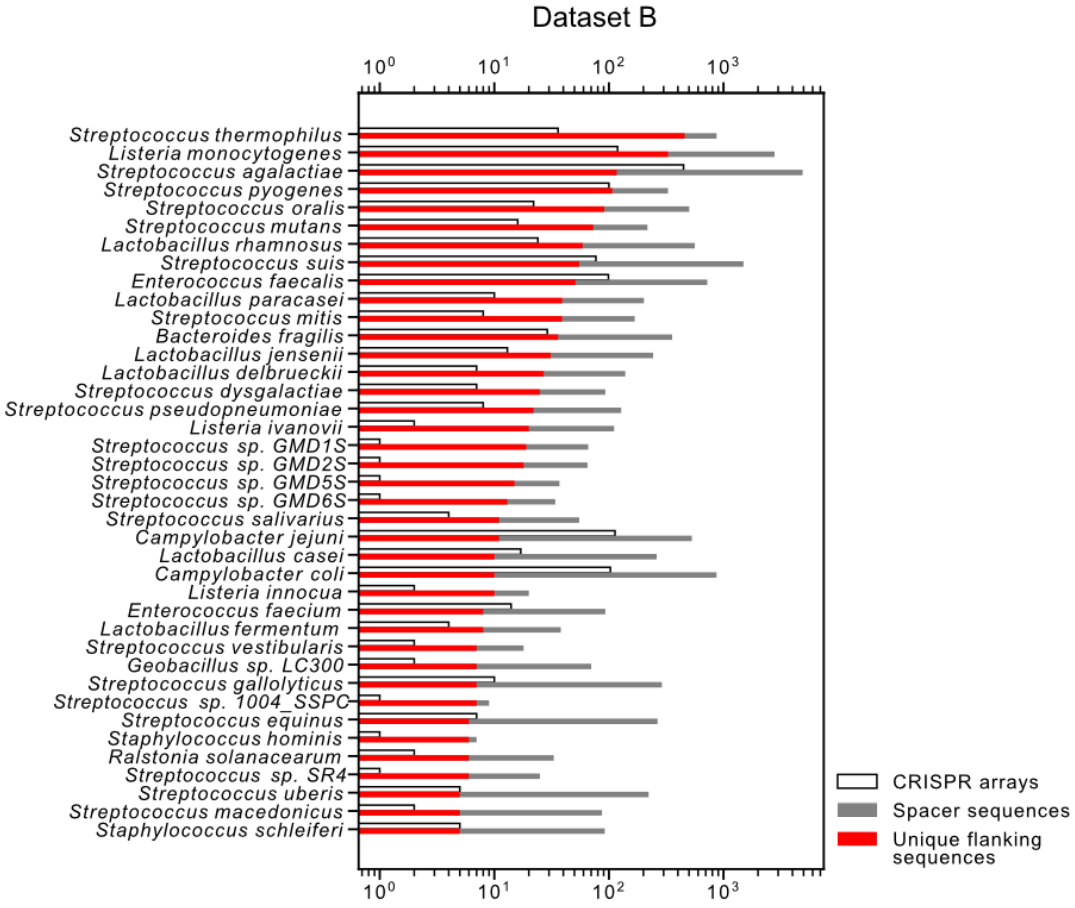


Figure 12 - Comparison between the number of unique target hits in viral database and the number of their query spacer sequences from Dataset B. Unique target hits in viral database are marked with red and query spacer sequences in grey. Each column represents a bacterial species which spacer sequences were extracted from.

Similarly, in Dataset B most species with the number of unique flanking sequences above the set thresholds belonged to *Streptococcus* genus. Despite the large starting dataset, the limited number of target hits considerably reduces the remaining species for which Cas9 PAM can be identified.

3.1.3 Validation

Streptococcus pyogenes was selected for the validation as its Cas9 ortholog (SpCas9) is well characterised and both Dataset A and Dataset B had comparable number of extracted spacers and target hits. Experimentally determined SpCas9 PAM is 5'-NGG-3' and the SpCas9 was classified as CRISPR Type II-A (Sternberg et al., 2014). This therefore acted as a positive control for my approach.

Validation of PAM identification pipeline relied on the PAM prediction methodology described in CASPERpam. Several different filtering approaches were used to determine the amount of background noise, represented by flanking sequences that do not have 5'-NGG-3' PAM present. The filtering used was to some degree different between the flanking sequences obtained from Dataset A and Dataset B, only due to the different available annotations for the two datasets.

Detection of CRISPR arrays in Dataset A was performed by using CRISPRDetect, which scores the quality of detected CRISPR array based on multiple criteria, such as the presence of annotated *cas1* or *cas2* genes in a provided annotated file, previously known or confidently predicted repeats, their identity, specific sequence motifs at the 3'-end, repeat length, spacer length and identity among spacers (Biswas et al., 2016). Based on this criteria a score and a descriptive label is given to the CRISPR array. Furthermore, the direction of CRISPR array is also determined by CRISPRDetect, which is important to accurately identify their cognate viral

sequences and to determine the right orientation of the target sequence and consequently the flank containing PAM.

Different combinations of selection criteria were used to estimate how different properties of *Streptococcus pyogenes* CRISPR arrays and the search parameters affect the amount of noise (false positive sequences) captured with the selection of the flanking regions. For both datasets 16 different criteria were selected based on whether the detected CRISPR arrays were annotated as Type II, the number of mismatches in the target sequence, and the strand of the target. In Dataset A only spacers from CRISPR arrays annotated as non-questionable (high scoring CRISPR arrays) and with high direction probability were selected.

The correct PAM was predicted for eight different subsets of flanking sequences in the Dataset A (Table 1, Figure 13). All filtering approaches that successfully predicted SpCas9 PAM included only flanking sequences from target hits on the positive strand. Collecting flanks from target hits both on positive and negative strand yielded only partial PAMs.

On average, both scores decreased with allowing more mismatches to be present in the target sequence. The highest scoring subset of *Streptococcus pyogenes* flanks from Dataset A were selected from CRISPR arrays annotated as CRISPR Type-II and with no allowed mismatches (R_{avdev} = 1.45, JSD_{sig} = 0.73). Considerably higher scores and qualitatively clearer PAMs (Table 1, Figure 13) were obtained by selecting only flanking sequences that were retrieved through spacers from annotated CRISPR Type II arrays. This was expected since *Streptococcus pyogenes* genome carries both Type I-C and Type II-A CRISPR type loci (Le Rhun et al., 2019). Including spacers to the analysis that come from CRISPR arrays of both types introduced false-positive hits and decreased the difference in the Shannon entropy between the two

significant positions, which form PAM, and the rest of the sequence (Figure 13). This also validates the utility of different selection criteria and using two different starting datasets in order to avoid the confounding effects.

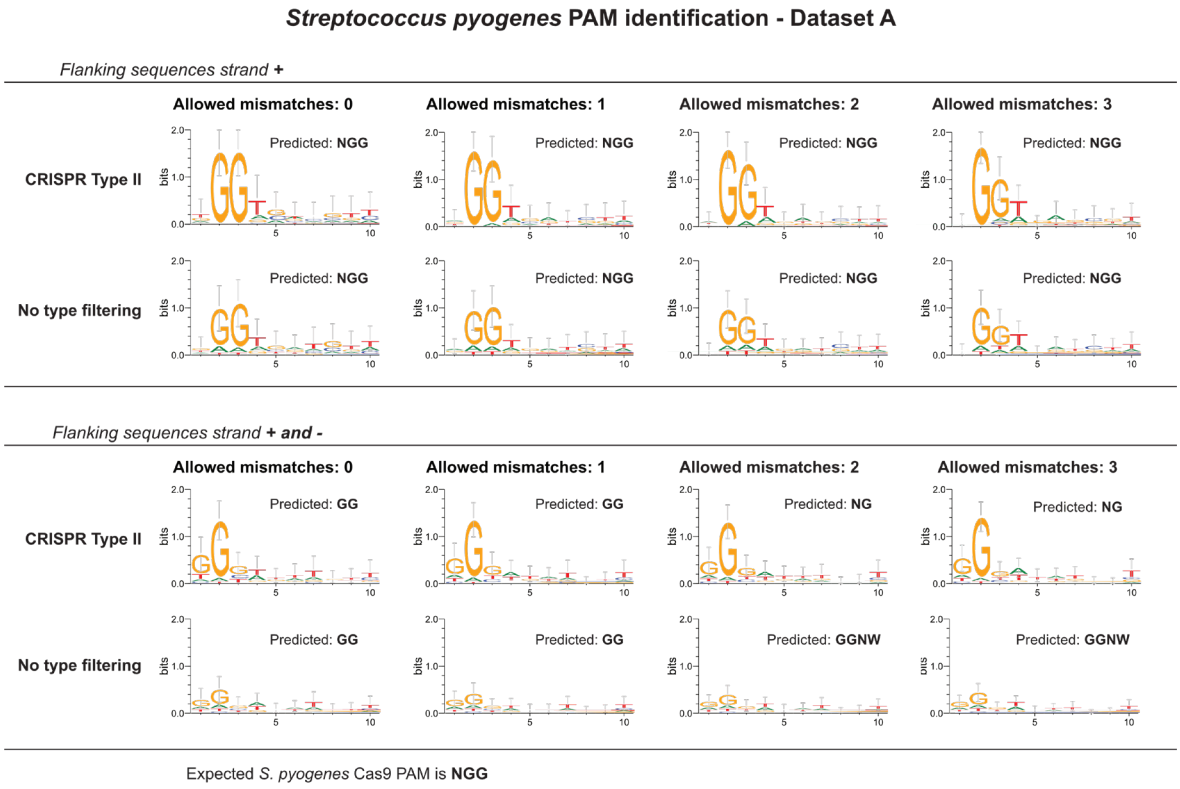


Figure 13 - Sequence logos from PAM identification pipeline for *Streptococcus pyogenes* Cas9 in Dataset A. 3'-end flanking sequences used for prediction of PAM were filtered based on the number of allowed mismatches, the strand of protospacers and annotation of CRISPR array type. PAM was predicted for each subset of flanks.

Table 1 - Predicted PAM sequences for *S. pyogenes* Cas9 based on filtering strategies for Dataset A and Dataset B. Flanking sequences upon which PAM was predicted were filtered by CRISPR type, protospacer strand, and the threshold number of mismatches (MM) in the protospacer. Correctly predicted PAM sequences are underlined.

Filtering			Dataset A			Dataset B		
Filtered for Type-II	Flank strand	<= MM	Predicted PAM	R_{avgdev}	JSD_{sig}	Predicted PAM	R_{avgdev}	JSD_{sig}
×	+	0	<u>NGG</u>	0.82	0.54	NRR	0.37	0.36
×	+	1	<u>NGG</u>	0.77	0.53	NRR	0.33	0.32
×	+	2	<u>NGG</u>	0.68	0.50	NRR	0.36	0.34
×	+	3	<u>NGG</u>	0.57	0.46	NRR	0.16	0.24
×	+ / -	0	GG	0.21	0.31	NRND	0.08	0.16
×	+ / -	1	GG	0.19	0.30	NRNND	0.09	0.15
×	+ / -	2	GGNW	0.13	0.24	NR	0.16	0.23
×	+ / -	3	GGNW	0.16	0.27	RR	0.05	0.15
✓	+	0	<u>NGG</u>	1.45	0.73	<u>NGG</u>	0.41	0.39
✓	+	1	<u>NGG</u>	1.37	0.72	<u>NGG</u>	0.39	0.37
✓	+	2	<u>NGG</u>	1.34	0.71	<u>NGG</u>	0.44	0.39
✓	+	3	<u>NGG</u>	1.11	0.63	NGGD	0.14	0.24
✓	+ / -	0	GG	0.71	0.51	<u>NGG</u>	0.13	0.23
✓	+ / -	1	GG	0.68	0.50	<u>NGG</u>	0.16	0.25
✓	+ / -	2	NG	1.11	0.58	NG	0.30	0.30
✓	+ / -	3	NG	1.18	0.60	<u>NGG</u>	0.06	0.15

The right PAM was predicted for six different subsets of flanking sequences from Dataset B. The highest scoring and correctly predicted PAM in Dataset B was determined based on spacers coming from Type II annotated CRISPR arrays and with no allowed mismatches in the target sequence (R_{avdev} = 0.41, JSD_{sig} = 0.39). The other correctly predicted PAMs were all derived from spacers of only Type II CRISPR arrays. Absence of filtering by type introduced sequence noise from spacers of CRISPR Type I and resulted in only partially correct PAMs (Figure 14). The validation in Dataset B also showed lower scores for the set of flanking regions that were extracted based on sequence hits from both positive and negative strand.

Flanking sequences from sequence hits on only the negative strand were also tested, on both 5'-end and 3'-end of the sequence hit, in both forward and reverse direction,

however no correctly predicted PAMs were observed, indicating that the orientation of the extracted spacer sequences plays important role in *in silico* PAM elucidation. All the combinations had scores lower than any listed in Table 1.

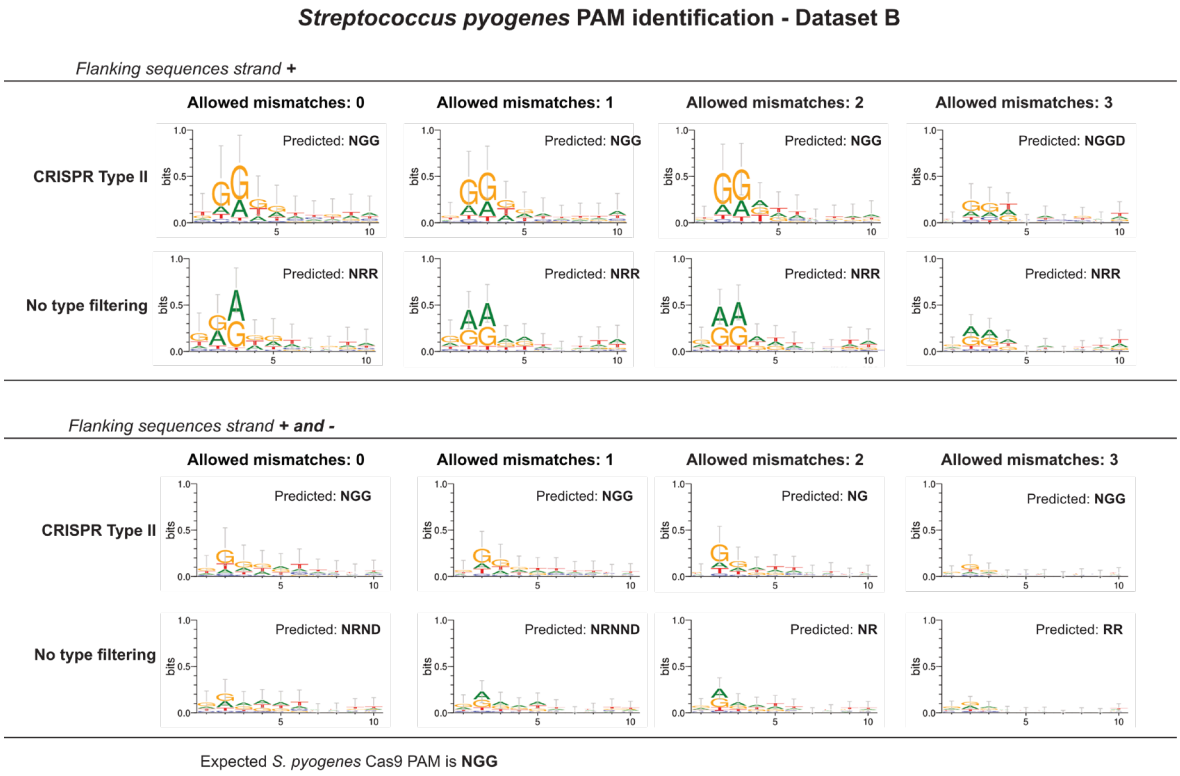


Figure 14 - Sequence logos from PAM identification pipeline for *Streptococcus pyogenes* Cas9 in Dataset B. 3'-end flanking sequences used for prediction of PAM were filtered based on the number of allowed mismatches, the strand of protospacers and annotation of CRISPR array type. PAM was predicted for each subset of flanks.

Overall, the conservation of PAMs derived from Dataset A was higher than conservation in Dataset B. The average Shannon entropy for the position 2 and 3 (constituting 5'-NGG-3' PAM for SpCas9) in the flanking sequences derived from Dataset A was 1.0 bit, whilst the average Shannon entropy of the same positions in flanks from Dataset B was 0.31 bit. The difference can be also observed in sequence logos between logos in Figure 13 and Figure 14 (as height of columns 2 and 3).

Lower overall Shannon entropy does not necessarily mean that PAM identification is less likely, however it does indicate that flanking sequences with lower Shannon entropy contain more false positives.

The novelty of the used pipeline for PAM identification was the stringency of the search algorithm, which allowed few mismatches, and in the set of filtering parameters that were used when selecting flanking sequences. These parameters (spacers from annotated CRISPR arrays, number of mismatches, strandedness of the sequence hit) allowed more granular control over the final data that was used for evaluation and visualisation.

The pipeline additionally correctly predicted the PAM sequences of previously characterised Cas9 orthologs from for *Streptococcus agalactiae* (5'-NGG-3'), *Listeria monocytogenes* (5'-NGG-3'), *Streptococcus thermophilus* CRISPR3 (5'-NGGNG-3'), *Streptococcus agalactiae* (5'-NGG-3') and *Streptococcus mutans* (5'-NGG-3') (Fonfara et al., 2014; Gasiunas et al., 2020; Horvath et al., 2008; Magadán et al., 2012; Mojica, 2009; van der Ploeg, 2009; Zhang et al., 2013).

To find novel Cas9 ortholog candidates for testing *in vitro* both Dataset A and Dataset B were evaluated jointly. Cas9 orthologs from 54 species remained after removing Cas9 orthologs that were already characterised in previous studies. Flanking sequences for the 54 Cas9 orthologs were submitted to same filtering strategies as previously described SpCas9 validation dataset. Furthermore, to remove poorly conserved putative PAM sequences, thresholds of R_{avdev} = 0.25, JSD_{sig} = 0.25 and minimum of 10 flanking sequences were imposed, which resulted in 40 orthologs with mean R_{avdev} of 1.35 and JSD_{sig} of 0.44. These Cas9 orthologs were belonging to 8 genera, namely: *Bacteroides, Campylobacter, Enteroccocus, Lactobacillus, Listeria, Ralstonia, Staphylococcus* and *Streptococcus*. Out of 40

orthologs only 3 showed consistent PAM across different filtering strategies: *Listeria ivanovii* (5'-NRNNC-3', mean R_{avdev} = 0.46, mean JSD_{sig} = 0.48), *Bacteroides fragilis* (5'-NRRWNC-3', mean R_{avdev} = 0.51, mean JSD_{sig} = 0.48) and *Lactobacillus rhamnosus* (5'-NNNAA-3', mean R_{avdev} = 0.58 and JSD_{sig} = 0.31). Upon qualitative inspection of PAM sequence logos we proceeded with further characterisation of *Bacteroides fragilis* and *Lactobacillus rhamnosus* Cas9 orthologs.

3.1.4 Putative PAM sequence identified in a previously uncharacterized Cas9 ortholog

The pipeline was used to investigate two Cas9 orthologs, from *Lactobacillus rhamnosus* and *Bacteroides fragilis.* They were selected on the basis of high R_{avdev} and JSD_{sig} values, which were in the range of values that were observed for validated SpCas9 PAM. The selection was further supported by qualitative inspection of sequence logos as they showed well conserved positions and by the fact that these two Cas9 orthologs were not characterised at the time the results were obtained.

Several *B. fragilis* strains were previously analysed for CRISPR-Cas systems and were shown to include I-B, II-C and III-B CRISPR types, yet no PAM specificities were determined. (Tajkarimi and Wexler, 2017) Using the pipeline, only CRISPR arrays from Dataset A, annotated as Type II were selected with up to 3 mismatches in the target sequence allowed. Only target hits from positive strand were kept. A putative consensus PAM 5'-NRRWNC-3' was predicted from 46 unique flanking sequences (R_{avdev} = 0.59, JSD_{sig} = 0.54), where R stands for either G or C and W stands for either A or T nucleotide in the sequence. For qualitative confirmation a sequence logo was produced (Figure 15).

The second Cas9 ortholog analysed was from *L. rhamnosus* (RefSeq GCF_000026505.1) belonging to CRISPR type II-A. The PAM motif predicted from my analysis (Figure 15) was rich in adenine which made this Cas9 an interesting alternative to other Cas9 orthologs that predominantly recognize G-rich PAMs. In Dataset A 117 unique flanking sequences were available for *L. rhamnosus*, after filtering steps to keep only flanks from target hits on the positive strand, 58 flanks remained and yielded a consensus PAM 5'-NNNAA-3' (Figure 15) with relatively high metrics supporting the significant positions (R_{avdev} = 0.62 and JSD_{sig} = 0.38). Since the NCBI taxonomy identification that was used for grouping of CRISPR arrays in the analysis does not distinguish between different strains, detected CRISPR arrays for *Lactobacillus rhamnosus* (taxid: 47715) were furthermore divided by strain between *L. rhamnosus* (taxid: 47715) and *L. rhamnosus* GG (taxid: 568703). A subset of 12 flanking sequences from *L. rhamnosus* GG showed high conservation in a form of a consensus PAM 5'-NRRAA-3' (R_{avdev} = 1.12 and JSD_{sig} = 0.62).

L. rhamnosus strains have been isolated from various environments, such as dairy fermented products, plants, human and animal intestine and human clinical samples, and can differentiate in strain-specif genes that allow them to adapt to a particular environment (Ceapa et al., 2016). Among them, *L. rhamnosus* GG is one of the most widely used probiotic strains and was shown to have various positive health effects. The strain was initially isolated from human fecal samples by Gorbach and Goldwin, hence letters GG in the name. It exhibits particularly strong mucus adhering properties compared to other *L. rhamnosus* strains (Segers and Lebeer, 2014).

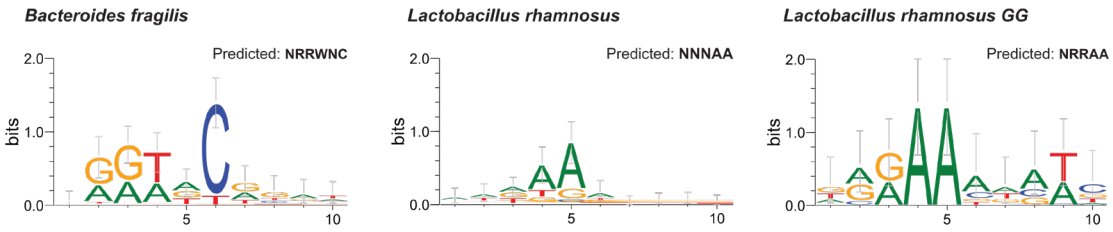


Figure 15 - A putative PAM sequences identified for Cas9 orthologs from *B. fragilis*, *L. rhamnosus* and *L. rhamnosus* GG strain using *in silico* PAM identification pipeline. Sequence logos represent conservation on nucleotide positions in the 3'-end flanking sequences of the target sites in the database of viral sequences.

Experimentally determined PAM for *L. rhamnosus* has been published since the commencement of this thesis, which showed 5'-NGAAA-3' PAM, soon after these *in silico* results were obtained. The authors tested the *in silico* predicted PAM by using interference plasmids that were generated by cloning a spacer and PAM sequence into a plasmid with a selectable marker. In case the PAM was correct and the CRISPR system was functional in the tested *Lactobacillus* species, the Cas9 cleaved the plasmid and no cells were recovered on the selective media. If the PAM was unrecognizable by Cas9, the protein was be unable to bind and cleave, thus allowing the plasmid to replicate and confer antibiotic resistance. In the study they determined 5′-NGAAAA-3′, 5′-NGAAA-3′, 5′-NTAACN-3′, 5′-NGGN-3′, and 5′-NTTAAT-3′ PAMs for *L. casei*, *L. rhamnosus*, *L. gasseri*, *L. jensenii*, and *L. pentosus*, respectively (Crawley et al., 2018).

3.1.5 *In vitro* confirmation of the specificities of the PAMs

Further *in vitro* confirmation of the specificities of the PAMs obtained from *L. rhamnosus* Cas9 (LrCas9) and *B. fragilis* Cas9 (BfCas9) was planned and is currently ongoing in the Crisanti lab (experiments and analysis are being carried out by Dr

Teresa Petrucci). Confirmation of the two putative PAM sequences *in vitro* could potentially mean that this approach to PAM identification is suitable to elucidate novel CRISPR endonucleases and test their activity. *In vitro* validation is performed using cell-free transcription-translation (TX-TL) method (Crooks et al. 2004). TX-TL assay was shown that the method can be a suitable approach for PAM determination of natural and engineered CRISPR nucleases (Marshall et al., 2018). In TX-TL the metabolic, transcriptional and translational machinery from *E. coli* lysates is used as TXTL mix, together with DNA encoding gRNA, and *cas* genes, resulting in gene expression in short period of time. To elucidate PAM specificity for encoded nuclease, a pooled PAM library is added in the reaction (Figure 16). The PAM library is constructed from nuclease targeted sequence and flanked by variable PAM region. Variable region is synthetised to have 7 random nucleotides adjacent to the target site, which means that the theoretical complexity of the variable region is 4^7. The PAM library would therefore represent 16384 unique PAM sequences that could be cleaved. After exposing PAM library to the translated nuclease, the pool of uncleaved targeted sequences is amplified and sequenced. Sequencing the starting library and the cleaved library allows a relative comparison between the two and subsequent quantification of depleted PAM sequences.

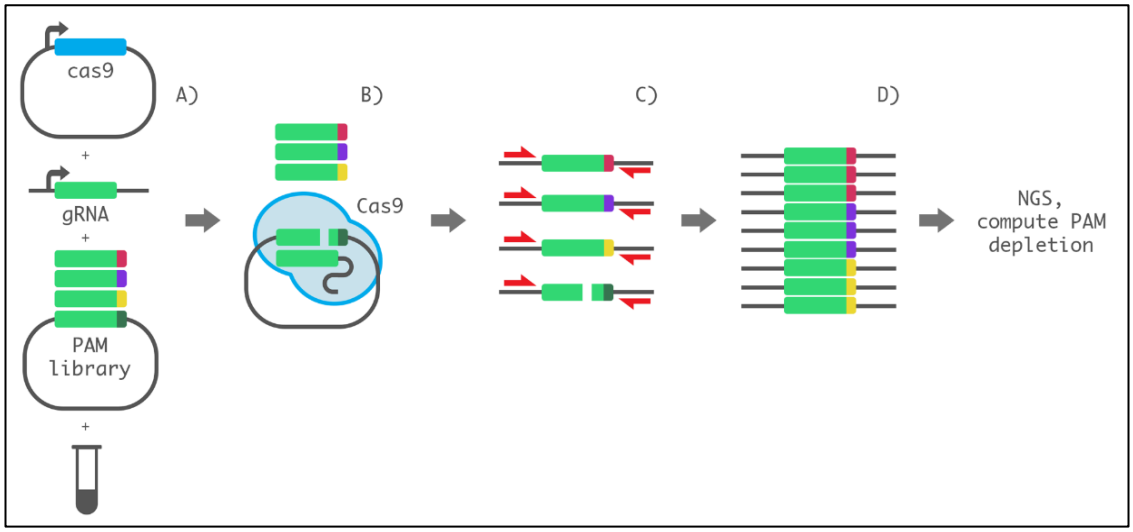


Figure 16 - Schematic of the TXTL method. A) In the reaction plasmid with CRISPR nuclease gene is used together with gRNA, PAM library and E. coli lysate. B) When PAM library is exposed to the expressed nuclease only sequences containing PAM recognized by the nuclease will be cleaved. C) Pool of uncleaved targeted sequences is amplified D) and sent to sequencing in order to compute PAM depletion.

3.2 Assessing the specificity of novel CRISPR-Cas nucleases in the *An. gambiae* genome

Upon obtaining putative PAM sequences for BfCas9 and LrCas9 from the *in silico* PAM identification pipeline I focused the analysis to *An. gambiae* and the specificity of the Cas9 orthologs therein. PAM specificity of a Cas9 endonuclease presents a limitation in accessing target sites for *in vivo* genome editing applications, including gene drive, as the desired target site needs to be in the presence of a PAM that is specific for the Cas9 ortholog in use.

A scan of the current *An. gambiae* PEST genome (AgamP4) was performed to determine the specificity of BfCas9 and LrCas9 based on their putative PAM sequences. For comparison SpCas9 was also included. The scan included potential 20 bp long target sites, which represent a spacer sequence or a genomic target with PAM sequence adjacent to them.

The size of the AgamP4 genome is around 264 Mb, however due to the highly repetitive and poorly assembled Y chromosome only chromosome arms 2L, 2R, 3L, 3R and X were included in the scan, amounting to the genome size of 230 Mb. There were over 21,5 million target sites found for SpCas9 with NGG PAM in the genome on both strands, which equals PAM density of 0.09 bp^{-1} (Table 2). Given the fact that both LrCas9 and BfCas9 have longer putative PAM sequences it was expected their targeting space to be narrower than SpCas9. The scan resulted in 14.0 million target sites for the putative LrCas9 NRRAA PAM, representing PAM density of 0.06 bp^{-1} in the genome. 13.5 million target sites were found for the putative 5'-NRRWNC-3' BfCas9 PAM, the lowest target site density (0.05 bp^{-1}) of the three Cas9 orthologs analysed.

Table 2 - Potential target sites in *An. gambiae* AgamP4 genome for three different Cas9 orthologs: SpCas9, LrCas9, BfCas9

	Whole genome (PAM density bp^{-1})	**Exons**	**Filtered and accessible in exons (% of all targets in exons)**
SpCas9	21,634,376 (0.09)	4,174,970	1,008,122 (24.1%)
LrCas9	13,954,702 (0.06)	1,488,706	314,368 (21.1%)
BfCas9	13,490,459 (0.05)	2,159,878	679,050 (31.4%)

The distribution of PAM density across the genome was determined by counting the PAM occurrences for each Cas9 ortholog in 100 kb windows (Figure 17). The abundance of SpCas9 target sites in each 100 kb window shows strong correlation with the GC-content across every chromosome arm (Pearson $R > 0.85$, $p<0.001$). Centromeric regions of chromosomes show lower GC-content and lower abundance of target sites. The number of LrCas9 and BfCas9 target sites does not correlate strongly with GC-content in the 100 kb windows across the genome (Pearson $R < 0.5$, $p<0.001$) and their density is not affected by the proximity of the centromere, reflecting the fact that their PAMs do not have a strong G-bias. In comparison SpCas9 PAM density was reduced accordingly towards centromeres due to lower GC-content in those regions.

Furthermore, the target sites were filtered to be present in exons to estimate their abundance in functional transcripts and potential genome editing target sites. Most abundant target sites in exonic regions are for SpCas9 (4,174,970 sites), followed by BfCas9 (2,159,878 sites) and LrCas9 (1,488,706 sites). The smallest fraction of target sites in exonic regions compared to target sites in whole genome was found for LrCas9 (Table 2). The explanation for a larger reduction of potential LrCas9 target sites, after filtering for exons, could be attributed to its more AT-rich PAM sequence and larger GC-content in exonic sequences compared to the whole genome (52% GC-content in exons and 44% in whole genome).

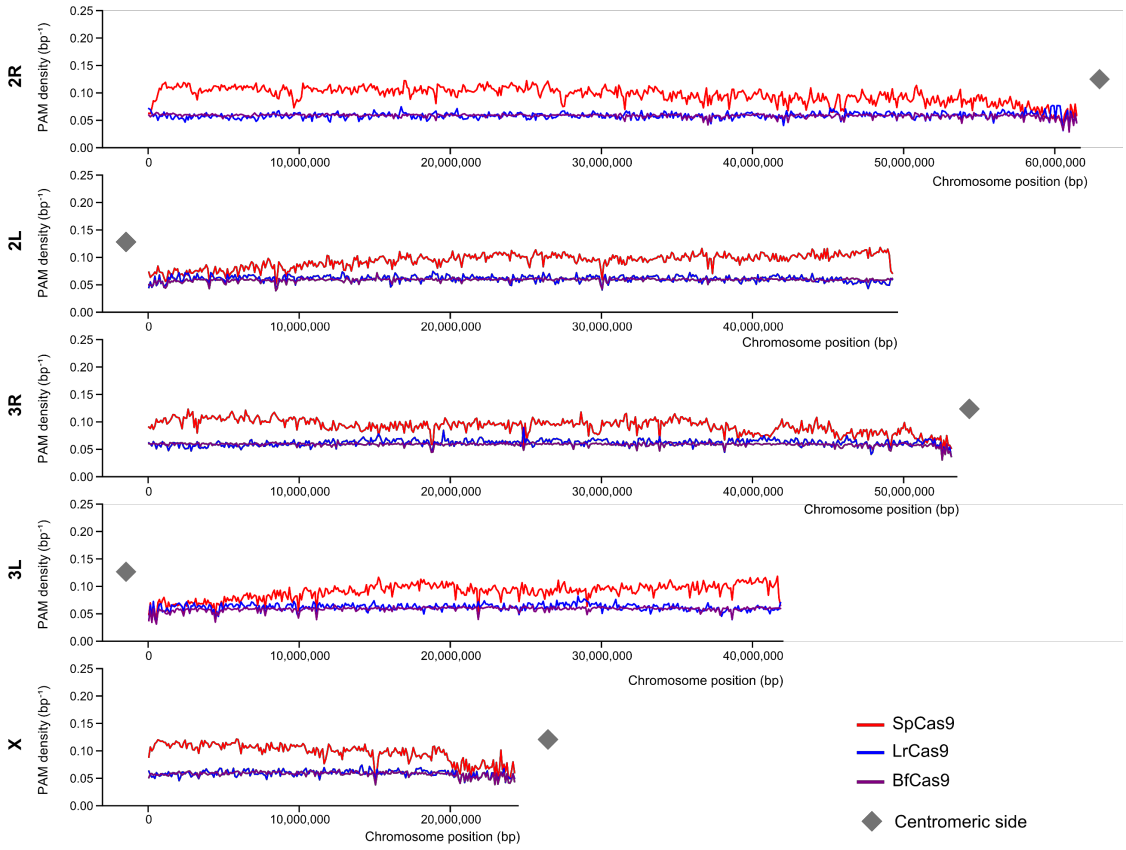


Figure 17 - PAM occurrence density calculated in 100 kb windows for SpCas9, LrCas9 and BfCas9 for each chromosome arm in *An. gambiae* genome AgamP4.

PAM properties of Cas9 orthologs also play important role when assessing the off-target effects. Longer and more specific PAM sequences reduce the number of potential off-target sites that can be recognised by gRNA in the genome. A search for off-targets was performed for each target site from exonic regions. Up to 3 mismatches in the 23 bp long target site were allowed. Target sites that were found in the repetitive, poorly assembled regions or regions of low sequence complexity were also removed, leaving only accessible target sites. For SpCas9 24.5% of target sites had one or more off-targets, whereas for LrCas9 and BfCas9 the percentage was smaller, 15.0% and 5.3%, respectively (Figure 18). However, the off-target

count for LrCas9 and BfCas9 could potentially be higher if the full sequence flexibility of their PAMs was accounted for. The off-target search was performed by using the full target sequence (including PAM) as query, meaning the 3 mismatches could be distributed along the gRNA but also within the flexible positions in the PAM, namely on positions labelled with N (allowing nucleotide to be either A, C, T, G), W (A or T) or R (A or G). Results allowing 3 mismatches therefore prevent direct comparison of off-targets between the three orthologs, but can serve as an approximation for the off-target abundance that each Cas9 ortholog could potentially have.

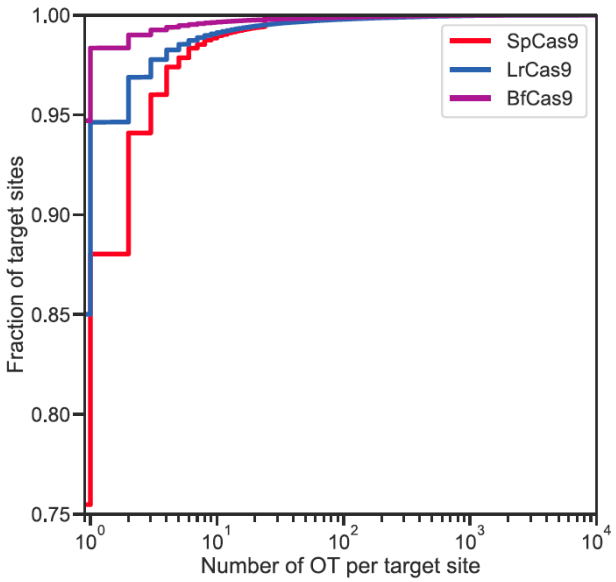


Figure 18 - Cumulative distribution function for the distribution of number of off-targets (OT) per target site in the AgamP4 genome for target sites of SpCas9, LrCas9 and BfCas9.

Additional filters were applied to estimate the abundance of potential targets sites (PTS) that are within likely functionally constrained exonic regions (based on conservation score – described in detail in Chapter 4.1) and that have low number

of off-targets. Either two off-targets with three mismatches or one off-target with two mismatches were allowed for each PTS. The filtering step reduced the target sites to ~20-30% from the initial number of exonic targets. The final number of filtered and accessible target sites was 1,008,122 for SpCas9, 314,368 for LrCas9 and 679,050 for BfCas9.

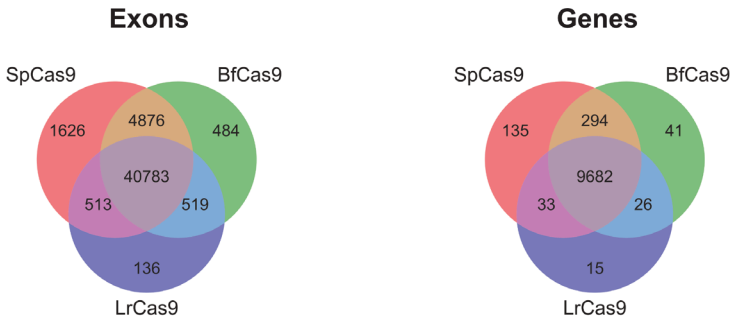


Figure 19 - Venn diagrams for exons and genes targetable by SpCas9, LrCas9 and BfCas9.

Out of 13,057 genes on autosomes and the X chromosome 10,226 genes (78.3%) contained PTS by any of the three Cas9 orthologs. The majority of genes with PTS have targets for all three Cas9 orthologs, however there are 135, 41 and 15 genes with PTS specific only to SpCas9, BfCas9 and LrCas9, respectively (Figure 19). The higher abundance of target sites for SpCas9 and BfCas9 within coding sequences could be attributed to higher GC content in genes compared to the rest of the genome.

3.3 Discussion

Starting with *in silico* PAM determination, a bioinformatic pipeline was set up to identify CRISPR arrays in the provided bacterial genomes, extract spacer sequences carrying information about invading sequences from foreign organisms and match them with sequences from viral databases using a search algorithm. The genomic locations obtained from matched sequences in viral databases were used to extract flanking sequences from viral source genomes. As PAMs are adjacent to the 3'-end of the spacer matching sequence for Type II CRISPR-Cas systems, extracted flanking regions should show a conserved motif in the first 10 bp of the flanking sequences. The pipeline was validated on previously characterised Cas9 orthologs and their PAM specificities.

Two different datasets were used in the analysis – Dataset A was collected in house through the detection of CRISPR arrays in publicly available genomes of bacterial species with annotated Cas9 proteins in InterPro database, and Dataset B that was assembled from two published and well annotated datasets. The rationale behind using Dataset B was to reduce the number of spacer sequences deriving from CRISPR loci that do not belong to CRISPR Type II. For that purpose the dataset with already annotated CRISPR loci and their associated spacers were used (S. A. Shmakov et al. 2017, 2018). CRISPRDetect, which was utilised to generate Dataset A, has a limited ability to predict the CRISPR type, only through a library of already annotated CRISPR repeats.

Using this approach only spacers that could be found in the CRISPR arrays of a particular CRISPR type were selected and used in the analysis. This is important as some bacterial genomes include multiple CRISPR loci that can be attributed to different CRISPR-Cas systems. Assuming that all CRISPR arrays, which can be found

in the same genome, belong to the same CRISPR type could lead to capturing spacer sequences that contribute to noisy, false positive targets and flanking sequences. Using the annotated dataset (Dataset B) expanded the set of CRISPR arrays found in Dataset A and consequently provided a larger pool of spacers that can be used for viral target identification. A larger pool of spaces theoretically increased the chance to obtain a true positive hit in viral database, however it also evidently decreased the overall conservation of observed PAM sequences.

An important step in determining PAMs is the transcription direction of CRISPR arrays. In the step of spacer acquisition, excised sequences are added to the 3'-end of the CRISPR array and the PAM sequence is recognized on the 3'-end of the gRNA sequence for Type II CRISPR-Cas systems (Sternberg et al., 2014). To be able correctly predict PAMs from aligned spacer flanking sequences, all sequences must be in the correct orientation. Adding flanking sequence extracted from the wrong side of the spacer to the flanking sequence stack can be a source of noise as this sequence will not contain a PAM. To avoid this both flanking regions, on the 5'- and 3'-ends were examined, for each CRISPR array before grouping them together, as outlined by Mojica (2009). Despite accounting for the directionality of the CRISPR array and the orientation of the protospacers, flanks extracted from the 3'-end of protospacers on the positive strand showed better PAM conservation. The mechanism of spacer integration that would support the observed bias in type II systems is still not completely clear. The acquisition machinery that includes a Cas1-Cas2 complex is needed to search for potential targets and is accompanied by Cas9 which specifies functional targets by recognising the PAM (Heler et al., 2015). It was also observed that new spacer sequences are integrated at leader-side CRISPR repeat and it was suggested that *in vivo*, spacers are almost always integrated in the same orientation as their parental protospacers (Xiao et al., 2017).

Arguably the major bottleneck in the *in silico* approach to PAM identification is the size of the viral database, more precisely the availability of bacteriophage sequences (Collias and Beisel, 2021; Karvelis et al., 2017). The ability to identify PAM sequence for a particular Cas9 ortholog *in silico* in the end comes down to obtaining the protospacer sequence's match in the viral database. It was shown recently that most of the spacers originate from the host-specific viromes (Shmakov et al., 2020). This could presumably explain the enrichment of viral database search hits in *Streptococcus* and *Lactobacillus* genus, as the phage-host interactions in those species are relevant in medicine and food industry and therefore the availability of sequencing data of their viromes is disproportionately larger (Chibani-Chennoufi et al., 2004). Additionally, the available target hits can further be reduced by imposing stringent search parameters, such as the level of sequence identity, which was shown to improve the final PAM resolution. The bioinformatics approach for PAM identification could potentially be more fruitful by obtaining new metagenomic samples of both bacteria and bacteriophages from a particular ecological niche (Burstein et al., 2017).

Nonetheless, a novel putative PAM sequences for a Cas9 ortholog was identified from *B. fragilis* (BfCas9). Additional Cas9 ortholog from *L. rhamnosus* GG was selected for its adenine-rich PAM specificities to be further tested *in vitro.* Both CRISPR loci encoding the two selected orthologs were categorized as CRISPR type II. The putative PAM sequence of *L. rhamnosus* Cas9 (LrCas9) contained two very well conserved adenines constituting a PAM sequence of 5'-NRRAA-3'. It could potentially expand a genome editing toolbox with an ortholog that could target A-rich sequences, unlike most currently widely used Cas9 variants that predominantly recognize G-rich PAM sequences. The putative PAM specificities of *B. fragilis* Cas9 showed an extended 5'-NGG-3' PAM, with a T and A preference on position 4 and a well conserved cytosine on position 6, constituting a putative 5'- NRRWNC-3' PAM

sequence. It could potentially increase the target specificity by offering a longer and more diverse PAM sequence. Based on the obtained results from the *in silico* pipeline described here, further research is being conducted to test the two Cas9 orthologs *in vitro* by using the TX-TL system (Teresa Petrucci, personal correspondence).

A genome scan was performed with the goal to put the two obtained putative PAMs into the context of *An. gambiae* genome and to compare their targeting space with SpCas9, which is currently most widely used in genome engineering in the mosquito. SpCas9 with the 5'-NGG-3' PAM was expected to be the least specific of the three orthologs. A considerable amount of target sites was found for SpCas9, in average every 11.1 bp (average NGG PAM density of 0.09 bp^{-1} in non-overlapping 100 kb windows). In comparison, a target site for putative LrCas9 PAM with 5'-NRRAA-3' sequence could be found in average every 16.6 bp and every 20 bp the least specific BfCas9 with 5'-NRRWNC-3' PAM sequence. In theory this presents ample target sites for use in genome engineering, or specifically for gene drive applications in *An. gambiae* genome. A similar analysis of targeting space in the *An. gambiae* genome was performed which accounted for the role of resistant alleles caused by naturally occurring SNPs in the wild populations that can prevent the gRNA binding within the genome of *An. gambiae*. Their results indicated that the flexibility of Cas9-based gene drives and the availability of target sites dismisses the concern of the standing genetic variation to be detrimental to deploying gene drives in the field. They conclude that approximately 90% of all protein-coding genes still have at least one target site available (Schmidt et al., 2020). However, the standing genetic variation is not the only source of variation at the target site, which needs to be further accounted for when selecting gene drive target sites. The analysis that was performed in this study resulted in slightly smaller fraction of targetable genes in the *An. gambiae* genome, most likely due to more stringent sequence conservation requirement than the one used in the analysis by Schmidt et al. (2020).

SpCas9 is a suitable endonuclease in terms of availability of target sites in the *An. gambiae* genome genome. The targeting space could be potentially extended by introducing BfCas9 and LrCas9 which would increase the targeting specificity with longer PAM sequences. Since the putative PAM sequence of BfCas9 is extended SpCas9 PAM, it could be used as a substitute for SpCas9 where the adjacent sequence would allow, potentially increasing specificity and reducing the likelihood of off-targets. However, the adaptation of a Cas9 ortholog for *in vivo* genome editing requires additional steps. First the PAM specificity for BfCas9 need to be successfully confirmed *in vitro*. Further *in vitro* experiments are required to fully understand the temperature profile of activity of both LrCas9 and BfCas9 as it could greatly impact their usability in the mosquito. The precise cutting position in the sequence and the overhangs created by the endonuclease need to be defined before moving the testing to *in vivo* context. The feasibility of their use in the mosquito for gene drive purposes also depends on their cutting efficiency as it would greatly affect the homing rate of the gene drive carrying the Cas9 ortholog.

3.4 Methods

3.4.1 Dataset

The initial attempt to set up the PAM determination pipeline focused on the Class II CRISPR-Cas system, specifically type II, as its signature effector protein, SpCas9, has been well characterized and its architecture is well suited to genome editing applications. To limit bacterial genomes to only the ones containing Class II, type II CRISPR-Cas system, all 1578 proteins from the Cas9 family of orthologs were selected from the EBI InterPro database under the accession id IPR028629 and the genomes that contained a Cas9 orthologue were downloaded from NCBI FTP for 248 bacterial species. 12144 CRISPR arrays were detected using CRISPRDetect, out of those 5863 were determined with high confidence and used in further analysis. The program discovers putative arrays in genomes and has an enhanced annotation of CRISPR array orientation, finds precise repeat-spacer boundaries, and small and large mutations (substitution, deletion and insertions) in spacers and repeats (Biswas et al., 2016). CRISPRDetect resulting reports with annotated CRISPR arrays were parsed to extract 65455 spacer sequences. This was done by using custom Python parser written for this purpose. The dataset obtained using this approach was named Dataset A.

Another dataset, here named Dataset B, was used which was a combination of two published datasets - a dataset containing 7282 annotated CRISPR type II arrays and a dataset containing 43090 spacer sequences. The first dataset was comprised of fully annotated CRISPR-Cas loci derived from all publicly available genomes from NCBI (Shmakov et al., 2018). The dataset contained the locations of CRISPR loci, CRISPR arrays and CRISPR-associated neighbouring genes, which were used to classify the locus into the right CRISPR-Cas type. The second dataset consisted of

spacer sequences extracted from CRISPR arrays from all archaeal and bacterial genomes (S. A. Shmakov et al., 2017).

The two published datasets were combined based on the NCBI identification and organism taxonomy id, and spacers were matched to their source CRISPR arrays as characterized in the first dataset. This allowed obtaining more information about CRISPR loci and their spacers and consequently refine the selection of spacers used in querying the viral and plasmid databases. The joint dataset of published spacer sequences and their corresponding CRISPR arrays dataset was named Dataset B.

Scripts used for data manipulation and analysis from this chapter are available on GitHub https://github.com/nkran/PAM-identification.

3.4.2 Target search

The spacer sequences extracted from Dataset A and Dataset B were used as query sequences for search in the database of viral (taxid: 10239) sequences to obtain the source sequences of the invading virus. Search was performed by using bowtie v1.2.3 using a seed length option of 23 (-l 23) and the number of allowed mismatches was set to 3 (-v 3). All target hits were reported (-a) and annotated with the number of mismatches, genomic location and strand (Langmead, 2010).

3.4.3 Extraction and visualization of flanking regions

With the genomic positions of viral DNA targets identified, 10 bp long sequences adjacent to the target, upstream and downstream of the matching sequence, were extracted using *samtools faidx* (Li et al., 2009). Flanking sequences were aligned relative to the beginning or end of the spacer sequence. A custom Python script was used to stack flanking sequences together and group them by the CRISPR-Cas types

of their CRISPR loci and by their organisms based on NCBI taxonomy id. This grouping was conducted to obtain specific PAMs for specific taxonomic groups which also had Cas9 orthologs annotated.

Duplicates with entirely identical flanking sequences were marked and removed to avoid false conservation coming from the multiple identical sequences. In order to visualize and discern PAM sequences WebLogo, a sequence logo generator was used (Crooks et al., 2004). Furthermore, to quantify the conservation of nucleotide positions in the flanking sequence, average deviation of Shannon entropy, which explains the sequence conservation at a particular position, was used as a metric. The calculation of average deviation of Shannon entropy was implemented as described and used in CASPERpam (Mendoza and Trinh, 2018). Calculation of divergence between two nucleotide frequency distributions by using Jensen-Shannon divergence as a measure of conservation was performed using Python package for discrete information theory - dit v1.2.3. (G. James et al., 2018) Significantly conserved positions contributing to PAM in the flanking sequences were predicted by the methodology used in CASPERpam, based on the comparison of relative nucleotide frequencies on each flanking position (Mendoza and Trinh, 2018).

3.4.4 BfCas9 and LrCas9 gRNA design and testing for *in vitro* activity

Diverse Cas9s require appropriate engineered single guide RNA with specific adapter guide sequence for the binding to the protein counterpart. The gRNA design requires the fusion of a repeat sequence from CRISPR array and its matching tracrRNA (Cong et al., 2013). The repeat sequence without spacer was therefore used as a query to find a target sequence, encoding tracrRNA, in the neighbourhood

of *cas9* locus with RNApredator (Eggenhofer et al., 2011). A matching sequence was found, and its RNA transcript was used in predicting the secondary structure when fused together with the crRNA (crRNA; interspaced repeat sequence in CRISPR array) as shown in Figure 20. The secondary structure of crRNA-tracrRNA dimer was predicted by RNAcofold and visualised by RiboSketch (Bernhart et al., 2006; Lu et al., 2018). Based on the sequence and secondary structure of the from *Streptococcus pyogenes* gRNA and published gRNA variants for *Lactobacillus gasseri* Cas9 (Crawley et al., 2018), 4 different gRNAs and crRNA:tracrRNA complex for each BfCas9 and LrCas9 were designed with modifications within the lower stem, bulge and nexus.

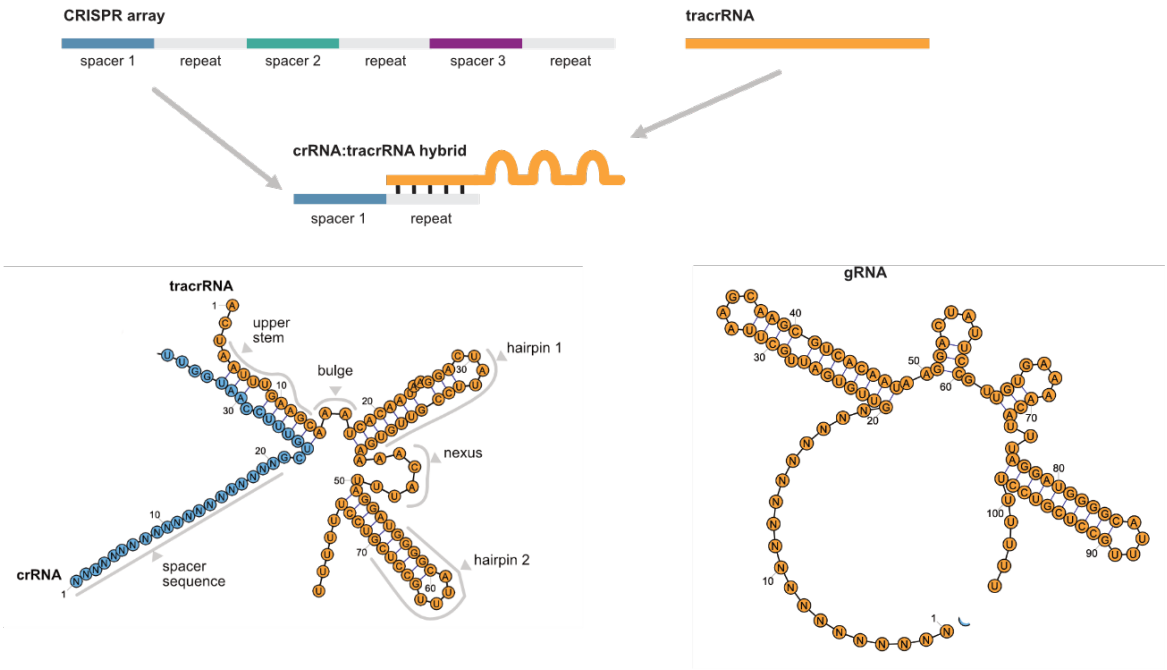


Figure 20 - crRNA and tracrRNA hybridization.

CRISPR is transcribed and processed into crRNA that contains spacer sequence and a CRISPR repeat. tracrRNA is encoded in the genomic neighborhood of CRISPR array and when expressed hybridizes with crRNA into crRNA:tracrRNA RNA complex. crRNA:tracrRNA complex then guides Cas9 endonuclease to the target site. Below are crRNA:tracrRNA and one version of fused gRNA used for testing of BfCas9 endonuclease.

3.4.5 Genome scan

Sequences from autosomal chromosome arms 2L, 2R, 3L and 3R and the X chromosome from *An. gambiae* PEST reference genome (AgamP4) were used for the scan for potential target sites with three different Cas9 orthologs – *Streptococcus pyogenes* Cas9 (SpCas9, 5'-NGG PAM), *Lactobacillus rhamnosus* GG Cas9 (LrCas9, 5'-NRRAA PAM) and *Bacteroides fragilis* Cas9 (BfCas9, 5'-NRRWNC PAM). The genome-wide scan was performed used Guido, a gRNA design software tool that I designed that is optimised for potential applications of CRISPR-based genome editing in the mosquito. Guido is described in detail in Chapter 4.3.1. A target site (or gRNA) was defined as 20 bp long sequence that is restricted by PAM sequence on the 5'- end of the target sequence. Bowtie v1.2.3 was used to search for off-targets for each gRNA, allowing up to 3 mismatches in the sequence. Potential gRNAs were filtered to be present only in exonic regions based on AgamP4 reference annotation (AgamP4.12). Exonic regions included 5'-UTR regions, coding sequences and 3'-UTR regions for each gene.

Only gRNAs that have the predicted cut site within the boundaries of annotated exons were kept. The filtering was performed using bedtools v2.30.0. (Quinlan and Hall, 2010) Furthermore, the target sites within exons were filtered based on the number of off-targets reported for each gRNA. Only two off-targets with 3 mismatches or one off-target with 2 mismatches were allowed. The accessibility dataset from Ag1000g phase II was used to filter out potential target sites that are found in poorly assembled regions, in regions that have repetitive sequences and regions that showed poor mapping quality of WGS data obtained in the Ag1000g phase II project (The Anopheles gambiae 1000 Genomes Consortium, 2021, 2020). To confine the selection of potential target sites to regions with higher sequence conservation, which could imply a higher degree of functional constraint, additional

filtering was performed based on conservation score calculated for AgamP4 reference genome. Cs value of 0.2 was used as the threshold for the final filtering step (Kranjc et al., 2021). The calculations and visualisations were obtained using NumPy v1.20.3, Pandas v1.2.4 and SciPy v1.6.3 Python packages. (Harris et al., 2020; Reback et al., 2021; Virtanen et al., 2020)

4 Rational selection of genomic target sites for CRISPR-based gene drives

4.1 Identification of ultra-conserved genomic regions

One of the gene drive resistance mitigation strategies is to target an evolutionary conserved genomic context that would upon any kind of disruption impose a very high fitness cost. Any mutations in such regions associated with high functional constraint are therefore removed. This process consequently reduces the probability of any sort of selection of gene drive resistant mutations that arise at the target site. In order to find such loci I analysed genetic diversity among individuals of the same species and between individuals of different species.

Genetic diversity is a direct consequence of biological evolution. It is a key component of adaptation to environmental changes over a longer period of time. Genetic diversity is created through DNA sequence polymorphisms that can appear in individuals spontaneous through errors introduced in cell processes involved in DNA replication and repair or through external factors that mutagenise DNA, such as radiation and chemicals. The rate at which mutations accumulate can be different across the genome and can vary between species (Hodgkinson and Eyre-Walker, 2011; Lynch, 2010; The Anopheles gambiae 1000 Genomes Consortium, 2020). Across generations some genetic variants (or alleles) can increase in frequency in the population and become fixed, whilst other alleles can get lost, both processes governed by natural selection. Genetic drift, which is the random fluctuation of allele frequency across generations, is another force that influences the observed variability in loci at any position in the genome. (Charlesworth, 2009) Negative (or puryfying) selection happens when deleterious mutations, that diminish individual's fitness, occur in the genome. In this process genetic variation at the locus carrying deleterious mutation, as well as at the linked neutral loci, gets

reduced (Charlesworth et al., 1993). The constraint is strongly dependent on fitness cost of the mutation. Mutations in functional sites with high fitness cost are under larger constraint and will be quickly removed, thus contributing very little to sequence diversity when compared with other individuals of the same species or across species (Cooper and Brown, 2008).

Previous gene drive experiments targeting genes with essential functions involved in female fertility showed strong selection pressure for gene drive resistant alleles that restored gene function, a problem that was long foreseen at inception (Burt, 2003; Hammond et al., 2016). With this in mind a CRISPR-Cas9 gene drive construct was created that targeted highly conserved female-specific exon 5 in *dsx* gene. Targeting *dsx* locus led to full suppression of caged populations without generating any gene drive resistant alleles that could restore function to disrupted female-specific *dsx* isoform. The deleterious mutations generated through DNA repair mechanisms were quickly removed from the population as they were not able to restore the female phenotype and fertility and therefore increase in frequency (Kyrou et al., 2018). This highlighted the requirement for meticulous selection of gene drive target sites that are under high functional constraint and therefore can not tolerate any generated mutations at the target site that would lead to gene drive resistance.

4.1.1 Conservation score calculation

An *in silico* analysis was employed to uncover functionally constrained regions in gene drive candidate genes and elsewhere in the *An. gambiae* genome based on sequence comparison between different organisms. The sequence conservation is

expected to decrease with the evolutionary distance between the species and the choice about which to include in the analysis relies on the desired stringency. The analysis I performed integrated conserved syntenic regions of 19 *Anopheles* species, including five species from *Anopheles gambiae* complex, and three more phylogenetically distant Dipteran species that span more than 100 million years of evolution (Figure 1). Mutations that could be found in highly-conserved genomic regions that would, for example, be shared across all the compared species, would very likely be associated with high fitness cost, which would imply functional constraint.

Representative reference genomes of the following *Anopheles* species were used in the analysis: *An. coluzzii*, *An. arabiensis*, *An. quadriannulatus*, *An. melas*, *An. merus*, *An. epiroticus*, *An. christyi*, *An. sinensis*, *An. minimus*, *An. maculatus*, *An. culicifacies*, *An. stephensi*, *An. funestus*, *An. atroparvus*, *An. dirus*, *An. farauti*, *An. darlingi* and *An. albimanus*. Two additional mosquito reference genomes were included, namely *Aedes aegypti* and *Culex quinquefasciatus*, as well as the phylogenetically most distantly related species to *An. gambiae*, genome of the fruitfly *Drosophila melanogaster*.

Pairwise genome alignments were performed between reference genome of each species and *An. gambiae*. Syntenic regions from each pair of aligned genomes were further scaned for regions of high sequence identity (>70%) using a sliding window of 30bp and 50bp. The identity percentages were obtained for each position of the autosomes and the X chromosomes in AgamP4 reference genome and stored in a large matrix.

The matrix of the identities for each species in comparison was normalised based on the relative phylogenetic distance (eq. 4.1). In eq 4.1, index *m* represents the

number of genomes in comparison, index n is the length of a chromosome and d_m is the phylogenetic distance between genome m and AgamP4. $I_{m,n}$ represents the sequence identity in the matrix between genome m and AgamP4 on the position n.

$$\tilde{I}_{m,n} = \begin{bmatrix} I_{1,1} & \cdots & I_{1,n} \\ \vdots & \ddots & \vdots \\ I_{m,1} & \cdots & I_{m,n} \end{bmatrix} \begin{bmatrix} d_1 \\ \vdots \\ d_m \end{bmatrix} \quad 4.1$$

Additionally, to account for selection forces applying in the genome phyloP ("phylogenetic P-values") score was calculated for each position in the genome. (Hubisz et al., 2010) It detects departure from the neutral rate of substitution in both positive way (acceleration) and negative way (conservation). The method was applied to the multiple alignment of species genomes obtained from pairwise alignments relative to AgamP4. During the calculation a statistical test is employed at each position to examine increases or decreases in the rate of substitution across all genomes used in the analysis (Pollard et al., 2010).

Intra-species variation was also considered in the calculation of conservation score. Positions of SNPs detected in 1142 wild-caught individual mosquitos of *An. gambiae* and *An. coluzzii* collected in the Ag1000g phase 2 project were mapped in the genome-wide matrix. SNP density σ was calculated in the 20 bp long sliding window on each position in the genome.

Conservation score (Cs) was calculated, utilising inter-species variation, intra-species variation and the selection forces applying at each genomic position. Average normalised identity Ĩ, as the measure of inter-species variation, was calculated for each position (n) in the genome across all the genome alignments (eq 4.2).

$$\tilde{I}_n = \frac{\sum_{i=0}^{m} \tilde{I}_{i,m}}{m} \quad 4.2$$

SNP density σ_n for each position n was transformed (eq 4.3) to obtain higher values for low SNP density.

$$\sigma_n = \frac{1 - \sigma}{1 + \sigma} \quad 4.3$$

PhyloP conservation score is reported by PHAST (Hubisz et al., 2010) as $-\log_2 pval$ and was for this reason transformed (eq 4.4) for each position n to obtain high p_n values for significantly conserved positions.

$$p_n = \frac{1}{2^{-p}} \quad 4.4$$

Finally, Cs was calculated as product of $\tilde{I}_n$, σ_n and p_n (eq. 4.5).

$$Cs_n = \tilde{I}_n \cdot \sigma_n \cdot p_n \quad 4.5$$

Cs were scaled to a range of values between 0 and 1 using MinMax scaler. Positions with high conservation correspond to values close to 1 and positions with values close to 0 show poor conservation.

4.1.2 Conservation across the *An. gambiae* genome

Storing data in a matrix with base pair resolution allowed insights into conservation of a particular locus in interest or, for example, to compare the genome-wide sequence identity across any given set of genomes used in the analysis.

To highlight differences in sequence identity and to validate the methodology a genome-wide comparison between two groups of genomes was conducted. In one group genomes of five sibling species of *Anopheles gambiae* complex were included, namely *An. coluzzii, An. arabiensis, An. quadriannulatus, An. melas, An. merus*. The other group consisted of five more distantly related species: *An. darlingi, An. albimanus, Aedes aegypti, Culex quinquefasciatus,* and *Drosophila melanogaster* (Figure 21).

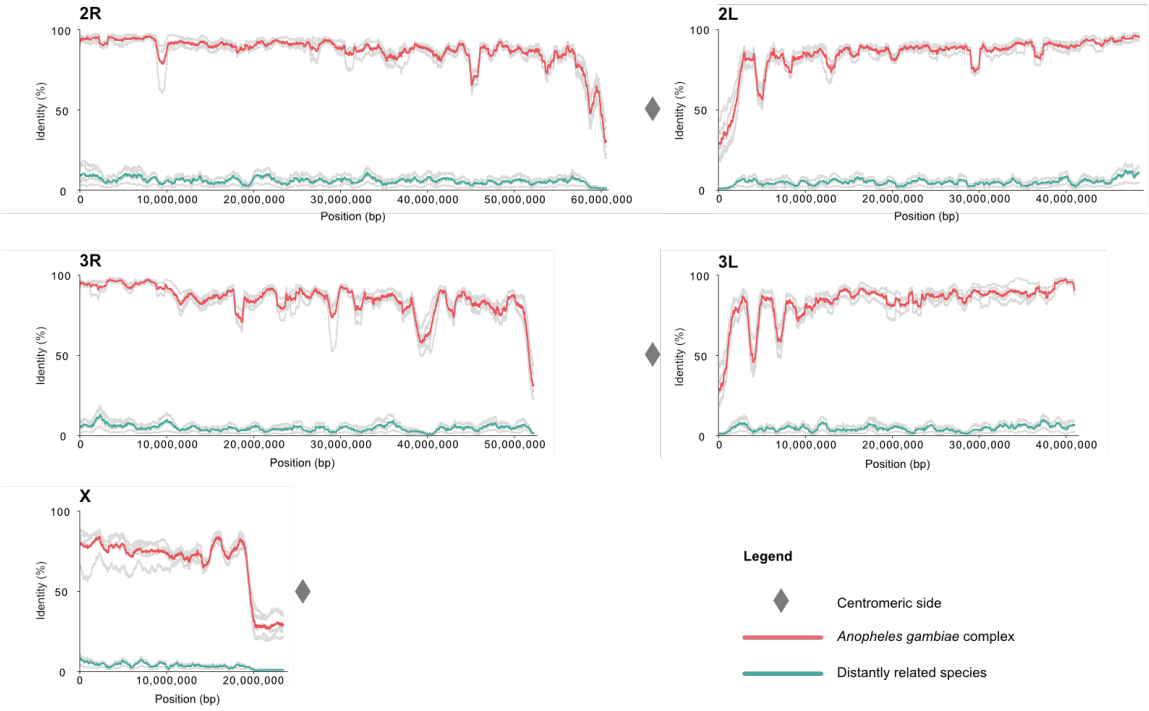


Figure 21 - Average sequence identity in *An. gambiae* AgamP4 reference genome. Two groups of genomes were created: 5 species from Anopheles gambiae complex (An. quadriannulatus, An. melas, An. merus) and 5 distantly related species (An. darlingi, An. albimanus, Aedes aegypti, Culex quinquefasciatus, Drosophila melanogaster) represented by red and green colour, respectively. Sequence identity (in percentage on y-axis) for species in both groups was calculated in a sliding window of 1 Mb represented by grey traces on the chart. Mean sequence identity was calculated for each group in a sliding window of the same size. Centromeric side of the chromosome arm is marked with ♦.

Average detected sequence identity of genomes of *Anopheles gambiae* complex species was 81.0%. The mean sequence identity was higher in autosomes (84.4%) than in X chromosome (67.2%). The mean sequence identity of distantly related species was considerably lower, 5.12% in autosomes and 3.69% in X chromosome, with mean sequence identity at 4.8%. Both groups showed decreased sequence conservation in the regions proximal to centromeres (<50% for *Anopheles gambiae* complex). The results validate the expected discrepancy in sequence identity between species that were previously shown to be evolutionary distant. Furthermore, the lower sequence identity that was observed on the X chromosome is in line with previous observations of higher rearrangement rates on the X (Neafsey et al., 2015). Based on observed differences in sequence identity deriving from multiple alignment the results were used as the base of the proposed conservation score calculation, which was further improved by including the intraspecies variation.

4.1.3 Conservation at the *doublesex* gene drive target site

The analysis was then focused on the gene drive target site in female-specific exon 5 of *doublesex* gene, where, to date in the laboratory, mutations that confer functional resistance to the gene drive have not been selected (Hammond et al., 2021b; Kyrou et al., 2018). As expected, the validated gene drive target site on the intron 4 – exon 5 junction showed very high sequence identity across most of the species included in the analysis (Figure 22). The sequence of the target site was 100% identical in all genomes of *Anopheles gambiae* complex species. High sequence conservation was detected among three of the genomes of distantly related species as well, namely *An. albimanus* (92.8%), *Aedes aegypti* (86.5%) and *Culex quinquefasciatus* (86.2%). Two genomes, *An. darlingi* and *Drosophila melanogaster*

did not show sequence conservation above the required threshold of detection in the locus.

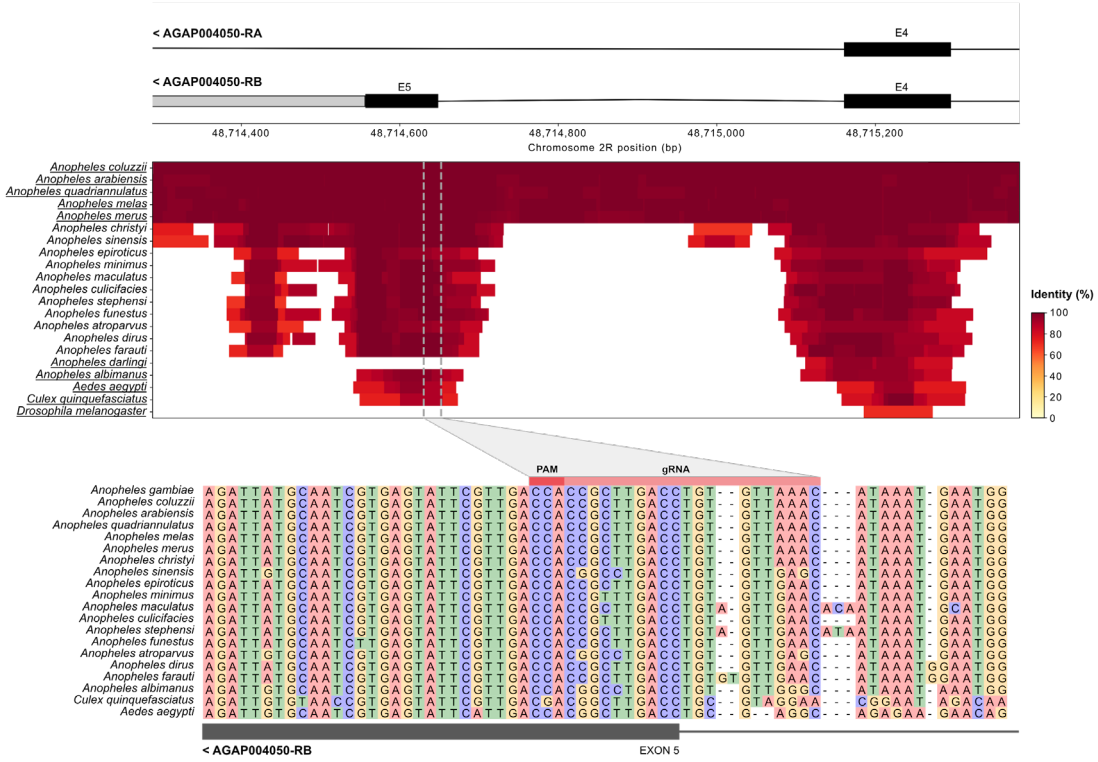


Figure 22 - Sequence identity at the gene drive target site locus in *doublesex*. The heatmap represents the sequence identity for each of the species in the analysis. Underlined are the species of *Anopheles gambiae* complex and the group of distantly related. Dashed grey lines mark the start and end position of gene drive target site in female-specific exon 5 described in Kyrou et al. (2018). The target site showed high sequence conservation (> 70%) across all except two species, namely *An. darlingi* and *Drosophila melanogaster*. Multiple sequence alignment is shown under the heatmap for the region around the gene drive target site.

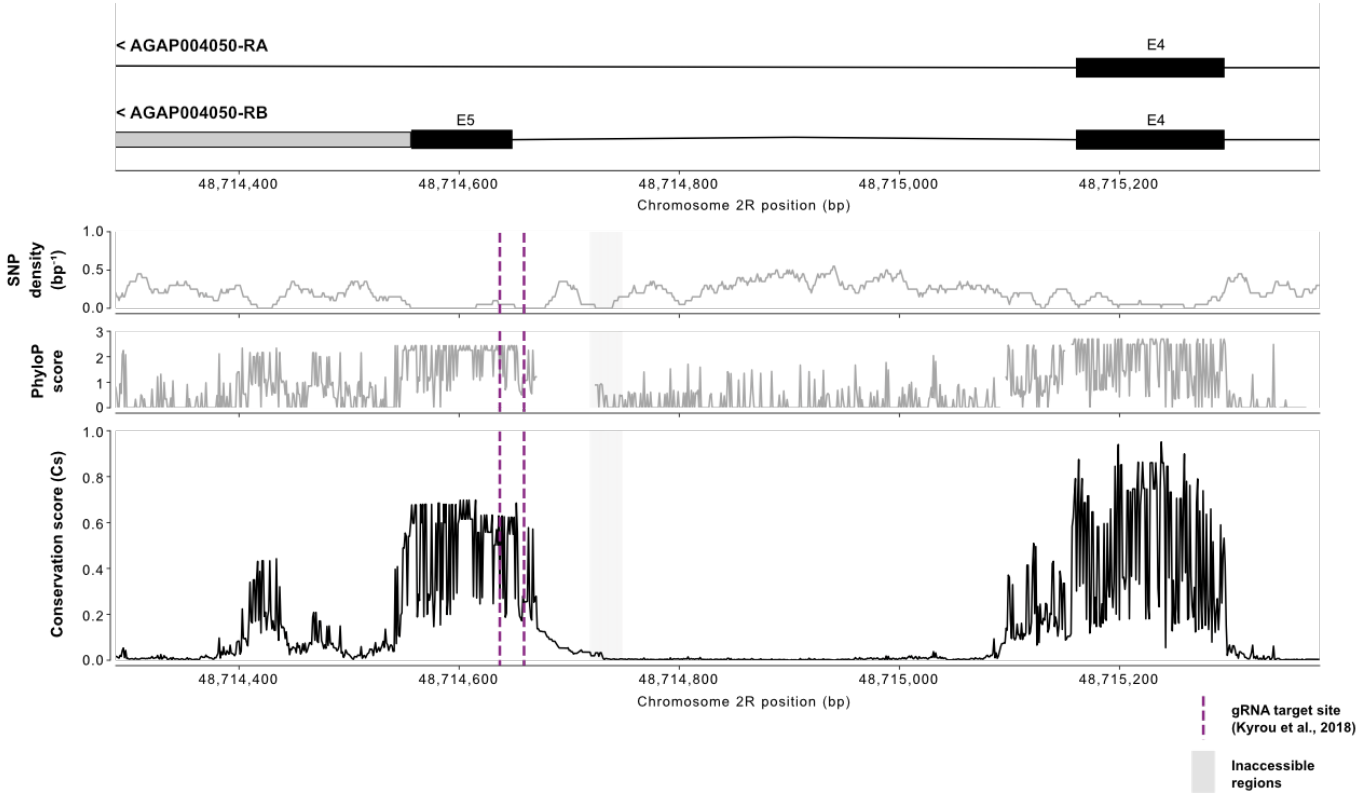


Figure 23 - SNP density, PhyloP and conservation score (Cs) at the *doublesex* target site. SNP density displayed in the second panel under gene structure was calculated by using sliding window of 20 bp. PhyloP score provides the metric for departure from the neutral rate of substitution and is used as a measure of conservation. Together with SNP density and normalised sequence identity they comprise conservation score (Cs) displayed in the bottom panel in black. High conservation score was observed at exon 4, which is shared between male (*AGAP004050-RA*) and female (*AGAP004050-RB*) isoform, and at exon 5, which is female-specific exon. Dashed purple lines represent start and end position of gRNA target site described in Kyrou et al. (2018). Grey box on the bottom three panels represents inaccessible region where quality of genotyping was shown to be low. (The Anopheles gambiae 1000 Genomes Consortium, 2017)

The target site locus also showed low SNP density of 0.03 bp^{-1} based on the data from 1142 wild-caught individual mosquitos (Figure 23). For comparison, mean SNP density in poorly conserved intron 4 was 0.25 bp^{-1}. The mean PhyloP score over the 23 bp of the target site was 1.69 with the two nucleotide positions adjacent to the cut site, scoring 2.44 and 2.16, which puts the two genomic positions in the 95th percentile of all PhyloP conservation scores, indicating high conservation.

High sequence identity across most of the genomes, low SNP density and high PhyloP score all amounted to high conservation score (Cs) at the *dsx* target site. The mean Cs at the target site was 0.43 (Cs_{min} = 0.17, Cs_{max} = 0.68), which puts the target site in the top 10th percentile of Cs from chromosome arm 2R (Figure 24). The mean Cs of the coding region of *dsx* exon 5 was 0.50, suggesting there could be potentially other alternative gene drive target sites with even higher conservation. Exon 4, which lies approximately 1 kb downstream of exon 5, and is shared between both male and female *dsx* isoforms showed even higher sequence conservation and resulted in high Cs, indicating high functional constraint of that particular exon.

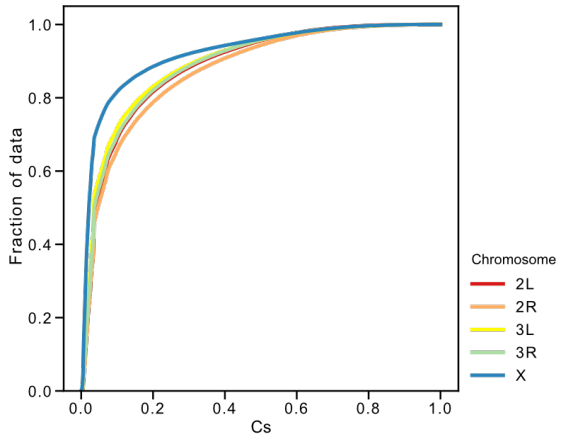


Figure 24 - CDF function for positive conservation score (Cs) values per chromosome arm.

Gene ontology (GO) enrichment analysis was performed to investigate in which cellular processes the most conserved genes are involved. Given a set of highly-conserved genes an enrichment analysis will find which GO terms are over-represented (or under-represented) using annotations that are attached to each gene in the set.

Highly-conserved genes were defined by calculating mean Cs in the exonic regions of genes. 653 genes that were in the 95th percentile of genes with the highest average Cs were submitted to the GO enrichment analysis for GO terms in molecular function and biological process (Figure 25). Genes that are essential for cadherin binding, rRNA binding and neurotransmitter binding activity were enriched the most, followed by genes performing many other molecular functions involved in transporter activity. Analysis of enriched terms for biological processes showed that the most conserved genes play essential roles in cell junction assembly, regulation of membrane potential and synaptic signaling.

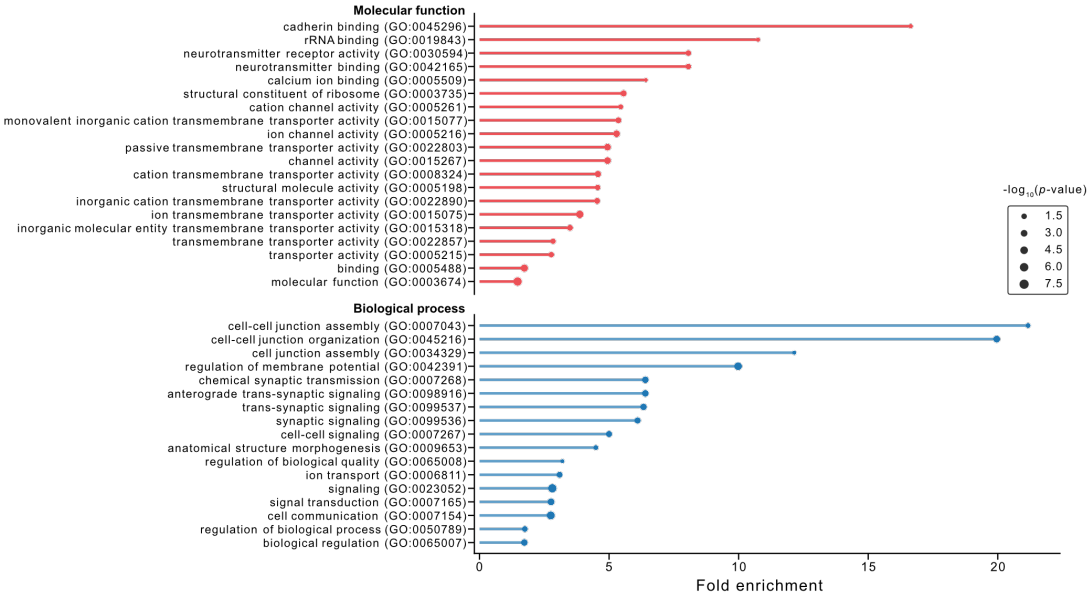


Figure 25 - Gene ontology enrichment for the most conserved genes in *An. gambiae* genome. GO enrichment was performed for the 95th percentile of genes with the highest average Cs. Enriched GO terms are divided into two categories: molecular function and biological process. Size of the dot for each enriched term represent statistical significance for the enrichment.

The top 10 genes with highest average Cs (Cs > 0.68) were all non-coding RNA. Six of them were tRNAs, namely: *tRNA-Arg*, *tRNA-Leu*, *tRNA-Ser*, *tRNA-Trp* and *tRNA-Val*. Two genes belonged to miRNAs: *iab-4* and *mir-184*. *iab-4* was shown to play a regulatory role in the expression of *ultrabithorax* (*ubx*) gene, which is a homeobox gene used in the regulation of patterning in morphogenesis (Ronshaugen et al., 2005). It was also previously shown that it's well conserved across several species of insects (Liu et al., 2017; Rodríguez-Sanchez et al., 2021). *mir-184* is well conserved between vertebrates and insects and is responsible for posttranscriptional regulation in germline and early embryonic development, including tissue fate establishment, differentiation or the maintenance of tissue identity (Aboobaker et al., 2005; Iovino et al., 2009; Li et al., 2011; Liu et al., 2017). Another ncRNA gene that was part of the top 10 most conserved genes in the genome was *U6*. *U6* is one of the five small nuclear RNA (snRNA) that are a part of

the spliceosome, a large macromolecular complex responsible for the splicing of introns. The components of spliceosome were shown to be well conserved across eukaryotes (Mount et al., 2007).

The most conserved coding gene was *AGAP001439*, which encodes heparan sulfate 6-O endosulfatase. Experiments in *Drosophila* showed that the ortholog of *AGAP001439* plays an important function in regulation of a morphogen gene *wingless*, which forms a gradient that helps to orchestrate adult wing development (Kleinschmit et al., 2010).

Essential roles that are associated with the most conserved genes that were identified based on high Cs confirm the suitability of Cs to be used as a proxy for functional constraint in the wider genomic context for both coding and non-coding genes.

4.2 Prediction of on-target site mutations

Double stranded breaks (DSB) induced by DNA editing with CRISPR-Cas9 were shown to produce different functional results at the cleavage site (Canver et al., 2015; Hammond et al., 2017; van Overbeek et al., 2016). For gene drive applications the most desired outcome is high frequency of HDR where the intact chromosome (with the gene drive construct) is used as a template to fix the cut. High frequency of HDR leads to high homing rate and introduces stronger bias in gene drive construct inheritance in progeny. In some cases, repair mechanisms other than HDR can be invoked that introduce heterogenous mutations in the form small insertions or deletions ('indels') at the target site, potentially causing gene drive resistance, as the target sequence cannot be cleaved by the Cas9-gRNA complex. Mutations generated at the target site can be the result of two end-joining repair mechanisms: NHEJ and MMEJ. Outcomes of both repair mechanisms (Figure 4) were shown to be dependent on the DNA sequence context around the cut site (van Overbeek et al., 2016). Since the MMEJ involves end-resection and the annealing of microhomologous sequences adjacent to the cut site, the resulting deletion can be "mechanistically" predicted *in silico* by always following the same rules that apply for MMEJ derived deletions (Bae et al., 2014a; Grajcarek et al., 2019). To estimate the propensity for a target site to generate functional resistant alleles through a repair mechanism is crucial in the gRNA designing process since, if they can be generated, their rate of generation through repair processes downstream of Cas9-induced cleavage is likely to far higher than the rate at which they arise through spontaneous mutation or exist as standing variation (Hammond et al., 2017).

4.2.1 Simulating MMEJ deletions

A precise position of the cut site can be determined for any given gRNA target site, 3 bp upstream from the first PAM position when Cas9 is utilised. To simulate MMEJ outcomes for a particular cut site the flanking sequences on both sides of the cut site need to be inspected for common sub-sequences that could be used as microhomologies (MH) in MMEJ repair.

The implemented algorithm to predict MMEJ deletions in this study (Figure 26) extracts flanking sequences of any length on either side of the cut site. Default flanking length of 70 bp was selected to support deletion sizes that are most frequently occurring upon Cas9-induced cleavage (<50 bp) (Lieber, 2010; van Overbeek et al., 2016). Upstream flanking sequence (x flank) is used as a template to search for MH. The scanning process starts at the beginning of flank x where a scanning window of 2 bp captures the first two nucleotides in the sequence and then searches for matching 2 bp sequences in the flank y. Positions of matching sequences are stored as pair of tuples; x(i, j) and y(i, j), where x and y are the upstream and downstream sequences from the cut site, respectively. Start and end position of the matching MH in flank x or y are denoted with i and j, respectively.

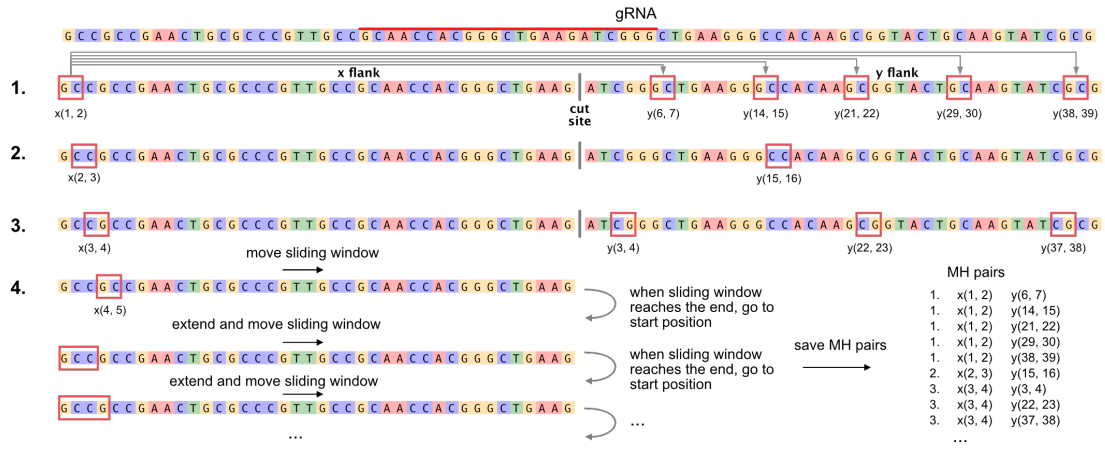


Figure 26 - Graphical representation of *in silico* MMEJ prediction algorithm. The cut position is first determined for a gRNA target site and the flanking regions are obtained (x flank, y flank). The first iteration (1) starts at the start of x flank where a sequence of 2 bp window is used to search for homologous sequences in flank y. Pairs of microhomology (MH) are stored in a temporary memory. In the next step of the iteration the 2 bp window is moved for 1 bp (2, 3) and the process of MH search is repeated. When the 2 bp window reaches the end of the x flank (4) it is returned to the beginning of the sequence and the window is extended for 1 bp. The process gets repeated until all subsequences of x flank are tested for MH in flank y.

Upon saving all pairs of matching MH for the first 2 bp of x flank, the scanning window is moved by 1 bp towards the cut site - x(i+1, j+1) and the process of matching and saving pairs of MH is repeated. When the scanning window reaches the end of the x flank the iteration restarts with extended scanning window of 3 bp. The scanning stops when all sub-sequences longer than 2 bp from x flank are exhausted.

Each saved MH pair could theoretically be employed in MMEJ and generate a deletion, which covers a sequence between the two pairs of MH, with one MH included in the deletion. The deletion length determines whether such a deletion could lead to translational frameshift, which in effect reduces the likelihood of generating a functional restorative resistant allele at gene drive target sites. Nonetheless, the previous experiments showed that among hundreds of

theoretically possible MH pairs at a target site much fewer are preferentially used in MMEJ. MH selection in the repair mechanism was shown to depend on the MH length, GC-content of the MH sequence and the length of the deletion. The GC content of microhomologous sequences might determine the repair pathway with a bias for MMEJ due to stronger binding in the annealing step of the MMEJ. Mismatches within the MH or low GC content were shown to reduce the rate of MMEJ, which indicates that MMEJ is associated with the thermostability of base pairing between the MH (Bae et al., 2014a; Sfeir and Symington, 2015; Sharma et al., 2015).

For these reasons a scoring system was proposed by Bae et al. (2014) based on the previously mentioned MH parameters.

MH score (MHs) is defined as the following equation (eq 4.3) where S and Δ represent microhomology index and deletion length in bp, respectively:

$$MHs = S \times e^{-\frac{\Delta}{20}} \qquad 4.3$$

Microhomology index S (eq 4.4) accounts for the length of MH pattern, l_{MH}, and the number of stronger G:C base pairs in the MH pattern, GC_{MH}.

$$S = l_{MH} + GC_{MH} \qquad 4.4$$

Examples of predicted MMEJ deletions and the elationship between MHs, deletion length, MH lenth and GC content is visually represented later in Figure 31.

4.2.2 Testing the MMEJ outcomes

MH scores were found to have positive correlation with MMEJ rate in HeLa cells (Bae et al., 2014a), which led us to set up an experiment to test whether the predicted MMEJ deletions and their MH scores correlate with observed mutation frequencies *in vivo* in *An. gambiae*. The d*oublesex* gene was arbitrarily selected as the locus to search for four target sites that had different MMEJ properties. The intention was to test the overall frequency of MMEJ-associated deletions on different target sites based on the sum score of 5 most highly ranked MH patterns. Additionally, the relative frequency between the first two highly scored MMEJ deletions was tested to see whether the difference matches the difference in MH scores.

Ideally, for gene drive purposes, a gRNA target site should be chosen where frequency of MMEJ deletions is high (high MHs_{sum}) relative to unpredictable NHEJ mutations and the MMEJ deletions at the target site produce frame shift. If the individual MHs also correlates strongly with the frequency of the MMEJ deletion, then ideally a gRNA target site should be selected where the highest-scoring MMEJ deletion represents the majority of MMEJ events at the target site (high MHs_{sum} and large MH_1-MH_2) and also causes frame shift.

For the purpose of analysing the observed MMEJ deletion frequency and their MHs, four different target sites were selected based on the MH_{sum} (sum of top 5 MH scores) and the difference between top two MH scores ($MH_1 - MH_2$) (Figure 27).

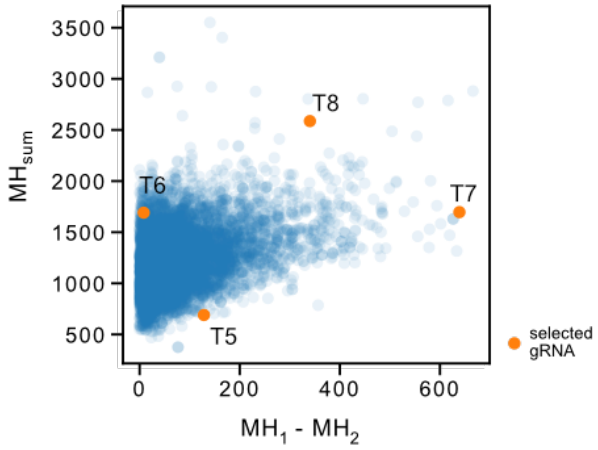


Figure 27 - Selection of gRNAs for *in vivo* testing based on their MMEJ properties. 4 gRNAs were selected in *doublesex* gene to be tested *in vivo* based on their MH_{sum} score and the difference between two highest scoring MH patterns (in orange). MH_{sum} (on Y-axis) is the sum of MH scores from the five highest scoring MH patterns predicted *in silico*. The difference between the first and the second MH score for each gRNA is showed on X-axis (MH_1 - MH_2). The selection was made to include two gRNAs with similar MH_{sum} score and large difference between MH_1 and MH_2 (T6 and T7) and two gRNAs with dissimilar MH_{sum} score and small difference between MH_1 and MH_2 (T5 and T8).

There were 8825 gRNA target sites found in *dsx* gene. Two selected targets, T6 and T7 had comparable MHs_{sum} values and large difference between MHs_1 and MHs_2. The other two target sites, T5 and T8, were selected to have smaller difference between their top two MH scores and larger difference in MHs_{sum}. Plasmids bearing gRNA targeting T5, T6, T7 and T8 were created and microinjected in *An. gambiae* embryos to obtain transgenic lines used for the experiment.

Three gRNA strains expressing a gRNA (T5, T6 or T7) under the constitutive U6 promoter and marked by 3xP3::dsRED, were obtained by random piggyBac integration, as previously outlined (Nolan et al., 2002). The T8 strain could not be obtained due to high larval mortality post-injection. Transgenic males of the T5, T6 and T7 strains were crossed in triplicate (50*3 males per strain) to females expressing vas2::Cas9, marked by 3xP3::YFP (Dong et al., 2018; Hammond et al.,

2016). Their larval progeny was pooled and had gDNA extracted *en masse* and using pooled amplicon sequencing to determine the identity of end-joining (EJ) mutations generated at each target site (Figure 28).

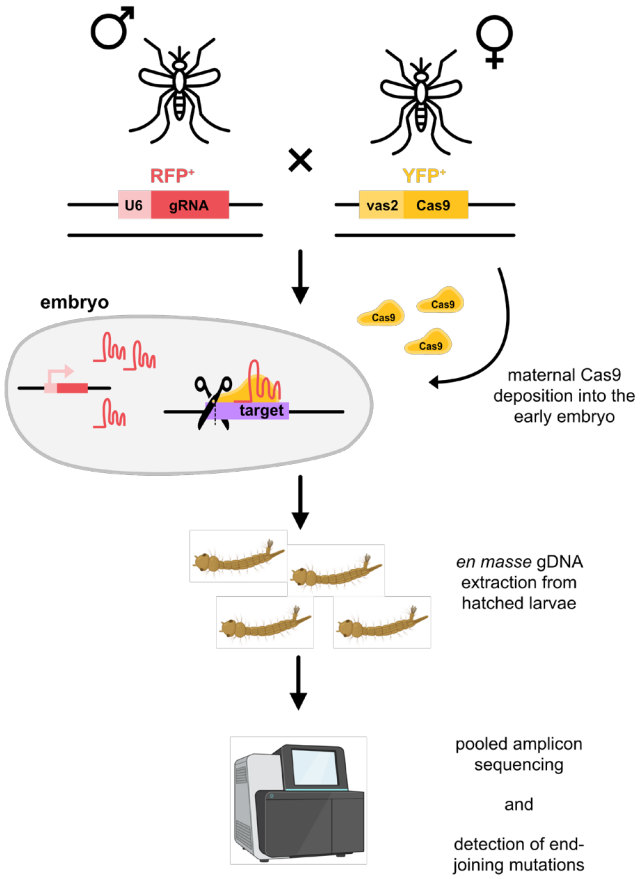


Figure 28 - Genetic crosses for *in vivo* MMEJ experiment. Male mosquitoes carrying an RFP+ marked transgene expressing a gRNA targeting the site of interest (T5, T6 or T7), under the control of the U6 promoter, were crossed to females carrying a YFP+ marked transgene, expressing Cas9 under the control of the *vas2* promoter that is known to induce maternal Cas9 deposition into the early embryo (Hammond et al., 2021a). Due to maternal Cas9 deposition into the embryo, in the presence of the RFP+ marked U6::gRNA, the gRNA and Cas9 combine and cleave the desirable target site on *doublesex* (T5, T6 or T7), producing a multitude of end-joining (EJ) mutations. After the embryos hatched, L1 larvae were pooled together, their gDNA was extracted *en masse* and using pooled amplicon sequencing the identity of EJ mutations generated at each target site were revealed, to determine whether they are influenced by nearby microhomologies.

Three biological replicates were performed for each cross, with three technical replicates in the gDNA extraction, amounting to nine samples for each target site.

Amplicon sequencing results showed abundant EJ mutations for the target sites T6 and T7, while T5 target site revealed only 0.004% of all reads at the target site in average with EJ mutations, of which only 0.002% reads contained MMEJ-associated deletions (Figure 29). Due to presumable low Cas9-gRNA combination activity in the embryo, which was indicated with very low rate of observed EJ mutations (both NHEJ and MMEJ deletions), T5 samples were omitted from further analysis. In comparison, in average 17.3% and 16.9% of all reads had MMEJ deletions in T6 and T7 samples, respectively. The overall frequency of MMEJ deletions was significantly higher in both T6 and T7 sample compared with the frequency of NHEJ deletions (independent t-test, $p<0.05$).

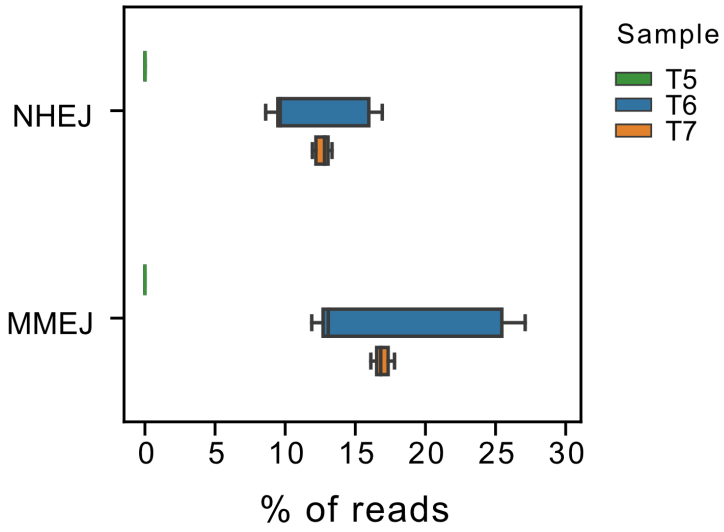


Figure 29 - Abundance of NHEJ and MMEJ deletions at the target sites T5, T6 and T7. All replicates of T5 samples showed negligible levels of NHEJ and MMEJ (< 0.01% or reads containing deletions). Frequency of NHEJ deletions was significantly lower in T6 and T7 samples (independent t-test, $p < 0.05$). In average 17.3% reads in T6 had MMEJ deletion and 16.9% in T7.

Closer inspection at the target site showed a wide range of EJ mutations, 37 unique deletions were generated at the T6 and 39 at the target site T7. Each target site had one predominantly abundant MMEJ deletion (Figure 30). A 4-bp deletion (101-TGCA-105) at the target site T6 and a 12-bp deletion (204-ATCGGGCTGAAG-216) at the target site T7 represented the large majority of EJ mutations. The 101-TGCA-105 deletion could be attributed to MMEJ by the presence of 3 bp long MH sequences on both sides of the cut site. It was present in all replicates with the mean frequency of 15.72%. Deletion 204-ATCGGGCTGAAG-216 was observed with the mean frequency of 16.0% and was predicted to be generated from the top scoring 10 bp long MH pattern at the target T7.

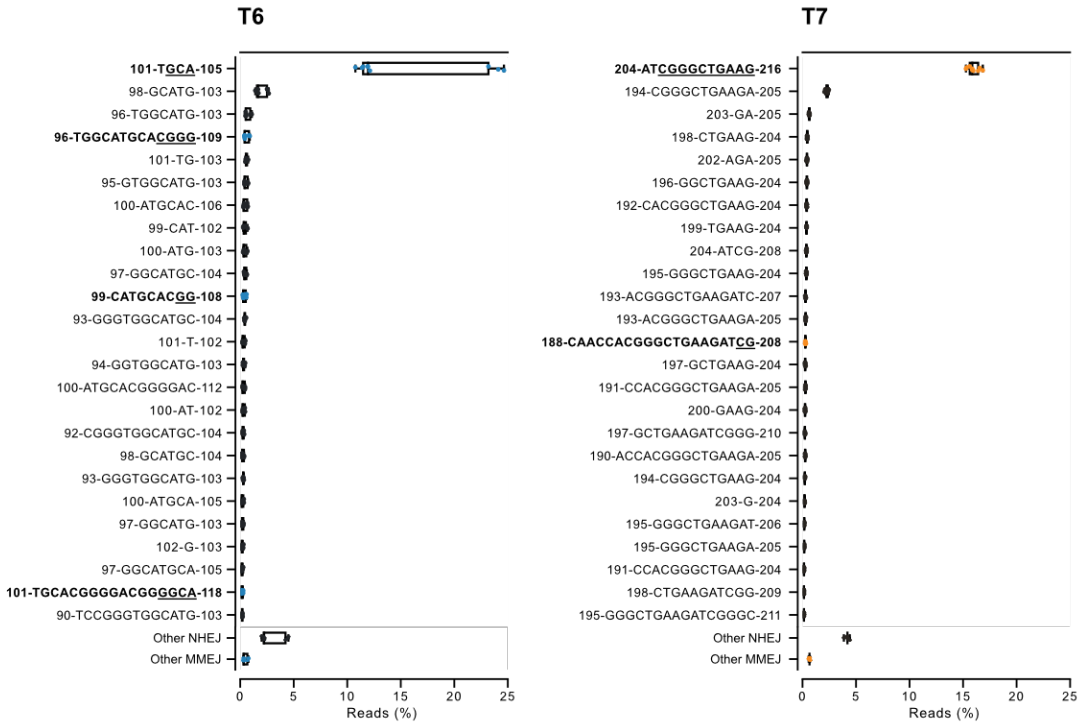


Figure 30 - EJ deletions observed at the target site T6 and T7. 25 most frequent deletions observed at the target sites T6 and T7 are displayed on Y-axis by their deletion tag, which consists of start position, end position and the deleted sequence. MMEJ-associated deletions are written in bold, and their MH sequence is underlined. MMEJ frequencies are marked with colour points (blue for T6 and orange for T7). A 4-bp long MMEJ deletion 101-TGCA-105 was found to be the most abundant mutation at the target site T6 with mean frequency of 15.72%. MMEJ at the target site T7 yielded the most abundant deletion 204-ATCGGGCTGAAG-216 generated after resection and annealing of 10 bp long MH. Its mean frequency was 16.0% and was observed in all replicates.

All other EJ mutations observed at the target sites were present in considerably lower frequencies (<5%). The second most abundant MMEJ deletion at the target site T6 was 96-TGGCATGCACGGG-109 with mean frequency of 0.56% and 20 bp deletion 188-CAACCACGGGCTGAAGATCG-208 with the mean frequency of 0.27% at the target site T7.

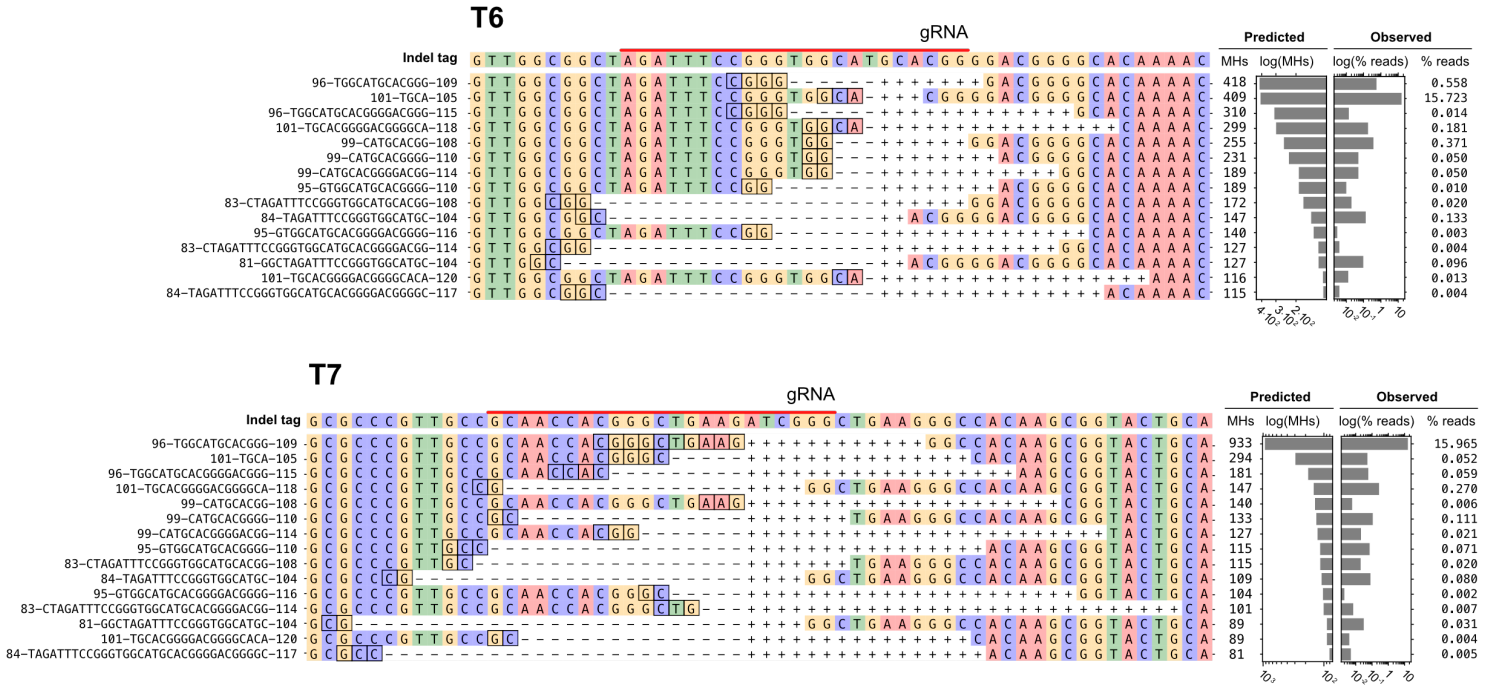


Figure 31 - Predicted MMEJ-associated deletions, their MH scores and observed frequency at the target sites T6 and T7. 15 highest scoring MMEJ deletions that were predicted *in silico* are shown with their MH scores (MHs), mean observed frequency (in percentage of reads containing the deletion), mean proportion of MMEJ deletions (in percentage) and their associated deletion tag with the deletion start position, end position and deletion sequence with MH pattern underlined. MMEJ deletions are presented in the sequence alignment around the gRNA (in red) where "-" stands for deleted nucleotides in the left flanking region from the cut site and "+" stands for deleted nucleotides in the right flanking region. MH sequence is marked with black border on the left side of the deletion. The highest-scoring MMEJ deletion 204-ATCGGGCTGAAG-216 at target site T7 was also the most abundant detected deletion (16%). This was not the case at the target site T6 where second highest scoring deletion 101-TGCA-105 was present in the highest frequency (15.7%).

The target site T6 was selected to have the difference in calculated MHs_1 and MHs_2 minimal (ΔMHs = 9), however the observed MMEJ-associated deletions for MH_1 (96-TGGCATGCACGGG-109) and MH_2 (101-TGCA-105) on that site had unexpectedly large difference between their observed mean frequency (15.2 p.p). In comparison, the difference between the two most abundant MMEJ-associated deletions in T7 in score was ΔMHs = 786 and 15.7 p.p. in their observed mean frequency (Figure 31). The differences in frequency between the two most abundant MMEJ deletions at either target site were not significant (independent t-test, p = 0.795), unlike their MH scores would suggest.

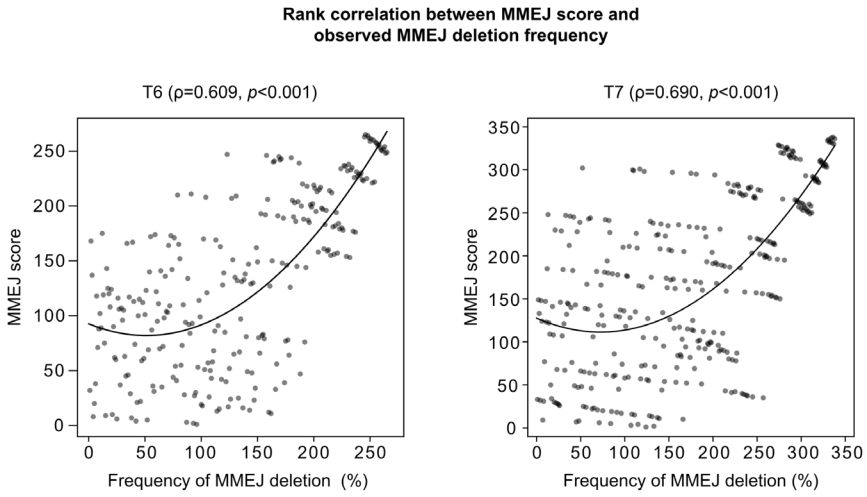


Figure 32 - Rank correlation between predicted MMEJ score and observed MMEJ deletion frequency for target sites T6 and T7. Strong Spearman correlation was observed for T6 (ρ = 0.609, p<0.001) and T7 (ρ = 0.690, p<0.001). The most abundant eight MMEJ deletions show particularly high rank correlation with their predicted scores and are consistent across all replicates.

The frequency of MMEJ deletions did not always correspond to their ranking based on MH score, as seen from the results. Nonetheless, a strong Spearman correlation between MH scores and their associated deletion frequencies was calculated for both target sites T6 (ρ = 0.609, p<0.001) and T7 (ρ = 0.690, p<0.001) (Figure 32).

Unfortunately, no data was obtained for target sites T5 and T8, which would give even more power in the comparison of overall MMEJ rate to NHEJ rate and the distribution of MMEJ-associated deletions.

4.3 Target site selection in the genes of interest

Current genome editing protocols recommend that researchers select guides carefully to minimise potential endonuclease activity on nonspecific target sites that can cause unintended genetic modifications. These nonspecific target sites are called off-targets (Ran et al., 2013). A large number of computational tools that support target selection (or gRNA design) for CRISPR-based genome editing exist. (Bradford and Perrin, 2019; Liu et al., 2020). Several bioinformatic tools were produced to help determine which gRNAs are the most efficient for general CRISPR-Cas genome editing purposes (Concordet and Haeussler, 2018; Labun et al., 2016; Park et al., 2015; Prykhozhij et al., 2015).

gRNA design tools first find potential target sites in a genome or a DNA sequence by locating the protospacer adjacent motif (PAM) which is a 3-6 bp sequence, specific to a Cas9 endonuclease, needed for initiation of DNA and gRNA-Cas9 binding. The presence (or absence) of a PAM sequence in a particular locus that is being targeted (i.e. sequence encoding a functional domain of a protein or a transcript splicing binding site) can play a major role in obtaining desired effects of the genome editing. For that particular purpose novel Cas9 endonucleases have been engineered and now offer researchers unprecedented flexibility in targeting almost any desired sequence (Walton et al., 2020). The flexibility to search for the presence of any PAM sequence in a genomic context is needed in a gRNA design tool.

Different scores and information are usually assigned to each gRNA that help researchers decide which gRNA in a given locus is most suitable for their application. Scores can be separated on off-target scores and on-target scores. Off-target scores try to estimate the off-target cleavage potential which depends on the similarity between the on-target sequence and predicted off-target sequences.

Predictions are first calculated on the level of individual putative off-target sites using the cutting frequency determination (CFD) off-target score (Doench et al., 2016). An off-target site with a high off-target score is more likely to be cleaved than one with a low score. The scores of all off-target sites of a guide are then summarised into the 'guide specificity score'. The other type of prediction is on-target scoring, which estimates the on-target cleavage efficiency and depends only on the target sequence (Canver et al., 2018).

Additional scoring systems were developed to account for end-joining mutations on the target site. As error-prone non-homologous end joining at the target site gives rise to small insertions and deletions and often causes frame shift mutations, a score was proposed to estimate MMEJ events at a target site. It ranks the most likely MMEJ patterns, giving the information about potential MMEJ events that could disrupt the ORF (Bae et al., 2014a). Other machine learning based predictive models were also developed to estimate the genome editing repair outcomes based on observed mutations in human and mouse cell lines, namely Lindel and InDelphi (Chen et al., 2019; Shen et al., 2018).

The relative importance of these scores depends on the context of the genome editing application. For example, some clinical applications might require high specificity, while others might require a target site that will produce favourable end-joining mutations, hence more focus will be put on MMEJ score of the target site.

Some tools have become the "industry standard" as they cater the needs for most of the researchers by including many of the essential and non-essential metrics packed in an intuitive user interface (Concordet and Haeussler, 2018; Labun et al., 2019, 2016). Nevertheless, one branch of the genome editing applications with its own

unique set of considerations in terms of target site selection are also CRISPR-based gene drives.

As previously mentioned, one of the essential parts of devising a gene drive application is to use rational approach to target selection that considers the potential for a gene drive target site to generate functional resistant alleles that could hinder the spread of a gene drive in a population and to reduce the likelihood of that happening.

Furthermore, to address the issue of resistance, a gRNA should target a functionally constrained site, where mutations are more likely to be deleterious and hence less prone to undergo selection. To identify such functionally constrained genomic regions sequence conservation could be used as proxy. *An. gambiae* genome was shown to have several ultra-conserved regions with minimal variation within the species and across other mosquito species, however no easily accessible genome-wide data existed that could evaluate a level of sequence conservation on each position of the *An. gambiae* genome (O'Loughlin et al., 2021; The Anopheles gambiae 1000 Genomes Consortium, 2020). Equally important is also consideration of any genetic variation at the target site locus that might be present in the local environment where gene drives will be released but is not present in the model reference genome, that is used for target selection. Surveys that sample local mosquito population and investigate their genetic variation are crucial in supplying information, not only to determine highly conserved genomic regions, or to account for resistant alleles already present in the wild, but also to design gene drives that can exploit pre-existing sequence differences between target and non-target populations (Willis and Burt, 2021).

Previous studies also showed that EJ mutations resulting from Cas9 endonuclease-induced DSB can be detrimental to successful spread of CRISPR-based gene drives (Champer et al., 2017; Hammond et al., 2017, 2016). Gene drive resistance caused by EJ mutations can, in part, be mitigated by selecting a target site, where the sequence context around the cleavage site will preferentially generate out-of-frame EJ deletions. MMEJ repair mechanism that is employed after DSB occurs can utilise microhomologies on each side of the cut site that generate predictable deletions. Effort should therefore be put towards selecting a target site with higher propensity for out-of-frame deletions, which are in effect non-functional and deleterious mutations, to avoid generating functional resistance.

In light of this, a novel computational tool was created that incorporates all forementioned aspects of gRNA design. It was employed in gene drive target site selection in female-fertility genes and sex-determination genes of *An. gambiae* that could potentially be used as refined or novel gene drive applications.

4.3.1 Guido - *in silico* gRNA designer

Guido was developed as a command-line interface (CLI) tool that can search for gRNA targets in any reference genome or DNA sequence. It integrates MMEJ prediction and scoring, off-target search, conservation score and fetches the variation data from Ag1000g project (The Anopheles gambiae 1000 Genomes Consortium, 2021).

The initial step is to build the necessary reference files with *guido-build* for the main program to use them as the sequence where available gRNA targets are searched in. *guido-build* takes a FASTA file as a sequence input and optionally a GTF or GFF3 genome annotation file. It creates a sequence index files, which allow Guido to

operate faster within a specific region of the genome and enable the off-target search across the genome.

The output file that *guido-build* creates contains the genome sequences and their associated index files, which are used to perform the search for gRNA target sites by the main program *guido*. This process offers the flexibility to generate all needed files for any DNA sequence researchers want to perform the search on, independently from the main program.

Guido requires a specified genomic region which is used to perform the gRNA targets search in. Alternatively, a gene name can be provided, which is then linked back to the genome annotation to determine the genomic location of the gene. The search can be restricted to only specific genomic feature, such as exons, introns, 5'-UTR and 3'-UTR regions and coding sequences (CDS). To find potential gRNA target site in the genomic region of interest the sequence is searched for the presence of PAMs on both strands. Allowing researchers to specify the PAM sequence extends the flexibility of using Guido with any type of endonucleases. Cut sites are then determined for each potential gRNA target site based on their PAM positions.

The search for target sites is followed by a step where MMEJ simulation is performed for each gRNA target site. MMEJ patterns, their scores, deletion sequences and information whether the frame shift is expected for a given MMEJ deletion are stored for each target site. Furthermore, a mean conservation score is calculated along the 23 bp of the gRNA. Despite the conservation score described in Chapter 4.1 was calculated only for *An. gambiae* AgamP4 reference genome, any numerical array of the size matching the genome length, on which the search is performed, can be used in place.

Another search for off-targets is performed for each gRNA sequence by using Bowtie search algorithm (Langmead, 2010). Bowtie allows searching for homologous sequences with up to three mismatches in the reference genome. The number of off-targets and the number of mismatches are scored to give a higher score penalty to off-targets with less mismatches compared to off-targets with more mismatches. In this way the target sites with higher likelihood to generate off-targets are scored lower.

In silico prediction of off-targets with Bowtie was compared with the results obtained in recent study of *in vitro* and *in vivo* off-target activity of gene drives targeting *dsx* (*AGAP004050*) and female fertility gene *AGAP007280*. (Garrood et al., 2021) Off-targets predicted by *in silico* tool Cas-OFFinder were added for comparison. Cas-OFFinder, unlike Bowtie, is not limited by the number of mismatches and allows variations in PAM sequences recognized by Cas9. (Bae et al., 2014b) Results in Table 3 show that only a fraction of *in silico* predicted off-targets with Guido or Cas-OFFinder can be found *in vivo* for *AGAP007280* target site and none for *dsx*. Despite likely over-representation of *in silico* off-targets, compared to their detected *in vivo* activity, selecting for gRNA with minimal number of predicted off-targets makes this step suitable and necessary in the gRNA design process.

Table 3 - Number of off-target sites predicted *in silico* and detected *in vitro* and *in vivo* at the gene drive target sites in *AGAP004050* and *AGAP007280*. Off-target analysis *in vitro* and *in vivo* performed by Garrood et. al. (2021) showed minimal off-target activity *in vivo*. Guido was able to predict only a fraction of off-targets compared to Cas-OFFinder, which is not limited to reporting off-target sites with only 3 mismatches to gRNA sequence like Guido. Only one *in vivo* off-target site for *AGAP007280* with 4 mismatches to gRNA was failed to be predicted by Guido.

			Canonical PAM (NGG)							
			Off-target sites with mismatches to gRNA							
Gene	**Method**	**Context**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
doublesex AGAP004050	Guido	*in silico*	0	0	0	0	/	/	/	/
	Cas-OFFinder	*in silico*	0	0	0	0	9	127	1142	7740
	CIRCLE-seq	*in vitro*	0	0	0	0	1	1	7	19
	in vivo		0	0	0	0	0	0	0	0
AGAP007280	Guido	*in silico*	0	0	2	8	/	/	/	/
	Cas-OFFinder	*in silico*	0	0	10	115	1139	8377	44763	190965
	CIRCLE-seq	*in vitro*	0	0	3	16	30	7	2	3
	in vivo		0	0	2	2	1	0	0	0

A scoring system was implemented in Guido that would penalise *in silico* predicted off-target sites with higher sequence identity to the target site as described in eq 4.5. A coefficient 5, 3, 2 and 1 were given to the number of off-target sites (n) with 0, 1, 2 and 3 mismatches to gRNA, respectively.

$$OT_{sum} = 5n_0 + 3n_1 + 2n_2 + n_3 \quad 4.5$$

One of the optional inputs in Guido allows importing a variation dataset in VCF, zarr or HDF5 formats. SNPs are fetched from the dataset based on the genomic location of the gRNA. Specifically for the purpose of target site selection in *An. gambiae* the variation dataset generated in Ag1000g Project phase 3 (Ag3) from 2,784 wild-caught mosquitoes collected from 19 countries in sub-Saharan Africa was integrated in Guido (The Anopheles gambiae 1000 Genomes Consortium, 2021). Guido fetches SNPs overlapping the genomic region of each potential gRNA and calculates the

overall frequency of SNPs for each position of the gRNA. Samples (individual mosquitoes) with SNPs that share the same position are also grouped by country where they were sampled from and by the species that they belong to (the dataset includes *An. gambiae*, *An. coluzzii* and *An. arabiensis* individuals). This feature allows researchers to account for pre-existing genetic variation at the target site that could impact gRNA-Cas9 binding to the target site as it was shown that any mismatches in the seed region of gRNA (10-12 bp from PAM) substantially reduce the cleavage activity (Jiang et al., 2015; Wu et al., 2014; Xu et al., 2017). Alternatively, SNPs that are present at the target site could be incorporated in the gRNA sequence if found in high frequency or fixed in the local population.

Finally, the gRNAs are ranked based the criteria deriving from predicted MMEJ outcomes, conservation score, predicted off-targets and reported variation at the gRNA target site. Relative weights were used in ranking to give more weight to gRNA with higher conservation score and predicted MMEJ out-of-frame deletions (Table 4).

Ranking is performed on two Multi Criteria Decision Making (MCDM) methods. MCDM refers to methods of decision making in a scenario where there are often multiple criteria or attributes that must be considered (Zavadskas et al., 2014). The ranking model implemented is the Technique for Order of Preference by Similarity to Ideal Solution - TOPSIS (Hwang et al., 1993). The main idea behind TOPSIS is to find the compromise solution, which is the closest to the ideal solution and the furthest from the worst solution based on the values of all criteria. TOPSIS algorithm firstly finds the ideal (best) gRNA alternative which maximizes benefit attributes (the best conservation score, the best MMEJ score, ...) and minimizes cost attributes (the number of off-targets, the number of variants in the PAM sequence, ...) and secondly finds the worst gRNA alternative which minimizes benefit attributes and

maximizes cost attributes. Final ranking is obtained based on the geometric distance from the ideal gRNA and the worst gRNA, where highest ranked gRNA has the shortest geometric distance from ideal gRNA and the longest geometric distance from the worst gRNA. Criteria and their relative weights listed in Table 4 were used for ranking.

Table 4 - List of criteria used in gRNA ranking. Ranking was employed based on criteria that included obtained MMEJ data (MHs_{sum}, MHs_{oof}), conservation score (Cs), predicted off-targets (OT_{sum}), and variation dataset (AC_{sum}, SNPs, AC_{pam}, AC_{seed}, AC_{small_seed}). A relative weight was assigned to each criterion to highlight the relative importance. Relative importance was determined arbitrarily, but with known gene drive limitations in mind, such as: decreased binding efficiency when mismatches are present in the seed sequence of gRNA, the potential for the rise of resistant alleles if the locus is not conserved or if the repair outcomes are not favourable. Attribute type, which was either benefit or cost, was accounted for when ranking algorithm searched for the compromise solution that has high values of benefit criteria and low values of cost criteria.

Criteria	Relative weight	Attribute type	Description
MHs_{sum}	0.5	Benefit	MHs_{sum} - Sum of MHs for five highest scoring MMEJ deletions at the target site
MHs_{oof}	0.8	Benefit	MHs_{oof} - Percentage of MH scores of out-of-frame MMEJ deletions (MHs_{oof} / MHs_{sum})
Cs	0.8	Benefit	Cs - Mean conservation score in the gRNA region
OT_{sum}	0.3	Cost	Off targets score - Mismatch-based score of off-target counts
AC_{sum}	0.3	Cost	SNP allele count - Sum of SNP allele counts in the gRNA region
SNPs	0.3	Cost	SNP count - Count of SNPs in the gRNA region
AC_{pam}	0.8	Cost	PAM SNP allele count - Sum of SNP allele counts on non-variable PAM positions (i.e. second and third position of NGG PAM)
AC_{seed}	0.4	Cost	Seed region SNP allele count - Sum of SNP allele counts in the 10 bp proximal to PAM
AC_{small_seed}	0.6	Cost	Shorter seed region SNP alleles - Sum of SNP allele counts in the 4 bp proximal to PAM

Guido outputs ranked gRNAs in a concise tabular text file accompanied with separate files with detailed report on gRNAs variation and their predicted MMEJ outcomes with the sequence around gRNA. If the genome annotation is provided, each gRNA will have information about genomic features overlapping the gRNA. Furthermore, genome accessibility annotation was provided for each gRNA. Genome accessibility dataset was generated in the Ag1000g project and highlights genomic regions that were annotated and masked as low-complexity genomic regions already at the stage of reference genome assembly and showed low coverage and poor mapping quality when performing the alignment of individual resequencing data for the project (Miles et al., 2017; Sharakhova et al., 2007). gRNAs from such masked regions should be considered cautiously due their high repeat content.

4.3.2 Gene drive target sites in the female-fertility genes and sex-determination pathway

Three female-fertility genes were proposed by Hammond et al. (2016) to be used for gene drive for population suppression: *AGAP007280*, *AGAP005958* and *AGAP011377*. They confer a recessive female-sterility phenotype upon disruption, which makes them suitable target genes for gene drive application. Additionally, a female-specific isoform of *AGAP004050* (*doublesex*) was selected as a gene drive target site by Kyrou et al. (2018).

Guido was used in attempt to find alternative target sites within the proposed genes and compare them with previously selected gRNAs. Only gRNAs with that are

accessible and have their cleavage site within the boundaries of exons were considered.

4.3.2.1 AGAP007280

AGAP007280 is an ortholog of a *Drosophila* gene called *nudel*, a gene expressed early in oogenesis in ovarian follicle cells and is maternally required for the development of dorsoventral axis in the embryo. The protein contains a protease domain that has an essential role in defining the dorsoventral polarity. It was shown that homozygous mutations affect the embryo viability or impair the structural integrity of the laid eggs, however heterozygous mutants did not show either of the two phenotypes (Hong and Hashimoto, 1996). *AGAP007280* was used as a gene drive target site but the gene drive did not spread in the population due to the emergence of functional resistant alleles (Hammond et al., 2017, 2016).

Guido reported 987 gRNA target sites in *AGAP007280*, out of which 182 target sites did not have any SNP present (based on the Ag1000g phase 3 dataset) in the PAM region of the gRNA. Five highest ranking gRNAs were found in exons 4, 6 and 10. (Figure 33) The highest-ranking gRNA-526 conservation score was Cs = 0.32, considerably higher than the target site targeted, here named, gRNA-579 selected by Hammond et al. (2016). In comparison, all 5 highest-scoring MH deletions at the target site gRNA-579 were predicted to produce in-frame deletions, as opposed to gRNA-526, where 4 out of 5 predicted MMEJ-associated deletion produce frame shift, giving the gRNA-526 a MHs_{oof} score of 84.7 (Table 5).

Analysis of genetic variation in the region of candidate gRNA-526 showed that a SNP was present in the PAM region at the target site, however only 3 alleles were sampled with this variant, one in Burkina Faso populations and two in Gambia populations, all present in *An. coluzzii*. This makes overall frequency of the

alternative allele in the SNP region to be under 0.05%. SNPs in the seed region of gRNA that could potentially decrease the binding efficiency of gRNA to the target site were also found in low frequency, overall under 0.5% and under 1% in local populations with one exception of G>A SNP that is present in 1.3% of all alleles in Burkina Faso, and shared between *An. coluzzii* and *An. gambiae* (Figure 34).

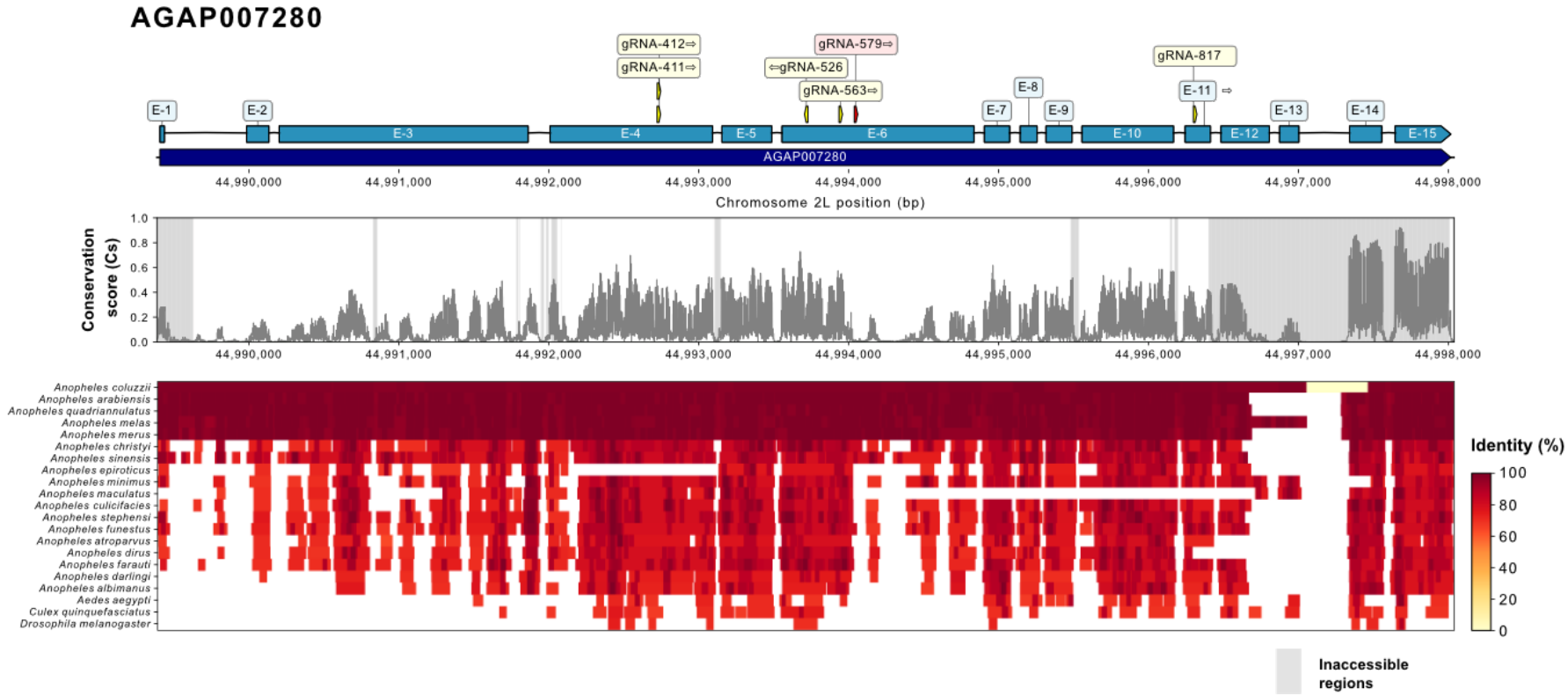


Figure 33 - Candidate gRNA target sites for *AGAP007280. AGAP007280* gene with 5 highest-ranking gRNAs in yellow and gRNA used by Hammond et al. (2016) in red. Conservation score is displayed on the middle panel in dark grey with inaccessible regions highlighted in light grey. Sequence identity (in percentages) is represented with heatmap for each species genome that was analysed for the calculation of conservation score.

Table 5 - Candidate gRNAs for *AGAP007280*. Genomic location and strand, gRNA sequence, the criteria used in ranking (Table 4), out-of-frame status (OOF), predicted off-targets (OT) and the final ranking are presented in the table. OOF indicates ordered 5 highest scoring MH deletions and their out-of-frame status denoted as + if deletion is out-of-frame or - if deletion is in-frame. OT denotes the number of off-targets with 0, 1, 2, 3 mismatches, respectively. gRNA used by Hammond et al. (2016) is highlighted in red.

gRNA	Genomic location and strand				gRNA sequence	Cs	SNPs	AC	AC_{pam}	AC_{seed}	$AC_{seed\ small}$	MHs_{sum}	Mhs_{oof}	OOF	OT	OT_{sum}	Rank
526	2L	44993677	44993700	-	CTGACACACAAACGGTCCACCGG	0.32	13	35	3	17	1	1248	84.6	+|+|+|-|+	0|0|0|3	3	1
411	2L	44992694	44992717	+	ATAATCCTACGCAGTGGGCAAGG	0.14	8	17	2	3	1	1186.1	75.0	+|-|+|+|+	0|0|0|1	1	2
412	2L	44992695	44992718	+	TAATCCTACGCAGTGGGCAAGGG	0.14	8	17	1	3	0	1193.3	62.7	-|+|+|+|+	0|0|1|0	2	3
817	2L	44996271	44996294	+	GCGGTGGGACAGGATAAGTACGG	0.20	13	28	1	11	5	1124.4	79.0	+|-|+|+|+		0	4
563	2L	44993907	44993930	+	CCACCGGGCAAGAAGTGTAACGG	0.19	13	28	1	5	1	883.8	83.2	+|+|+|-|+		0	5
579	2L	44994009	44994032	+	GAGGAAGAAAGTGAGGAGGAGGG	0.05	21	188	68	87	50	1752.8	0	-|-|-|-|-	0|0|2|8	12	714

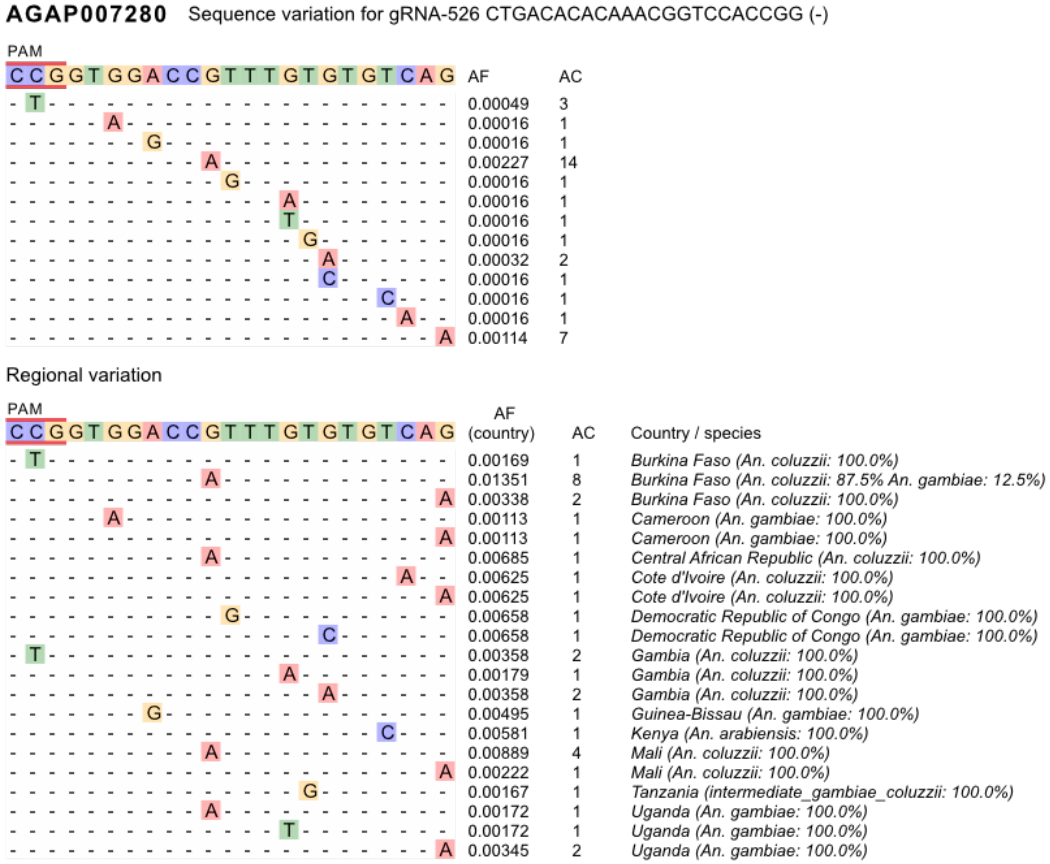


Figure 34 - Variation at target site of the highest-ranking gRNA-526 in *AGAP007280*. SNPs found in the sequence targeted by gRNA-526 are displayed with their allele frequency (AF) and allele counts (AC) based on the data obtained from Ag1000g phase 3 (The Anopheles gambiae 1000 Genomes Consortium, 2021). Observed SNPs were grouped by country where samples with the variant were collected. Relative allele frequency (AF), allele count (AC) and proportion of species where the alternative allele was found in (in percentage) are listed for each SNP in the country.

4.3.2.2 AGAP005958

AGAP005958 is an ortholog of *yellow-g* gene in *Drosophila.* Experiments showed that the gene is critical for follicle cell function and homozygous mutations in the gene lead to disruption of eggshell integrity. (Claycomb et al., 2004) The gene is female specific, haplosufficient and causes female sterility when both alleles of the gene are disrupted, which makes it a suitable gene drive target gene (Hammond et al., 2021a, 2016; Simoni et al., 2020).

gRNA search resulted in 167 potential target sites in *AGAP005958*. Four candidate gRNAs that were selected by ranking (gRNA-130, gRNA-131, gRNA-132, gRNA-133) were all clustered together in a 26 bp long region of exon 3 (Figure 35). They all had high MHs_{sum} scores with all of them producing predominantly out-of-frame predicted MMEJ deletions. The highest-ranking gRNA-132 had Cs 0.18 with no SNPs in PAM or small seed region and only one alternative allele in the larger seed region with frequency of 0.016% (Figure 36). gRNA-10 that did not cluster together with the other 4 candidate gRNAs was targeting the most conserved region in the exon 1 of the gene, which attributed to higher conservation score of Cs = 0.39 (Table 6). Overall, the coding region of *AGAP005958* showed high level of sequence conservation across the compared species. Previously proposed gRNA-74 had favourable predicted MMEJ outcomes but was ranked lower due to its high alternative allele count in the seed region of the gRNA.

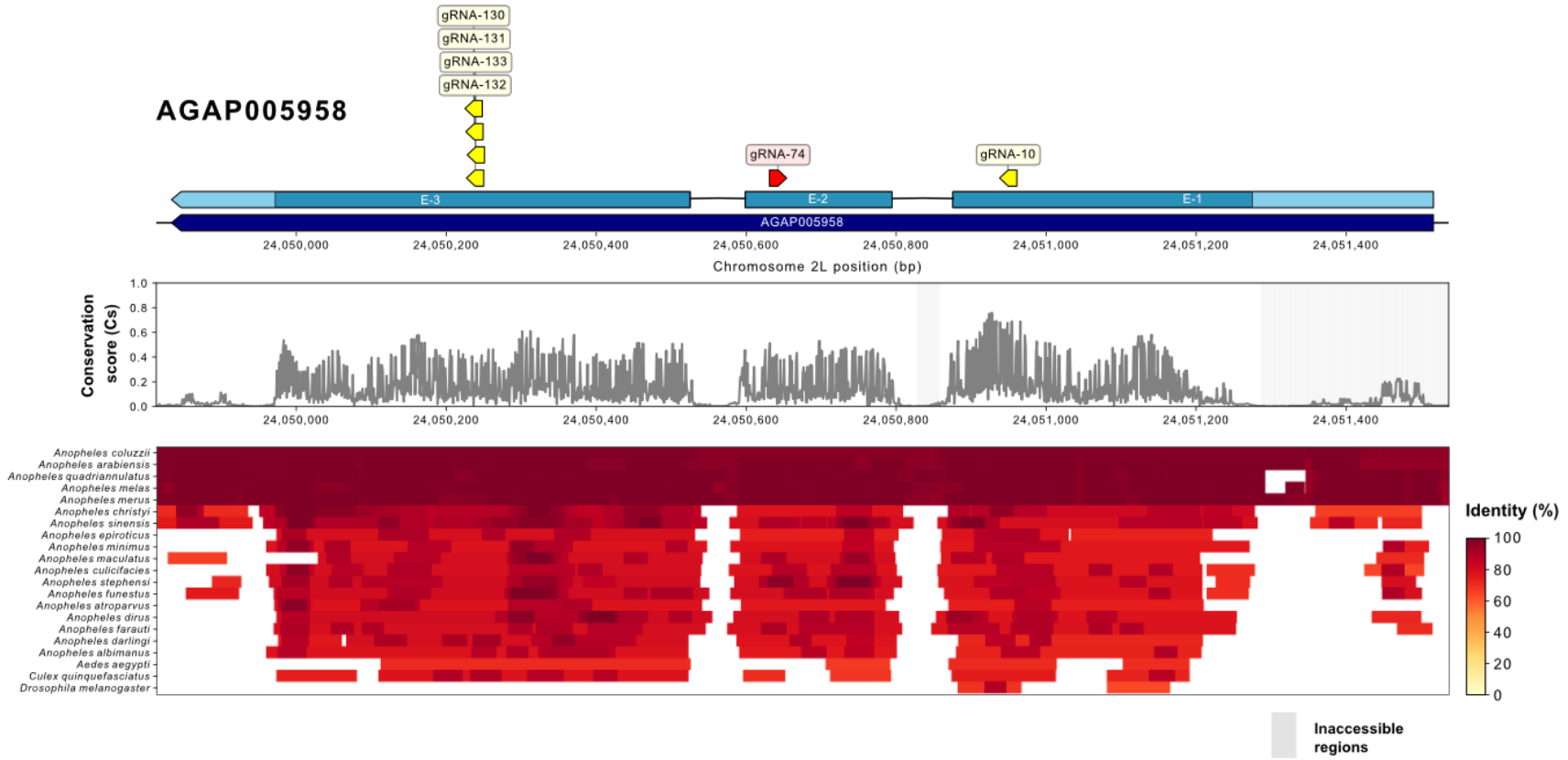


Figure 35 - Candidate gRNA target sites for *AGAP005958*. *AGAP005958* gene with 5 highest-ranking gRNAs in yellow and gRNA used by Hammond et al. (2016) in red. Conservation score is displayed on the middle panel in dark grey with inaccessible regions highlighted in light grey. Sequence identity (in percentages) is represented with heatmap for each species genome that was analysed for the calculation of conservation score.

Table 6 - Candidate gRNAs for *AGAP005958*. Genomic location and strand, gRNA sequence, the criteria used in ranking (Table 4), out-of-frame status (OOF), predicted off-targets (OT) and the final ranking are presented in the table. OOF indicates ordered 5 highest scoring MH deletions and their out-of-frame status denoted as + if deletion is out-of-frame or - if deletion is in-frame. OT denotes the number of off-targets with 0, 1, 2, 3 mismatches, respectively. gRNA used by Hammond et al. (2016) is highlighted in red.

gRNA	Genomic location and strand				gRNA sequence	Cs	SNPs	AC	AC_{pam}	AC_{seed}	$AC_{seed\ small}$	MHs_{sum}	Mhs_{oof}	OOF	OT	OT_{sum}	Rank
132	2L	24050229	24050252	-	CTCGGCACGGACGGTGGCAAGGG	0.18	12	24	0	1	0	1629.4	100	+|+|+|+|+	0|0|1|2	4	1
133	2L	24050230	24050253	-	GCTCGGCACGGACGGTGGCAAGG	0.18	12	24	2	2	0	1443.4	80.9	+|-|+|+|+	0|0|0|2	2	2
10	2L	24050940	24050963	-	CATTAAGCCGTACCCTTGCTGGG	0.39	6	15	0	4	0	1280.2	100	+|+|+|+|+		0	3
131	2L	24050228	24050251	-	TCGGCACGGACGGTGGCAAGGGG	0.17	13	189	166	2	2	1629.4	100	+|+|+|+|+	0|0|0|5	5	4
130	2L	24050227	24050250	-	CGGCACGGACGGTGGCAAGGGGG	0.17	13	189	166	2	2	1629.4	100	+|+|+|+|+	0|0|0|5	5	5
74	2L	24050633	24050656	+	GAGATACTGGAGCCGCGAGCTGG	0.18	17	143	2	105	3	1435.1	100	+|+|+|+|+		0	42

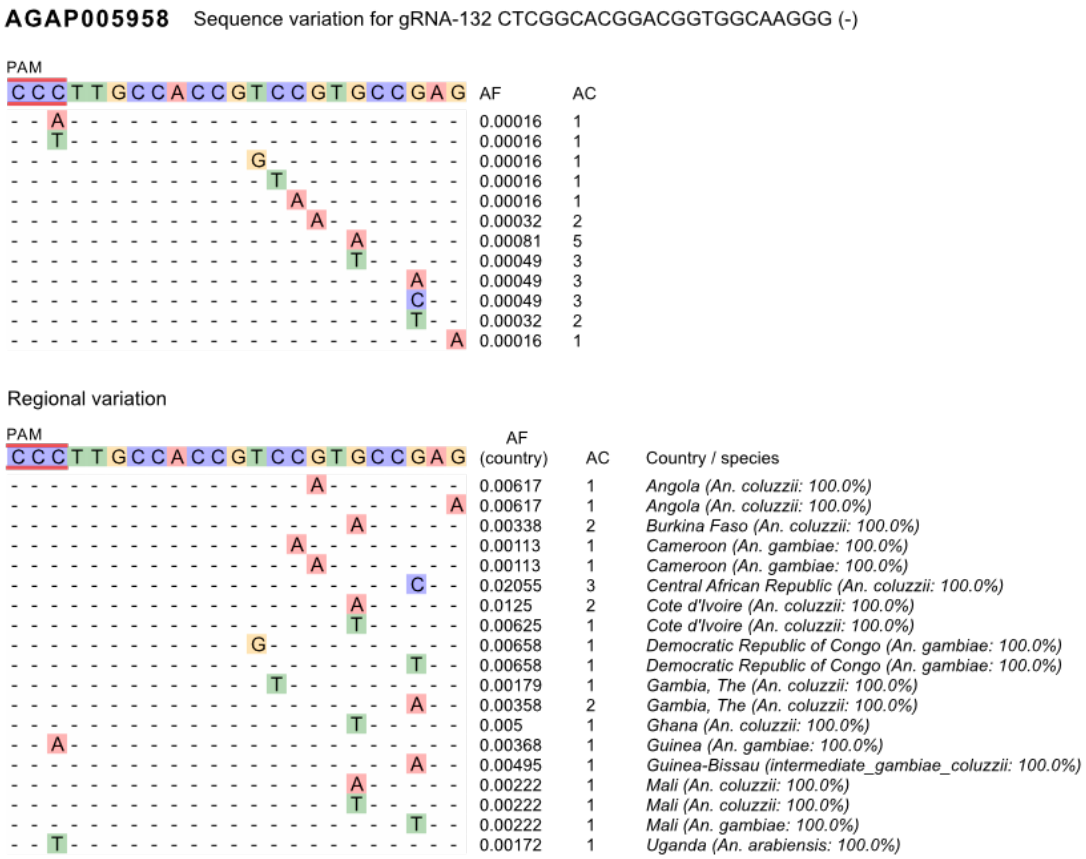


Figure 36 - Variation at target site of the highest-ranking gRNA-526 in gene *AGAP005958.* SNPs found in the sequence targeted by gRNA-132 are displayed with their allele frequency (AF) and allele counts (AC) based on the data obtained from Ag1000g phase 3 (The Anopheles gambiae 1000 Genomes Consortium, 2021). Observed SNPs were grouped by country where samples with the variant were collected. Relative allele frequency (AF), allele count (AC) and proportion of species where the alternative allele was found in (in percentage) are listed for each SNP in the country.

4.3.2.3 AGAP011377

AGAP011377 does not have a direct ortholog in *Drosophila* but contains a probable chitin-binding domain putatively involved in eggshell formation. The gene showed female-specific expression and knock-out experiments showed that females with homozygous mutation in the gene did not lay eggs, while heterozygous females' fertility was not affected. Based on those finding the gene was selected as a candidate gene for gene drive by Hammond et al. (2016).

100 gRNA target sites were found in *AGAP011377* by Guido. gRNA-64 was ranked the highest with relatively low conservation score Cs = 0.06, MHs_{sum} of 1153 and only two predicted MMEJ out-of-frame deletions and one off-target with 3 mismatches (Figure 37, Table 7). gRNA-70 was ranked as the second option and showed the highest conservation score in the gene. It also had better predicted outcomes from MMEJ repair, however its overall MHs_{sum} was lower, which decreased its ranking. Comparison in sequence variation between gRNA-64 and gRNA-70 also showed that gRNA-64 has a SNP present in the PAM region in frequency of 0.4%, while gRNA-70 does not contain any SNPs in PAM region. However, it has a G>A SNP at the PAM-distal side of the gRNA (position 21) that is present in high frequency (20.3%) in the sampled Angolan *An. coluzzii* population (Figure 38). The presence of a variant on the very PAM-distal end of the gRNA should have minor effect on the target recognition, but its potential to impair the cleavage should nonetheless be considered when selecting gRNA.

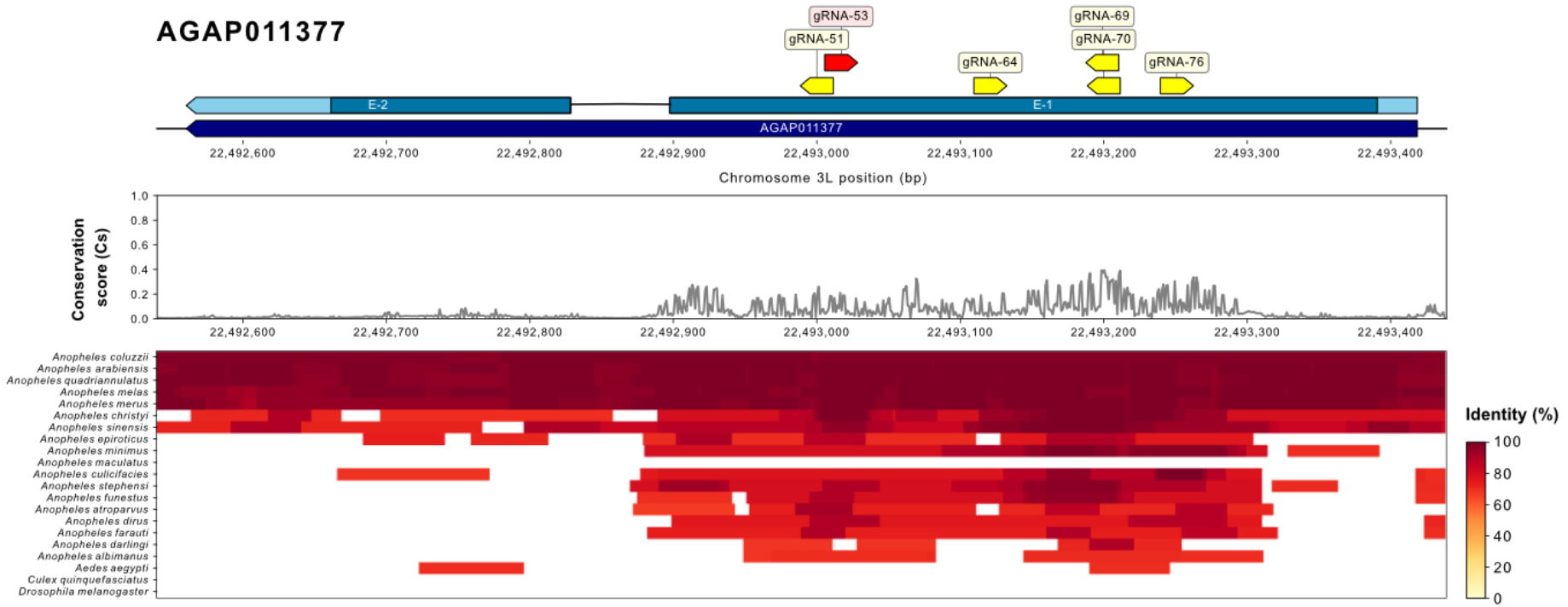


Figure 37 - Candidate gRNA target sites for *AGAP011377*. *AGAP011377* gene with 5 highest-ranking gRNAs in yellow and gRNA used by Hammond et al. (2016) in red is displayed in the plot. Conservation score is displayed on the middle panel in dark grey with inaccessible regions highlighted in light grey. Sequence identity (in percentages) is represented with heatmap for each species genome that was analysed for the calculation of conservation score.

Table 7 - Candidate gRNAs for *AGAP011377*. Genomic location and strand, gRNA sequence, the criteria used in ranking (Table 4), out-of-frame status (OOF), predicted off-targets (OT) and the final ranking are presented in the table. OOF indicates ordered 5 highest scoring MH deletions and their out-of-frame status denoted as + if deletion is out-of-frame or - if deletion is in-frame. OT denotes the number of off-targets with 0, 1, 2, 3 mismatches, respectively. gRNA used by Hammond et al. (2016) is highlighted in red.

gRNA	Genomic location and strand				gRNA sequence	Cs	SNPs	AC	AC_{pam}	AC_{seed}	$AC_{seed\ small}$	MHs_{sum}	Mhs_{oof}	OOF	OT	OT_{sum}	Rank
64	3L	22493111	22493134	+	CTGCGTAAAGTCACACGTACAGG	0.06	8	39	29	2	2	1153.1	38.7	-|+|-|+|-	0|0|0|1	1	1
70	3L	22493190	22493213	-	CGCTACTACAAGTGTACCGATGG	0.24	13	82	0	7	2	941.1	65.9	+|+|+|-|-		0	2
51	3L	22492990	22493013	-	CTGCTACAACGACGGATATGCGG	0.09	12	631	389	4	1	1525	100	+|+|+|+|+		0	3
69	3L	22493189	22493212	-	GCTACTACAAGTGTACCGATGGG	0.23	14	91	11	23	18	903.5	64.5	+|+|-|-|+	0|0|0|1	1	4
76	3L	22493241	22493264	+	ACCTGCAGCAGTGGTGCCGACGG	0.15	18	45	4	7	1	1259.4	60.6	+|+|-|-|+	0|0|0|4	4	5
53	3L	22493007	22493030	+	TAGCAGACGTAGAAATTTTCCGG	0.09	15	331	4	259	3	900.8	100	+|+|+|+|+		0	23

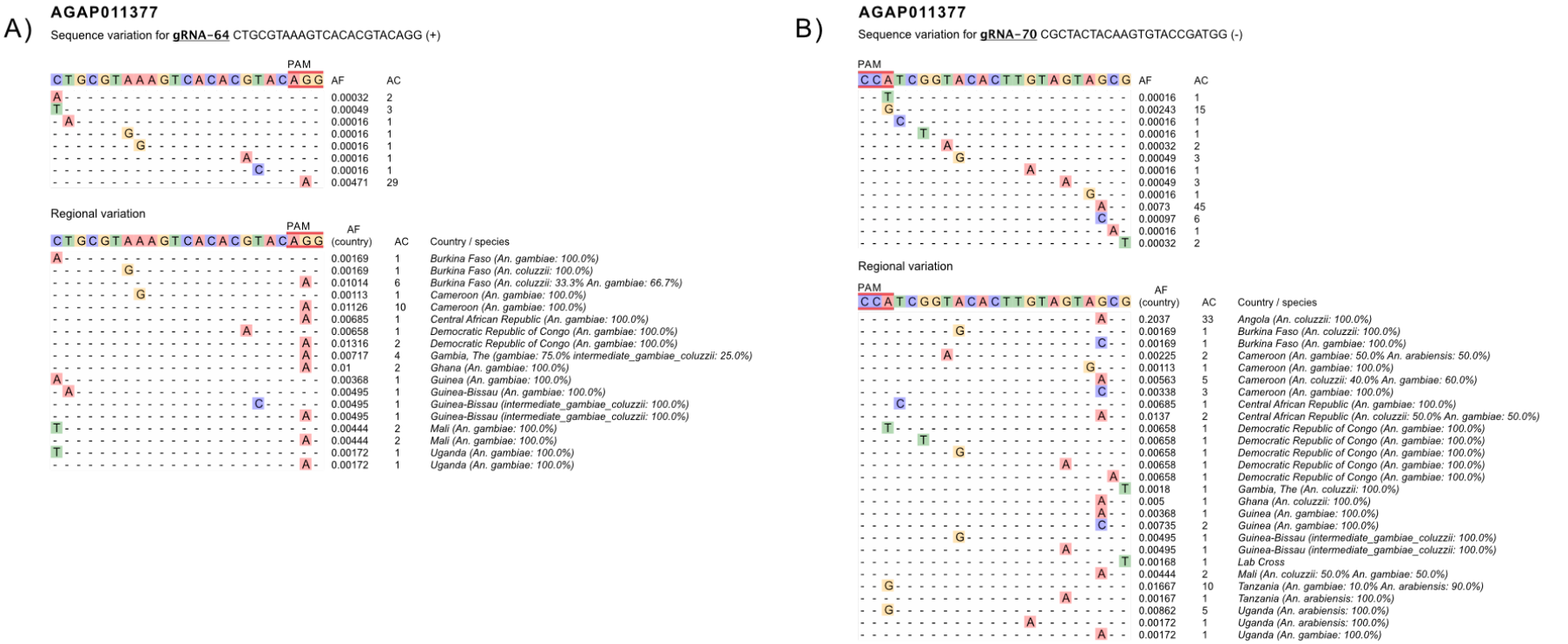


Figure 38 - SNPs found in the sequence targeted by gRNA-64 and gRNA-70 in gene *AGAP011377.* They are displayed with their allele frequency (AF) and allele counts (AC) based on the data obtained from Ag1000g phase 3 (The Anopheles gambiae 1000 Genomes Consortium, 2021). Observed SNPs were grouped by country where samples with the variant were collected. Relative allele frequency (AF), allele count (AC) and proportion of species where the alternative allele was found in (in percentage) are listed for each SNP in the country.

4.3.2.4 AGAP004050 - *doublesex*

Doublesex (*dsx* or *AGAP004050*) acts as the final binary switch between male and female sex development and was selected as a gene drive target site by Kyrou et al. (2018). Functionally, *dsx* activates or represses the genes necessary for the differentiation of sexually dimorphic traits. It produces two alternatively spliced isoforms: male-specific *dsxM* and female-specific *dsxF*. The female-specific segment corresponds to exon 5 and is unique to *dsxF*. The target site recognised by the gene drive in Kyrou et al lies in the female-specific exon 5 and overlaps the splicing junction at the edge of intron 4. It is highly conserved (Figure 39) and in caged population invasion experiments the gene drive was able to spread to fixation, causing complete population suppression due to a lack of viable females.

In addition to the gRNA targeting the locus in the gene drive experiments, here named gRNA-1152, 8 other potential gRNAs were found in *dsx* exon 5 (Figure 39, Table 8). All putative gRNAs had high mean conservation score (Cs>0.4) as expected from the results of the sequence conservation analysis. gRNA-1151 that was ranked the highest also had the highest mean Cs 0.6 compared to the mean Cs 0.4 of previously used gRNA-1152. Its seed region contained only one SNP that was found in very low frequency (0.02%) in a laboratory colony that was used for lab crosses in the Ag1000g project (Figure 40). No variation at the gRNA-1151 target site was observed in the sampled wild populations in Africa. gRNA-1152 small seed complementary sequence contained one SNP present in 1.1% of alleles and another SNP in the seed region present with frequency of 0.02%. *In vitro* cleavage assay by Kyrou et al. (2018) showed that cleavage still occurs with the sequence containing the variant in the small seed region of the gRNA. Despite encouraging results achieved by using gRNA-1152, alternative target sites, which showed arguably even better properties, could prove to be of use in the future experiments.

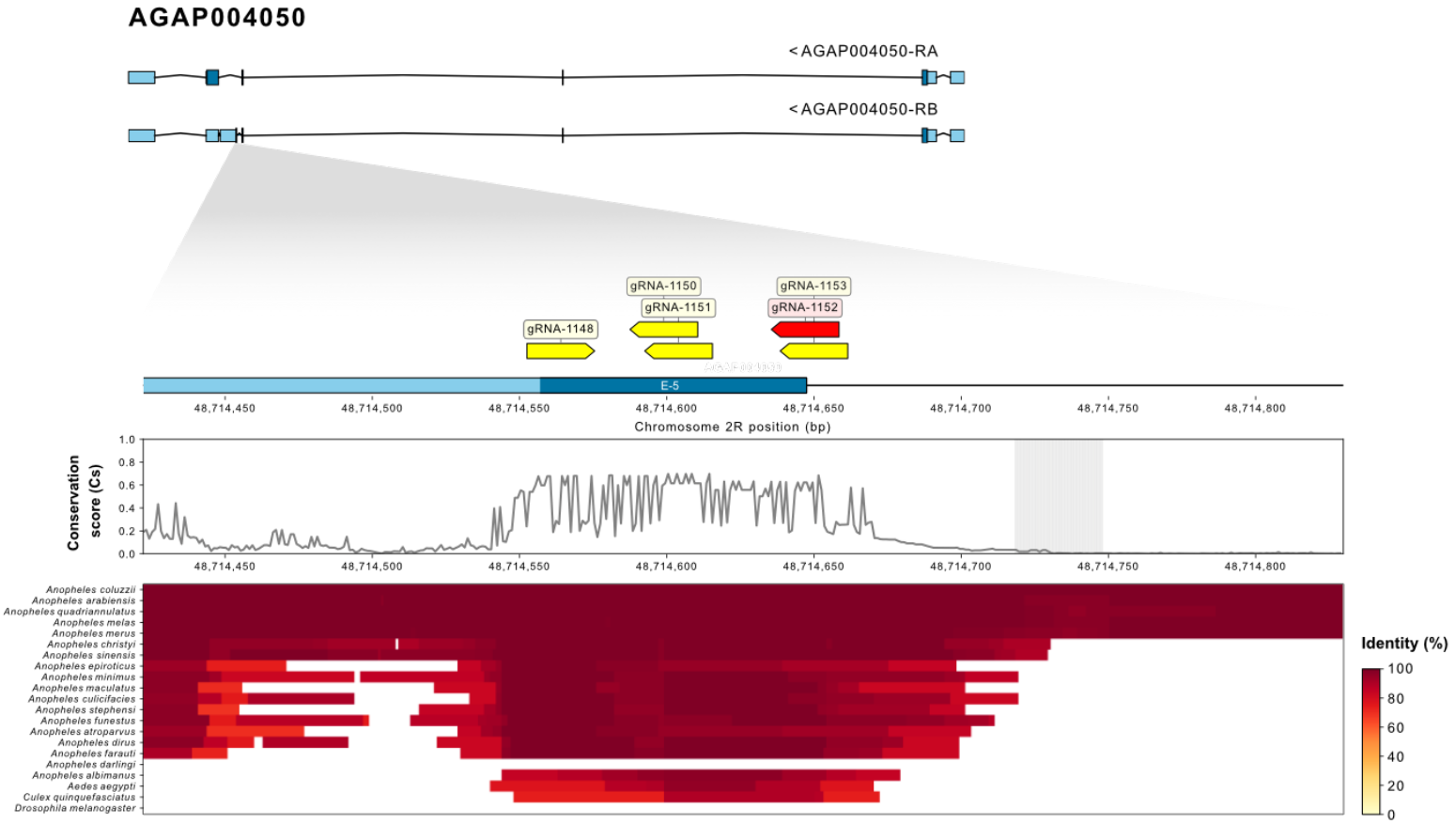


Figure 39 - Candidate gRNA target sites for AGAP004050-RB (*dsxF*). *AGAP004050* gene with 5 highest-ranking gRNAs (yellow) in the female specific exon 5 and gRNA used by Kyrou et al. (2018) in red. Conservation score is displayed on the middle panel in dark grey with inaccessible regions highlighted in light grey. Sequence identity (in percentages) is represented with heatmap for each species genome that was analysed for the calculation of conservation score.

Table 8 - Candidate gRNAs for *AGAP004050*. Genomic location and strand, gRNA sequence, the criteria used in ranking (Table 4), out-of-frame status (OOF), predicted off-targets (OT) and the final ranking are presented in the table. OOF indicates ordered 5 highest scoring MH deletions and their out-of-frame status denoted as + if deletion is out-of-frame or - if deletion is in-frame. OT denotes the number of off-targets with 0, 1, 2, 3 mismatches, respectively. gRNA used by Kyrou et al. (2018) is highlighted in red.

gRNA	Genomic location and strand				gRNA sequence	Cs	SNPs	AC	AC_{pam}	AC_{seed}	$AC_{seed\ small}$	MHs_{sum}	Mhs_{oof}	OOF	OT	OT_{sum}	Rank
1151	2R	48714593	48714616	-	CATAATCTGAACATGTTTGATGG	0.60	1	1	0	1	0	1141.4	83.4	+|+|+|-|+	0|0|0|14	14	1
1148	2R	48714553	48714576	+	TTATCATCCACTCTGACGGGTGG	0.48	2	2	0	1	0	912	82.9	+|+|+|+|-		0	2
1150	2R	48714588	48714611	-	TCTGAACATGTTTGATGGCGTGG	0.56	2	2	0	1	1	1013.7	62.5	+|+|+|-|-		0	3
1152	2R	48714636	48714659	-	GTTTAACACAGGTCAAGCGGTGG	0.44	2	71	0	70	70	1262.2	88.9	+|+|-|+|+		0	4
1153	2R	48714639	48714662	-	TATGTTTAACACAGGTCAAGCGG	0.40	2	71	70	1	0	1235.9	70.7	+|+|-|-|+		0	5

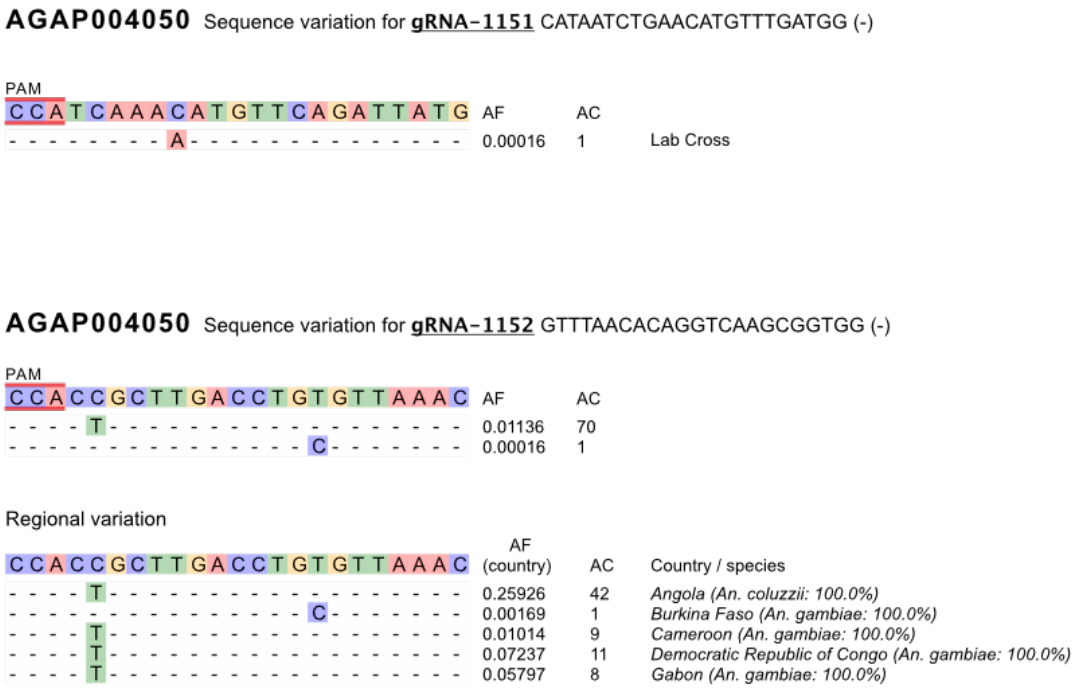


Figure 40 - Variation at target site of the highest-ranking gRNA-1151 and gRNA-1152 used by Kyrou et al. (2018) in exon 5 of gene *AGAP004050 (dsx)*. SNPs found in the sequence targeted by gRNA-1151 and gRNA-1152 are displayed with their allele frequency (AF) and allele counts (AC) based on the data obtained from Ag1000g phase 3 (The Anopheles gambiae 1000 Genomes Consortium, 2021). Observed SNPs were grouped by country where samples with the variant were collected. Relative allele frequency (AF), allele count (AC) and proportion of species where the alternative allele was found in (in percentage) are listed for each SNP in the country.

4.4 Discussion

A novel bioinformatic pipeline was created as part of my research to obtain a genome-wide conservation score in *An. gambiae* reference genome (AgamP4). Conservation score was devised with goal to identify conserved genomic regions that could be used as gene drive target sites that are less likely to tolerate resistant mutations as they are subjected to high functional constraint. In addition to interspecies variation obtained from the 21 genomes, genome-wide SNP density from 1,142 wild-caught mosquitoes was included in the conservation score calculation to account for intraspecies genome variability. The conservation score was supported also by inclusion of phyloP scores that indicated whether positions within highly identical genomic regions are under negative or positive selection, based on the phylogenetic model of neutral evolution (Hubisz et al., 2010). The conservation score can serve as proxy for structurally or functionally constrained sites, however it is important to note, that level of conservation score does not directly indicate the actual level of functional constraint applied to the locus. One locus, for example, might have 5% fitness cost associated with mutations and another locus could have 80% fitness cost associated with mutations there, yet both might result in the same level of ultraconservation of the sequences. The challenge remains how to distinguish these loci. The conservation score metric is obtained through measuring sequence variability, which can be reduced by other evolutionary forces than only puryfing selection, such as selective sweep or background selection (Charlesworth, 2009; Charlesworth et al., 1993). This was partially addressed by including species that span more than 100 million years of evolution (Neafsey et al., 2015). Nonetheless, the force of purifying selection could be sufficient to reduce genetic variation in loci where changes had no significant reductions in fitness as it would at sequence changes that were nonviable. Conservation score should also be interpreted cautiously as the indicator of functional constraint in the regions with high content of repetitive sequences or other low-complexity genomic regions. Level of

the conservation of such sequences is maintained by concerted evolution rather than high functional constraint (Liao, 2008). For that purpose genome accessibility annotation generated in Ag1000g project was added for each inspected locus. The genome accessibility annotation includes masked genomic regions containing various families of repetitive sequences and regions with low coverage and poor mapping quality (Miles et al., 2017; Sharakhova et al., 2007). The conservation score obtained for *An. gambiae* genome complements the effort to investigate ultra-conserved genomic regions in the genome and extends it by assigning a measure of sequence conservation to every locus in the genome (O'Loughlin et al., 2021).

Equally important for resistance mitigation is to account for propensity of a target site to generate mutations that would retain the function of the targeted gene but prevent the target site recognition by Cas9-gRNA complex. Deletions arising from MMEJ repair mechanism can be predicted by identifying two microhomologous sequences used in the annealing process that are present in the flanking regions on both sides of the cut site. The resected DNA between them is the resulting deletion that can be detected by genotyping the target site after cleavage. A computational alghoritm that predicts these deletions was implemented as part of my research project. The scoring system that associates predicted MMEJ deletions with their observed frequencies was modelled first by Bae et al. (2014) from the results obtained from sequence context and MMEJ frequencies in mammalian cell lines. To test how MMEJ deletions and their MH scores correlate with observed mutation frequencies *in vivo* in *An. gambiae* we selected 4 gRNAs with different MMEJ properties. In attempt to obtain transgenic lines and record the generated mutations at 4 different target sites, three lines were viable and only two showed levels of end-joining that were suitable for further analysis (target sites T6 and T7). Both target sites showed significantly higher overall mean frequency of MMEJ deletions (~1.4-fold higher) compared to NHEJ deletions. T6 and T7 target sites were selected with similar overall MMEJ sum score, which indicates the propensity for MMEJ

repair mechanism based on the sequence context at the target site, but with different distribution of MH scores across the highest scoring MH deletions at each target site. The observed frequency of MMEJ deletions did not follow the relative difference in the top two MH scores which were selected to be dissimilar between T6 and T7. However, the correlation between all MH scores and observed MMEJ frequencies at both target sites showed strong positive correlation (Spearman correlation, $\rho>0.6$, $p<0.001$). Importantly, it was shown that the 5 highest-scoring predicted MMEJ deletions on both target sites amounted to more than 97% of all observed MMEJ mutations that were generated at the target sites, which confirms that it is essential to select the target sites with consideration on their MMEJ outcomes.

Experiments in zebrafish embryos showed that MH scores that cluster closely together (i.e. smaller differences in scores between consecutive MMEJ deletions) negatively impacted the MMEJ activation at the target site. They hypothesise that MMEJ activation is more favourable at loci with one or two predominant microhomology pairs rather than many strong microhomology pairs. (Ata et al., 2018) In our *in vivo* experiments the two targets T6 and T7 were selected on similar basis and showed each a predominant MH pair despite their difference in MH score distribution. Nonetheless, the size of our experiment was unfavourable for making definite conclusions. To further adjust the MMEJ scoring model to better fit the observed frequency of MMEJ deletions many more target sites would need to be tested. A machine learning model similar to inDelphi, which was trained on the outcomes of repair mechanisms in mammalian cells at 2,000 target sites by Shen et al. (2018), would be undeniably useful for predicting genome editing outcomes in mosquito germline and somatic cells. However, a mosquito cell culture would need to be used to parallelize the process of screening as the laborious process of generating mosquito transgenic lines for such extensive screening experiment would be practically impossible.

Conservation score and MMEJ prediction algorithm described herein were packaged in a computational tool called Guido, which searches a sequence or a genomic region for potential CRISPR-Cas target sites. Although it was built with gene drive in mind, particularly in *An. gambiae*, it remains genome agnostic and highly customisable. Its modular architecture allowed integration of various variation datasets, which inform the resarcher about pre-existing sequence variation at the target site (Miles et al., 2017; The Anopheles gambiae 1000 Genomes Consortium, 2021, 2020). In light of the research on PAM specificites of different Cas9 orthologs described in Chapter 3, custom PAM preferences can be used with Guido to find novel gRNA targets for genome editing. *In silico* prediction of off-target activity is also provided with Guido. Predicted off-targets included most observed *in vivo* gene drive off-target effects that were observed in study performed in our lab (Garrood et al., 2021). However, the same *in vivo* study of off-target effects showed that pre-existing genome variation can enhance off-target activity by reducing the number of mismatches to the target site or convert it to a more cleavable site (i.e., mutations that create a PAM site). In future, the impact of genetic variation on off-target activity should be addressed already at the stage of gRNA selection by integration of variation-aware *in silico* off-target site prediction (Cancellieri et al., 2020).

The field of genome editing has been advancing rapidly and it has brought a myriad of computational tools to support it (Bae et al., 2014b; Heigwer et al., 2014; Moreno-Mateos et al., 2015; Park et al., 2015; Stemmer et al., 2015). Many of these tools purposed for gRNA design and selection include extra features, such as inclusion of on-target specificity and efficiency scores, primer design for target site amplification, and come with interactive visual interface (Concordet and Haeussler, 2018; Labun et al., 2019). Noneteless, Guido stemmed from the need to account for potential gene drive resistance at the target site and has shown to complement the existing gRNA tools and is used by the researchers in our lab.

Guido was used for the search of alternative target sites in the previously selected candidate genes for gene drive applications. Three female fertility genes selected by Hammond et al. (2016) and the female-specific exon 5 in *doublesex* selected by Kyrou et al. (2018) were analysed. Guido was able to find alternative target sites that are arguably more suitable for gene drive in each selected locus. The ranking of putative gRNA target sites was performed by including weighted criteria that incorporated the level of conservation at the target site (through conservation score), pre-existing sequence variation in the regions that could affect gRNA-Cas9 binding, predicted MMEJ outcomes and predicted off-target activity. The distribution of weights across criteria could potentially be changed depending on the genome drive application. For example, when using multiplexing, where more than one gRNA is used at the same time to target a locus to mitigate resistance, lower weights could be used for criteria considering MMEJ outcomes and higher weights could be used for criteria that account for off-target activity (Champer et al., 2020). Similarly, if Guido would be used for selecting a target site for population replacement gene drive, where the target site does not need to be in the coding region of a gene, criteria with predicted MMEJ outcomes could be omitted and weights could be redistributed to select for a conserved non-coding locus with minimal off-targets. Current implementation of criteria that accounts for the number of alternative alleles found within the seed region, small seed region and PAM region of the target site could also be extended further. Separate weights could be assigned to each nucleotide in the gRNA binding region based on the effect that a variant there would have on gRNA-Cas9 binding (Xu et al., 2017). Efforts to sample and monitor genetic variability in local mosquito populations also bring the opportunity to tailor the gRNA design to target a haplotype or an allele that is specific to a local population and leverage it to geographically limit the spread of a gene drive, as suggested by Willis and Burt (2021). Implementing such variation-aware algorithm into Guido that would find private alleles and use them to design novel, population-specific gRNAs, would further extend its flexibility and usability.

Increasing demand for suitable target sites for genetic vector control was the basis for the research described in this study. Novel data resources and methodologies were implemented in gRNA design to offer a solid base for rational gene drive target site selection. Furthermore, the computational tools built in this study offer the flexibility to be extended for the future needs that are likely to arise.

4.5 Methods

4.5.1 Identification of ultra-conserved genomic regions

4.5.1.1 Genomic data

Reference genomes of 18 Anopheles species, *Aedes aegypti*, *Culex quinquefasciatus* and *Drosophila melanogaster* were used in the analysis. AgamP4 (*An. gambiae*), AcolM1 (*An. coluzzii*), AaraD1 (*An. arabiensis*), AquaS1 (*An. quadriannulatus*), AmelC2 (*An. melas*), AmerM2 (*An. merus*), AepiE1 (*An. epiroticus*), AchrA1 (*An. christyi*), AsinC2 (*An. sinensis*), AminM1 (*An. minimus*), AmacM1 (*An. maculatus*), AculA1 (*An. culicifacies*), AsteI2 (*An. stephensi*), AfunF1 (*An. funestus*), AatrE3 (*An. atroparvus*), AdirW1 (*An. dirus*), AfarF2 (*An. farauti*), AdarC3 (*An. darlingi*), AalbS2 (*An. albimanus*), AaegL5 (*Aedes aegypti*), CpipJ2 (*Culex quinquefasciatus*) reference genomes were obtained from VectorBase (Giraldo-Calderon et al., 2015). The fruitfly genome of *Drosophila melanogaster* DmelP6 was obtained from Flybase (Larkin et al., 2021).

4.5.1.2 Genome alignment

CNEr package was used to streamline the process of large-scale identification of conserved noncoding elements (CNEs). (Tan et al., 2019) In the process of CNE detection, CNEr uses LASTZ genome aligner for the initial genome alignment. Every reference genome used in the analysis was aligned to *An. gambiae* reference genome AgamP4. Each alignment was further processed with *axtChain*, *chainMergeSort*, *chainPreNet*, *chainNet*, *netSyntenic*, *netToAxt* and *axtSort* tools for chaining and netting steps, for which the default parameters were used (C=0 E=30 H=0 K=3000 L=3000 M=50 O=400 T=1 Y=9400).

4.5.1.3 Sequence identity scans

Upon the alignment the selected genomes were scanned for highly identical regions relative to *An. gambiae* AgamP4 genome with the pipeline provided by CNEr. Two rounds of scans were performed. In the first round identity thresholds 90%, 97% and 100% were used over a scanning window of 30 bp and in the second round identity thresholds of 70%, 90%, 96%, 98%, 100% were used over a scanning window 50bp, respectively. Although CNEr was designed for detection of conserved noncoding elements, by omitting the filtering step it retained detected conserved elements in the coding regions as well. This procedure yielded a list of conserved genomic regions between each selected genome and *An. gambiae* where the sequence identity was higher than a threshold set in each scan. The minimal detected identity was set to 70%. For example, an identity value of 98% corresponds to detecting 49 identical positions in the 50bp alignment window.

4.5.1.4 Conservation score calculation

Genomic intervals with their corresponding sequence identity values for each pairwise alignment were mapped to a matrix of the size of each chromosome, allowing for per-base pair resolution of identity scores and easier comparison across different species. The matrix was implemented as array in Zarr v2.4.0 (Miles et al., 2020).

To calculate an overall conservation score (Cs), identity values were normalised with the phylogenetic distance between each species and *An. gambiae*. The phylogenetic tree was generated from the genome alignments using PHAST phyloFit (Hubisz et al., 2010). Phylogenetic distances within the tree were calculated using ETE Toolkit based on the phylogenetic tree resulting from phyloFit analysis (Huerta-Cepas et al., 2016).

Average normalised identity values were calculated for each position and SNP density data were used to skew values higher when SNP density was low. Furthermore, phyloP

was used to calculate scores with likelihood ratio test (--method LRT) representing conservation (--mode CON) for each genomic position. REV substitution model was used.

All calculations were done using Python modules Pandas, Numpy, Scipy, Zarr, h5py, scikit-learn and scikit-allel (Miles et al., 2019). The data was stored in HDF5 and Zarr format and plotted using Matplotlib and Seaborn. Genome accessibility data was obtained from Ag1000g phase 2 (The Anopheles gambiae 1000 Genomes Consortium, 2021, 2020).

The produced dataset was published in Kranjc et al. (2021) and can be accessed at https://zenodo.org/record/4304586. A Python tool for easier accessibility of the results was created and can be found with more detailed instructions together with the scripts used to generate the dataset at: https://github.com/nkran/AgamP4_conservation_score.

4.5.1.5 Gene ontology (GO) enrichment analysis

A list of exonic intervals was created from AgamP4.12 genome annotation and Cs was extracted for each position in exons. Mean Cs was then calculated for all exons that belong to their respective genes and sorted to indicate their mean conservation score. The 95th percentile of genes was selected based on their mean conservation score to obtain the most conserved genes in the genome. The selected genes were then used in GO enrichment analysis using PANTHER database (v14) (Mi et al., 2019).

4.5.2 Prediction of on-target site mutations

4.5.2.1 Molecular cloning

To introduce the gRNAs of interest in the mosquito genome we used the p125 plasmid. The p125 plasmid contains a gRNA expression cassette made up of a U6 ubiquitous PolIII promoter, *Bsa*I restriction sites allowing the Golden Gate cloning of any gRNA spacer at

that site, providing it is flanked by compatible overhangs, a gRNA scaffold and U6 terminator. It also contains an RFP marker whose expression is controlled by the 3xP3 neuronal promoter and SV40 terminator. These are flanked by 3' and 5' piggyBac repeats that allow the random integration of gRNA and linked fluorescent marker into the mosquito genome provided the plasmid is co-injected with a helper plasmid expressing a transposase. However, to avoid potential silencing of gRNA expression by nearby regulatory elements due to the random genomic integration of the constructs, we decided to introduce Gypsy insulators immediately after the piggyBac repeats to flank the gRNA and RFP marker. Gypsy insulators have been used previously to dampen the effect of nearby regulatory elements and aid the precise expression of transgenes in *Anopheles stephensi* (Carballar-Lejarazú et al., 2013).

The p125 plasmid (7559 bp long) was digested with *Bcu*I and *Bgl*II FastDigest restriction enzymes, following the FastDigest protocol (Thermofisher). *Bcu*I cuts between the 3'piggyBac repeat and the 3xP3 promoter. *Bgl*III cuts between the U6 terminator and the 5'piggyBac repeat. The digest yielded two fragments 4967 and 2592 bp long, that were purified directly after the digestion using the Wizard® SV Gel and PCR Clean-Up kit (Promega) to remove contaminants.

Gypsy insulators were amplified, using a Phusion® DNA polymerase mastermix (NEB), from an attB-CFP-Gypsy-UAS14 plasmid using a 60ºC annealing temperature and the below primer pairs for the 5' and 3' Gypsy elements, respectively:

- RightGypsy-p125-GA-F
  ttgcggcacatacgcgtcggacctctttgaCACGTAATAAGTGTGCGTTG
- RightGypsy-p125-GA-R
  ccgagtctctgcactgaacattgtcagatcGGAATTGATCGGCTAAATGG
- LeftGypsy-p125-GA-F
  gtcttcctactgcaggcggccgctctagaaCACGTAATAAGTGTGCGTTG

- LeftGypsy-p125-GA-R
  tgatgacctgcagcccgggggatccactagGGAATTGATCGGCTAAATGG

Primers contain overhangs (lowercase letters) complementary to the sequences upstream and downstream of the p125 restricted sites to allow Gibson assembly of the two p125 digested fragments and the two Gypsy insulators. Both PCRs yielded the expected 418 bp PCR products that were purified using the Wizard® SV Gel and PCR Clean-Up kit (Promega). Finally, a Gibson assembly of the four fragments was performed, using a 1:3 molar ratio of p125 fragments to Gypsy insulators and the NEB Gibson Assembly® cloning kit protocol, followed by bacterial transformation of the assembled product into NEB chemically competent cells.

The final p125-Gypsy plasmid was verified by diagnostic digestion using FastDigest enzymes *Sca*I and *Sma*I and Sanger sequencing (Eurofins), using primers p125-seq-R1 TGCTCTTCTGCAAAGCGATG and p125-seq-F1 CTTACTTGGCTTGTTTGGGA.

In the meantime T5, T6, T7 and T8 ssDNA oligos (Table 9) were annealed to produce dsDNA with overhangs complementary to *Bsa*I-digested p125-Gypsy plasmid (FastDigest, Thermofisher), to allow GoldenGate cloning of the gRNAs into the plasmid as previously done by Hammond et al. (2018). The resulting plasmids were verified by Sanger sequencing using the U6prom-F primer TGCGCTTGAAGGGTTGATCG, and prepared for mosquito embryo microinjection injection using the ZymoPURE™ II Plasmid Maxiprep Kit (ZymoResearch).

Table 9 - Oligo pairs used for GoldenGate cloning of T5, T6, T7 and T8 gRNAs.

gRNA target site	Oligo pairs	Annealed oligos
T5	TGCTGTCTATTTTGTTCTACCGCAA	TGCTGTCTATTTTGTTCTACCGCAA CAGATAAAACAAGATGGCGTTCAAA
	AAACTTGCGGTAGAACAAAATAGAC	
T6	TGCTGATTTCCGGGTGGCATGCA	TGCTGATTTCCGGGTGGCATGCA CTAAAGGCCCACCGTACGTCAAA
	AAACTGCATGCCACCCGGAAATC	
T7	TGCTGCAACCACGGGCTGAAGATC	TGCTGCAACCACGGGCTGAAGATC CGTTGGTGCCCGACTTCTAGCAAA
	AAACGATCTTCAGCCCGTGGTTGC	
T8	TGCTGAATACAAAAGACGGCGGCGG	TGCTGAATACAAAAGACGGCGGCGG CTTATGTTTTCTGCCGCCGCCCAAA
	AAACCCGCCGCCGTCTTTTGTATTC	
Note: Overhangs are shown in **bold**, and extra G was added in the gRNA spacer sequence in some cases to allow expression from the U6 PolIII promoter.		

4.5.2.2 Mosquito microinjection

Freshly laid embryos of the *Anopheles gambiae* G3 "wild-type" strain were microinjected with the relevant p125-Gypsy plasmid (300 ng/µl) bearing a gRNA targeting T5, T6, T7 or T8, together with a helper plasmid (300 ng/µl) containing a piggyBac transposase, expressed under the *vasa2* promoter. This is essential to allow random integration of the piggyBac-flanked region on p125-Gypsy into the mosquito genome. Surviving injected individuals were crossed to wild-type mosquitoes. Their RFP fluorescent offspring were selected as positive transformants. A single healthy transformant from each injection set was used to eventually establish each of the transgenic strains obtained: dsx-T5-gRNA, dsx-T6-gRNA and dsx-T7-gRNA. After 8 set of injections with the plasmid containing the T8 gRNA, not a single transgenic strain could be established and very high larval mortality was observed, despite purifying the plasmid on three separate occasions.

4.5.2.3 Genetic crosses, gDNA extractions and sequencing

50 RFP+ males of each of the three gRNA strains (dsx-T5-gRNA, dsx-T6-gRNA and dsx-T7-gRNA) were crossed to 50 YFP+ *vas2::Cas9*-containing females. The same cross was performed in triplicate for each strain, and the resulting progeny were collected straight after hatching for *en masse* gDNA extractions using the DNeasy® Blood and Tissue kit (Qiagen). Using 100ng starting gDNA a 200-500 bp locus was amplified spanning the relevant target site (T5, T6, T7) using the KAPA HiFi HotStart Ready Mix PCR kit (Kapa Biosystems), and primers containing partial Illumina AmpEZ adapters (underlined):

- **4050-T5-AmpEZ-F**: ACACTCTTTCCCTACACGACGCTCTTCCGATCTCAATAATCATGTTTTCTGTTCG
- **4050-T5-AmpEZ-R**: GACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTAATAGATCGATTGGAAATTCAC
- **4050-T6-AmpEZ-F**: ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGCACATGTACCATTTAAAAAG
- **4050-T6-AmpEZ-R**: GACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTCAAAGATTAGTTTGTGTTGGTG
- **4050-T7-AmpEZ-F**: ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGGTATGTGAATGTTAGCTACC
- **4050-T7-AmpEZ-R**: GACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTAGAGCTTGTTTGGAAGACATAC

PCRs were run with an annealing temperature of 62ºC for 20 cycles only (non-saturating conditions). PCR products were purified using the Wizard® SV Gel and PCR Clean-up System (Promega) and sequenced using Genewiz's AmpEZ NGS service.

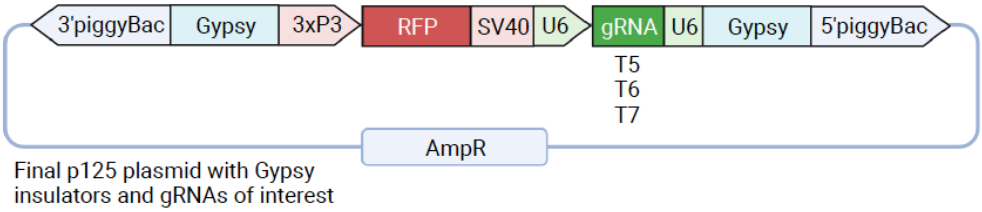


Figure 41 - Schematic of p125-Gypsy plasmid. Final plasmid that was used in the microinjections containing Gypsy insulator, RFP marker and T5, T6 and T7 gRNAs.

4.5.2.4 Data analysis

Pooled amplicon sequencing reads were analysed using CRISPResso2 v2.1.3. (Clement et al., 2019) Insertions and deletions were included if they altered a window of 1 bp on each side of the cleavage site. Allele frequencies were calculated by grouping together individual insertion or deletion events across all haplotypes on which they were found and summing them together. MMEJ deletions were predicted with Guido and matched with the deletions detected in pooled amplicon sequencing. This allowed the separation of detected MMEJ- and NHEJ-associated deletions at the target site. To test the difference between the MMEJ and NHEJ rates and the difference between MMEJ deletion frequencies at the target sites, independent two-sample *t*-test was performed with SciPy. (Virtanen et al., 2020) NumPy v1.18.4 (Harris et al., 2020) and Pandas v1.2.4 (Reback et al., 2021) was used for data processing. Plots were generated using Seaborn v0.11 and Matplotlib v3.3.

4.5.3 Guido

Guido was written as a Python package and consists of two main modules; *guido-build* is used for building the reference sequence that can subsequently be used with the main program *guido*. *guido-build* uses bowtie v1.2.3 (Langmead, 2010) to build the index of the reference sequence for the search of off-targets and *pyfaidx* (Shirley, 2014) to build the

index for querying the sequence for a genomic region where gRNA search is performed. Guido uses NumPy v1.18.4 (Harris et al., 2020), zarr v2.4 (Miles et al., 2020), jinja2 v2.11.2 , h5py v2.10, decipy v0.1 and scikit-allel v1.2.1 (Miles et al., 2019) for data processing and rendering output files.

Source code and instructions for installation and usage are available at:
https://github.com/nkran/guido

4.5.4 Target sites in genes of interest

Target sites for the three fertility genes and *dsx* were searched by using Guido. The following genes were used in search: *AGAP007280* (2L:44989399-44998080), *AGAP005958* (2L:24049746-24051543), *AGAP011377* (3L:22492551-22493453), *AGAP4050* (2R:48714423-48714830). Guido was run run with setting '*-f exons*', which limited the search only to exonic regions of the specified genes. Output from Guido was used for further processing and visualisations. gRNAs that were marked as inaccessible were omitted from further analysis and ranking. Genome accessibility data was obtained from Ag1000g phase 2 (The Anopheles gambiae 1000 Genomes Consortium, 2021, 2020). Variation dataset was obtained separately from Ag1000g phase 3 for the genes of interest (The Anopheles gambiae 1000 Genomes Consortium, 2021). For data analysis, Pandas v1.2.4, NumPy v1.18.4. and SciPy v1.5.2 were used (Harris et al., 2020; Reback et al., 2021; Virtanen et al., 2020). Plots were generated using Seaborn v0.11, Matplotlib v3.3 and FeaturesViewer v3.0.3. (Zulkower and Rosser, 2020)

Conclusions and Future Directions

Gene drives have shown great potential for the control of insect pests and vector populations. The field of gene drives has advanced considerably in the five years since I started the research project described in this thesis. We were able to overcome most of the technical barriers that seemed, at the time, to pose a serious risk to the succesfull application of this technology. The rational selection of target sites has been a key step in the development by surmounting gene drive resistance in *An. gambiae*.

The methodology described in this thesis could be used for applications beyond the generation of gene drive systems for the suppression of *An. gambiae* populations. The computational framework on which Guido was built can be repurposed to investigate biological data of other species and include them in the target site evaluation without sacrificing the flexibility to support the needs of researchers in other fields of population control. In order to achieve wider usage of such tools among the community of predominantly wet-lab biologists, a better and more user-friendly interface would likely be required.

Target site selection is an important step in the development of gene drives for both population suppression and population replacement. The basic requirement for the target site to produce favourable mutational outcomes is universal to any other genetic control application based on homing endonucleases. Further characterisation of mutational profiles, for example by high-throughput genetic screening in various genetic contexts, will be needed to fine-tune the predictive power. This will be required to correctly estimate the range of mutations in different tissues (somatic and germline) as well as any effects the sequence (and standing variation) around the cut site may have.

With costs of genome sequencing decreasing, insights into standing genetic variation of the targeted species are becoming more and more achievable. This contributes vital information to the process of target site selection. Including genetic variability data in the target site selection process is crucial for any gene drive application that aims to progress beyond the proof of principle stage.

As gene drives move further towards field application in the future, there will be a growing need to account for standing variation in local populations. If such variation is not accounted for it could prevent the activity of the endonuclease and hinder the spread of a drive. Researchers may even be able to take advantage of genetic variability at a target site. Differences between local populations could be exploited to limit the range of spread of gene drive systems and locally confine them.

The development of further resources providing information about existing variation in nature, as well as the computational tools that can integrate such information into the process of rational gene drive target selection, will play an important role in the success of genetic control strategies.

Bibliography

Aboobaker, A.A., Tomancak, P., Patel, N., Rubin, G.M., Lai, E.C., 2005. Drosophila microRNAs exhibit diverse spatial expression patterns during embryonic development. Proc Natl Acad Sci U S A 102, 18017–18022. https://doi.org/10.1073/pnas.0508823102

Achan, J., Talisuna, A.O., Erhart, A., Yeka, A., Tibenderana, J.K., Baliraine, F.N., Rosenthal, P.J., D'Alessandro, U., 2011. Quinine, an old anti-malarial drug in a modern world: role in the treatment of malaria. Malaria Journal 10, 144. https://doi.org/10.1186/1475-2875-10-144

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., George, R.A., Lewis, S.E., Richards, S., Ashburner, M., Henderson, S.N., Sutton, G.G., Wortman, J.R., Yandell, M.D., Zhang, Q., Chen, L.X., Brandon, R.C., Rogers, Y.-H.C., Blazej, R.G., Champe, M., Pfeiffer, B.D., Wan, K.H., Doyle, C., Baxter, E.G., Helt, G., Nelson, C.R., Gabor, G.L., Miklos, Abril, J.F., Agbayani, A., An, H.-J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R.M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E.M., Beeson, K.Y., Benos, P.V., Berman, B.P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M.R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K.C., Busam, D.A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J.M., Cawley, S., Dahlke, C., Davenport, L.B., Davies, P., Pablos, B. de, Delcher, A., Deng, Z., Mays, A.D., Dew, I., Dietz, S.M., Dodson, K., Doup, L.E., Downes, M., Dugan-Rocha, S., Dunkov, B.C., Dunn, P., Durbin, K.J., Evangelista, C.C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A.E., Garg, N.S., Gelbart, W.M., Glasser, K., Glodek, A., Gong, F., Gorrell, J.H., Gu, Z., Guan, P., Harris, M., Harris, N.L., Harvey, D., Heiman, T.J., Hernandez, J.R., Houck, J., Hostin, D., Houston, K.A., Howland, T.J., Wei, M.-H., Ibegwam, C., Jalali, M., Kalush, F., Karpen, G.H., Ke, Z., Kennison, J.A., Ketchum, K.A., Kimmel, B.E., Kodira, C.D., Kraft, C., Kravitz, S., Kulp, D., Lai, Z., Lasko, P., Lei, Y., Levitsky, A.A., Li, J., Li, Z., Liang, Y., Lin, X., Liu, X., Mattei, B., McIntosh, T.C., McLeod, M.P., McPherson, D., Merkulov, G., Milshina, N.V., Mobarry, C., Morris, J., Moshrefi, A., Mount, S.M., Moy, M., Murphy, B., Murphy, L., Muzny, D.M., Nelson, D.L., Nelson, D.R., Nelson, K.A., Nixon, K., Nusskern, D.R., Pacleb, J.M., Palazzolo, M., Pittman, G.S., Pan, S., Pollard, J., Puri, V., Reese, M.G., Reinert, K., Remington, K., Saunders, R.D.C., Scheeler, F., Shen, H., Shue, B.C., Sidén-Kiamos, I., Simpson, M., Skupski, M.P., Smith, T., Spier, E., Spradling, A.C., Stapleton, M., Strong, R., Sun, E., Svirskas, R., Tector, C., Turner, R., Venter, E., Wang, A.H., Wang, X., Wang, Z.-Y., Wassarman, D.A., Weinstock, G.M., Weissenbach, J., Williams, S.M., Woodage, T., Worley, K.C., Wu, D., Yang, S., Yao, Q.A., Ye, J., Yeh, R.-F., Zaveri, J.S., Zhan, M., Zhang, G., Zhao, Q., Zheng, L., Zheng, X.H., Zhong, F.N., Zhong, W., Zhou, X., Zhu, S., Zhu, X., Smith, H.O., Gibbs, R.A., Myers, E.W., Rubin, G.M., Venter, J.C., 2000. The Genome Sequence of Drosophila melanogaster. Science 287, 2185–2195. https://doi.org/10.1126/science.287.5461.2185

Ahmad, A., Robinson, A.R., Duensing, A., van Drunen, E., Beverloo, H.B., Weisberg, D.B., Hasty, P., Hoeijmakers, J.H.J., Niedernhofer, L.J., 2008. ERCC1-XPF Endonuclease Facilitates DNA Double-Strand Break Repair. Molecular and Cellular Biology 28, 5082–5092. https://doi.org/10.1128/MCB.00293-08

Akbari, O.S., Matzen, K.D., Marshall, J.M., Huang, H., Ward, C.M., Hay, B.A., 2013. A Synthetic Gene Drive System for Local, Reversible Modification and Suppression of Insect Populations. Current Biology 23, 671–677. https://doi.org/10.1016/j.cub.2013.02.059

Alphey, L., 2014. Genetic Control of Mosquitoes. Annual Review of Entomology 59, 205–224. https://doi.org/10.1146/annurev-ento-011613-162002

Alphey, L., Benedict, M., Bellini, R., Clark, G.G., Dame, D.A., Service, M.W., Dobson, S.L., 2009. Sterile-Insect Methods for Control of Mosquito-Borne Diseases: An Analysis. Vector-Borne and Zoonotic Diseases 10, 295–311. https://doi.org/10.1089/vbz.2009.0014

Anders, C., Niewoehner, O., Duerst, A., Jinek, M., 2014. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. Nature 513, 569–573. https://doi.org/10.1038/nature13579

Andreasen, M.H., Curtis, C.F., 2005. Optimal life stage for radiation sterilization of Anopheles males and their fitness for release. Medical and Veterinary Entomology 19, 238–244. https://doi.org/10.1111/j.1365-2915.2005.00565.x
Ariey, F., Witkowski, B., Amaratunga, C., Beghain, J., Langlois, A.-C., Khim, N., Kim, S., Duru, V., Bouchier, C., Ma, L., Lim, P., Leang, R., Duong, S., Sreng, S., Suon, S., Chuor, C.M., Bout, D.M., Ménard, S., Rogers, W.O., Genton, B., Fandeur, T., Miotto, O., Ringwald, P., Le Bras, J., Berry, A., Barale, J.-C., Fairhurst, R.M., Benoit-Vical, F., Mercereau-Puijalon, O., Ménard, D., 2014. A molecular marker of artemisinin-resistant Plasmodium falciparum malaria. Nature 505, 50–55. https://doi.org/10.1038/nature12876
Aryan, A., Anderson, M.A.E., Myles, K.M., Adelman, Z.N., 2013. Germline excision of transgenes in Aedes aegypti by homing endonucleases. Sci Rep 3, 1603. https://doi.org/10.1038/srep01603
Ashley, E.A., Dhorda, M., Fairhurst, R.M., Amaratunga, C., Lim, P., Suon, S., Sreng, S., Anderson, J.M., Mao, S., Sam, B., Sopha, C., Chuor, C.M., Nguon, C., Sovannaroth, S., Pukrittayakamee, S., Jittamala, P., Chotivanich, K., Chutasmit, K., Suchatsoonthorn, C., Runcharoen, R., Hien, T.T., Thuy-Nhien, N.T., Thanh, N.V., Phu, N.H., Htut, Y., Han, K.-T., Aye, K.H., Mokuolu, O.A., Olaosebikan, R.R., Folaranmi, O.O., Mayxay, M., Khanthavong, M., Hongvanthong, B., Newton, P.N., Onyamboko, M.A., Fanello, C.I., Tshefu, A.K., Mishra, N., Valecha, N., Phyo, A.P., Nosten, F., Yi, P., Tripura, R., Borrmann, S., Bashraheil, M., Peshu, J., Faiz, M.A., Ghose, A., Hossain, M.A., Samad, R., Rahman, M.R., Hasan, M.M., Islam, A., Miotto, O., Amato, R., MacInnis, B., Stalker, J., Kwiatkowski, D.P., Bozdech, Z., Jeeyapant, A., Cheah, P.Y., Sakulthaew, T., Chalk, J., Intharabut, B., Silamut, K., Lee, S.J., Vihokhern, B., Kunasol, C., Imwong, M., Tarning, J., Taylor, W.J., Yeung, S., Woodrow, C.J., Flegg, J.A., Das, D., Smith, J., Venkatesan, M., Plowe, C.V., Stepniewska, K., Guerin, P.J., Dondorp, A.M., Day, N.P., White, N.J., 2014. Spread of Artemisinin Resistance in Plasmodium falciparum Malaria. New England Journal of Medicine 371, 411–423. https://doi.org/10.1056/NEJMoa1314981
Ata, H., Ekstrom, T.L., Martínez-Gálvez, G., Mann, C.M., Dvornikov, A.V., Schaefbauer, K.J., Ma, A.C., Dobbs, D., Clark, K.J., Ekker, S.C., 2018. Robust activation of microhomology-mediated end joining for precision gene editing applications. PLOS Genetics 14, e1007652. https://doi.org/10.1371/journal.pgen.1007652
Bae, S., Kweon, J., Kim, H.S., Kim, J.-S., 2014a. Microhomology-based choice of Cas9 nuclease target sites. Nature Methods 11, 705–706. https://doi.org/10.1038/nmeth.3015
Bae, S., Park, J., Kim, J.-S., 2014b. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. Bioinformatics 30, 1473–1475. https://doi.org/10.1093/bioinformatics/btu048
Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., Horvath, P., 2007. CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. Science 315, 1709–1712. https://doi.org/10.1126/science.1138140
Bartoloni, A., Zammarchi, L., 2012. Clinical Aspects of Uncomplicated and Severe Malaria. Mediterr J Hematol Infect Dis 4. https://doi.org/10.4084/MJHID.2012.026
Beeman, R., Friesen, K., Denell, R., 1992. Maternal-effect selfish genes in flour beetles. Science 256, 89–92.
Beeman, R.W., Friesen, K.S., 1999. Properties and natural occurrence of maternal-effect selfish genes (‘ Medea ’ factors) in the Red Flour Beetle, Tribolium castaneum. Heredity 82, 529–534. https://doi.org/10.1038/sj.hdy.6885150
Bernardini, F., Galizi, R., Wunderlich, M., Taxiarchi, C., Kranjc, N., Kyrou, K., Hammond, A., Nolan, T., Lawniczak, M.N.K., Papathanos, P.A., Crisanti, A., Windbichler, N., 2017. Cross-Species Y Chromosome Function Between Malaria Vectors of the Anopheles gambiae Species Complex. Genetics 207, 729–740. https://doi.org/10.1534/genetics.117.300221
Bernhart, S.H., Tafer, H., Mückstein, U., Flamm, C., Stadler, P.F., Hofacker, I.L., 2006. Partition function and base pairing probabilities of RNA heterodimers. Algorithms Mol Biol 1, 3. https://doi.org/10.1186/1748-7188-1-3
Beshir, K., Sutherland, C.J., Merinopoulos, I., Durrani, N., Leslie, T., Rowland, M., Hallett, R.L., 2010. Amodiaquine Resistance in Plasmodium falciparum Malaria in Afghanistan Is Associated with the pfcrt SVMNT Allele at Codons 72 to 76. Antimicrobial Agents and Chemotherapy 54, 3714–3716. https://doi.org/10.1128/AAC.00358-10
Bhatt, S., Weiss, D.J., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., Battle, K.E., Moyes, C.L., Henry, A., Eckhoff, P.A., Wenger, E.A., Briët, O., Penny, M.A., Smith, T.A., Bennett, A.,

Yukich, J., Eisele, T.P., Griffin, J.T., Fergus, C.A., Lynch, M., Lindgren, F., Cohen, J.M., Murray, C.L.J., Smith, D.L., Hay, S.I., Cibulskis, R.E., Gething, P.W., 2015. The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015. Nature 526, 207–211. https://doi.org/10.1038/nature15535

Birnbaum, J., Scharf, S., Schmidt, S., Jonscher, E., Hoeijmakers, W.A.M., Flemming, S., Toenhake, C.G., Schmitt, M., Sabitzki, R., Bergmann, B., Fröhlke, U., Mesén-Ramírez, P., Soares, A.B., Herrmann, H., Bártfai, R., Spielmann, T., 2020. A Kelch13-defined endocytosis pathway mediates artemisinin resistance in malaria parasites. Science 367, 51–59. https://doi.org/10.1126/science.aax4735

Biswas, A., Staals, R.H.J., Morales, S.E., Fineran, P.C., Brown, C.M., 2016. CRISPRDetect: A flexible algorithm to define CRISPR arrays. BMC Genomics 17. https://doi.org/10.1186/s12864-016-2627-0

Bolotin, A., Quinquis, B., Sorokin, A., Ehrlich, S.D., 2005. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology (Reading, Engl.) 151, 2551–2561. https://doi.org/10.1099/mic.0.28048-0

Bozas, A., Beumer, K.J., Trautman, J.K., Carroll, D., 2009. Genetic analysis of zinc-finger nuclease-induced gene targeting in Drosophila. Genetics 182, 641–651. https://doi.org/10.1534/genetics.109.101329

Bradford, J., Perrin, D., 2019. A benchmark of computational CRISPR-Cas9 guide design methods. PLOS Computational Biology 15, e1007274. https://doi.org/10.1371/journal.pcbi.1007274

Braff, J.L., Yaung, S.J., Esvelt, K.M., Church, G.M., 2016. Characterization of Cas9–Guide RNA Orthologs. Cold Spring Harb Protoc 2016, pdb.top086793. https://doi.org/10.1101/pdb.top086793

Burstein, D., Harrington, L.B., Strutt, S.C., Probst, A.J., Anantharaman, K., Thomas, B.C., Doudna, J.A., Banfield, J.F., 2017. New CRISPR–Cas systems from uncultivated microbes. Nature 542, 237–241. https://doi.org/10.1038/nature21059

Burt, A., 2003. Site-specific selfish genes as tools for the control and genetic engineering of natural populations. Proceedings of the Royal Society of London B: Biological Sciences 270, 921–928. https://doi.org/10.1098/rspb.2002.2319

Cancellieri, S., Canver, M.C., Bombieri, N., Giugno, R., Pinello, L., 2020. CRISPRitz: rapid, high-throughput and variant-aware in silico off-target site identification for CRISPR genome editing. Bioinformatics 36, 2001–2008. https://doi.org/10.1093/bioinformatics/btz867

Canver, M.C., Haeussler, M., Bauer, D.E., Orkin, S.H., Sanjana, N.E., Shalem, O., Yuan, G.-C., Zhang, F., Concordet, J.-P., Pinello, L., 2018. Integrated design, execution, and analysis of arrayed and pooled CRISPR genome-editing experiments. Nature Protocols 13, 946–986. https://doi.org/10.1038/nprot.2018.005

Canver, M.C., Smith, E.C., Sher, F., Pinello, L., Sanjana, N.E., Shalem, O., Chen, D.D., Schupp, P.G., Vinjamur, D.S., Garcia, S.P., Luc, S., Kurita, R., Nakamura, Y., Fujiwara, Y., Maeda, T., Yuan, G.-C., Zhang, F., Orkin, S.H., Bauer, D.E., 2015. BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. Nature 527, 192–197. https://doi.org/10.1038/nature15521

Carballar-Lejarazú, R., James, A.A., 2017. Population modification of Anopheline species to control malaria transmission. Pathogens and Global Health 111, 424–435. https://doi.org/10.1080/20477724.2018.1427192

Carballar-Lejarazú, R., Jasinskiene, N., James, A.A., 2013. Exogenous gypsy insulator sequences modulate transgene expression in the malaria vector mosquito, Anopheles stephensi. PNAS 110, 7176–7181.

Carrami, E.M., Eckermann, K.N., Ahmed, H.M.M., C, H.M.S., Dippel, S., Marshall, J.M., Wimmer, E.A., 2018. Consequences of resistance evolution in a Cas9-based sex conversion-suppression gene drive for insect pest management. PNAS 115, 6189–6194. https://doi.org/10.1073/pnas.1713825115

Carvalho, D.O., McKemey, A.R., Garziera, L., Lacroix, R., Donnelly, C.A., Alphey, L., Malavasi, A., Capurro, M.L., 2015. Suppression of a Field Population of Aedes aegypti in Brazil by Sustained Release of Transgenic Male Mosquitoes. PLOS Neglected Tropical Diseases 9, e0003864. https://doi.org/10.1371/journal.pntd.0003864

Casini, A., Olivieri, M., Petris, G., Montagna, C., Reginato, G., Maule, G., Lorenzin, F., Prandi, D., Romanel, A., Demichelis, F., Inga, A., Cereseto, A., 2018. A highly specific SpCas9 variant

is identified by in vivo screening in yeast. Nat. Biotechnol. 36, 265–271. https://doi.org/10.1038/nbt.4066
Catteruccia, F., Nolan, T., Loukeris, T.G., Blass, C., Savakis, C., Kafatos, F.C., Crisanti, A., 2000. Stable germline transformation of the malaria mosquito Anopheles stephensi. Nature 405, 959–962. https://doi.org/10.1038/35016096
Ceapa, C., Davids, M., Ritari, J., Lambert, J., Wels, M., Douillard, F.P., Smokvina, T., de Vos, W.M., Knol, J., Kleerebezem, M., 2016. The Variable Regions of Lactobacillus rhamnosus Genomes Reveal the Dynamic Evolution of Metabolic and Host-Adaptation Repertoires. Genome Biol Evol 8, 1889–1905. https://doi.org/10.1093/gbe/evw123
Champer, J., Buchman, A., Akbari, O.S., 2016. Cheating evolution: engineering gene drives to manipulate the fate of wild populations. Nature Reviews Genetics 17, 146–159. https://doi.org/10.1038/nrg.2015.34
Champer, J., Liu, J., Oh, S.Y., Reeves, R., Luthra, A., Oakes, N., Clark, A.G., Messer, P.W., 2018. Reducing resistance allele formation in CRISPR gene drive. PNAS 115, 5522–5527. https://doi.org/10.1073/pnas.1720354115
Champer, J., Reeves, R., Oh, S.Y., Liu, C., Liu, J., Clark, A.G., Messer, P.W., 2017. 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. https://doi.org/10.1371/journal.pgen.1006796
Champer, S.E., Oh, S.Y., Liu, C., Wen, Z., Clark, A.G., Messer, P.W., Champer, J., 2020. Computational and experimental performance of CRISPR homing gene drive strategies with multiplexed gRNAs. Science Advances 6, eaaz0525. https://doi.org/10.1126/sciadv.aaz0525
Chan, S.H., Yu, A.M., McVey, M., 2010. Dual Roles for DNA Polymerase Theta in Alternative End-Joining Repair of Double-Strand Breaks in Drosophila. PLOS Genetics 6, e1001005. https://doi.org/10.1371/journal.pgen.1001005
Chan, Y.-S., Naujoks, D.A., Huen, D.S., Russell, S., 2011. Insect Population Control by Homing Endonuclease-Based Gene Drive: An Evaluation in Drosophila melanogaster. Genetics 188, 33–44. https://doi.org/10.1534/genetics.111.127506
Chan, Y.-S., Takeuchi, R., Jarjour, J., Huen, D.S., Stoddard, B.L., Russell, S., 2013. The Design and In Vivo Evaluation of Engineered I-OnuI-Based Enzymes for HEG Gene Drive. PLOS ONE 8, e74254. https://doi.org/10.1371/journal.pone.0074254
Chang, H.H.Y., Pannunzio, N.R., Adachi, N., Lieber, M.R., 2017. Non-homologous DNA end joining and alternative pathways to double-strand break repair. Nature Reviews Molecular Cell Biology 18, 495–506. https://doi.org/10.1038/nrm.2017.48
Charlesworth, B., 2009. Effective population size and patterns of molecular evolution and variation. Nat Rev Genet 10, 195–205. https://doi.org/10.1038/nrg2526
Charlesworth, B., Morgan, M.T., Charlesworth, D., 1993. The effect of deleterious mutations on neutral molecular variation. Genetics 134, 1289–1303.
Charlwood, J.D., Jones, M.D.R., 1979. Mating behaviour in the mosquito, Anopheles gambiae s.1.save. Physiological Entomology 4, 111–120. https://doi.org/10.1111/j.1365-3032.1979.tb00185.x
Chen, C.-H., Huang, H., Ward, C.M., Su, J.T., Schaeffer, L.V., Guo, M., Hay, B.A., 2007. A Synthetic Maternal-Effect Selfish Genetic Element Drives Population Replacement in Drosophila. Science 316, 597–600. https://doi.org/10.1126/science
Chen, J.S., Dagdas, Y.S., Kleinstiver, B.P., Welch, M.M., Sousa, A.A., Harrington, L.B., Sternberg, S.H., Joung, J.K., Yildiz, A., Doudna, J.A., 2017. Enhanced proofreading governs CRISPR–Cas9 targeting accuracy. Nature advance online publication. https://doi.org/10.1038/nature24268
Chen, W., McKenna, A., Schreiber, J., Haeussler, M., Yin, Y., Agarwal, V., Noble, W.S., Shendure, J., 2019. Massively parallel profiling and predictive modeling of the outcomes of CRISPR/Cas9-mediated double-strand break repair. Nucleic Acids Research 47, 7989–8003. https://doi.org/10.1093/nar/gkz487
Chibani-Chennoufi, S., Bruttin, A., Dillmann, M.-L., Brüssow, H., 2004. Phage-Host Interaction: an Ecological Perspective. Journal of Bacteriology 186, 3677–3686. https://doi.org/10.1128/JB.186.12.3677-3686.2004

Cho, S.W., Kim, S., Kim, J.M., Kim, J.-S., 2013. Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. Nature Biotechnology 31, 230. https://doi.org/10.1038/nbt.2507
Claycomb, J.M., Benasutti, M., Bosco, G., Fenger, D.D., Orr-Weaver, T.L., 2004. Gene Amplification as a Developmental Strategy: Isolation of Two Developmental Amplicons in Drosophila. Developmental Cell 6, 145–155. https://doi.org/10.1016/S1534-5807(03)00398-8
Clejan, I., Boerckel, J., Ahmed, S., 2006. Developmental Modulation of Nonhomologous End Joining in Caenorhabditis elegans. Genetics 173, 1301–1317. https://doi.org/10.1534/genetics.106.058628
Clement, K., Rees, H., Canver, M.C., Gehrke, J.M., Farouni, R., Hsu, J.Y., Cole, M.A., Liu, D.R., Joung, J.K., Bauer, D.E., Pinello, L., 2019. CRISPResso2 provides accurate and rapid genome editing sequence analysis. Nat Biotechnol 37, 224–226. https://doi.org/10.1038/s41587-019-0032-3
Coetzee, M., Hunt, R.H., Wilkerson, R., Torre, A.D., Coulibaly, M.B., Besansky, N.J., 2013. Anopheles coluzzii and Anopheles amharicus , new members of the Anopheles gambiae complex. Zootaxa 3619, 246–274. https://doi.org/10.11646/zootaxa.3619.3.2
Collias, D., Beisel, C.L., 2021. CRISPR technologies and the search for the PAM-free nuclease. Nat Commun 12, 555. https://doi.org/10.1038/s41467-020-20633-y
Coluzzi, M., Sabatini, A., Petrarca, V., Di Deco, M.A., 1979. Chromosomal differentiation and adaptation to human environments in the Anopheles gambiae complex. Trans R Soc Trop Med Hyg 73, 483–497. https://doi.org/10.1016/0035-9203(79)90036-1
Concordet, J.-P., Haeussler, M., 2018. CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. Nucleic Acids Res 46, W242–W245. https://doi.org/10.1093/nar/gky354
Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., Zhang, F., 2013. Multiplex Genome Engineering Using CRISPR/Cas Systems. Science 339, 819–823. https://doi.org/10.1126/science.1231143
Cooper, G.M., Brown, C.D., 2008. Qualifying the relationship between sequence conservation and molecular function. Genome Res. 18, 201–205. https://doi.org/10.1101/gr.7205808
Cowman, A.F., Berry, D., Baum, J., 2012. The cellular and molecular basis for malaria parasite invasion of the human red blood cell. J Cell Biol 198, 961–971. https://doi.org/10.1083/jcb.201206112
Cowman, A.F., Galatis, D., Thompson, J.K., 1994. Selection for mefloquine resistance in Plasmodium falciparum is linked to amplification of the pfmdr1 gene and cross-resistance to halofantrine and quinine. Proc Natl Acad Sci U S A 91, 1143–1147. https://doi.org/10.1073/pnas.91.3.1143
Crawley, A.B., Henriksen, E.D., Stout, E., Brandt, K., Barrangou, R., 2018. Characterizing the activity of abundant, diverse and active CRISPR-Cas systems in lactobacilli. Scientific Reports 8, 11544. https://doi.org/10.1038/s41598-018-29746-3
Crooks, G.E., Hon, G., Chandonia, J.-M., Brenner, S.E., 2004. WebLogo: a sequence logo generator. Genome Res. 14, 1188–1190. https://doi.org/10.1101/gr.849004
Curtis, C., 1968. Possible use of translocations to fix desirable genes in insect pest populations. Nature 218, 368–369.
D, B., J, K., K, N.-B., A, T., S, C., T, H., 1996. Quinine-tetracycline for multidrug resistant falciparum malaria. Southeast Asian J Trop Med Public Health 27, 15–18.
Datoo, M.S., Natama, M.H., Somé, A., Traoré, O., Rouamba, T., Bellamy, D., Yameogo, P., Valia, D., Tegneri, M., Ouedraogo, F., Soma, R., Sawadogo, S., Sorgho, F., Derra, K., Rouamba, E., Orindi, B., Lopez, F.R., Flaxman, A., Cappuccini, F., Kailath, R., Elias, S., Mukhopadhyay, E., Noe, A., Cairns, M., Lawrie, A., Roberts, R., Valéa, I., Sorgho, H., Williams, N., Glenn, G., Fries, L., Reimer, J., Ewer, K.J., Shaligram, U., Hill, A.V.S., Tinto, H., 2021. Efficacy of a low-dose candidate malaria vaccine, R21 in adjuvant Matrix-M, with seasonal administration to children in Burkina Faso: a randomised controlled trial. The Lancet 397, 1809–1818. https://doi.org/10.1016/S0140-6736(21)00943-0
Davidson, G., 1962. Anopheles gambiae complex. Nature 196, 907. https://doi.org/10.1038/196907a0
Davies, T.G.E., Field, L.M., Usherwood, P.N.R., Williamson, M.S., 2007. DDT, pyrethrins, pyrethroids and insect sodium channels. IUBMB Life 59, 151–162. https://doi.org/10.1080/15216540701352042

Deng, S.K., Gibb, B., de Almeida, M.J., Greene, E.C., Symington, L.S., 2014. RPA antagonizes microhomology-mediated repair of DNA double-strand breaks. Nat Struct Mol Biol 21, 405–412. https://doi.org/10.1038/nsmb.2786
Deredec, A., Burt, A., Godfray, H.C.J., 2008. The Population Genetics of Using Homing Endonuclease Genes in Vector and Pest Management. Genetics 179, 2013–2026. https://doi.org/10.1534/genetics.108.089037
Derua, Y.A., Alifrangis, M., Magesa, S.M., Kisinza, W.N., Simonsen, P.E., 2015. Sibling species of the Anopheles funestus group, and their infection with malaria and lymphatic filarial parasites, in archived and newly collected specimens from northeastern Tanzania. Malaria Journal 14, 104. https://doi.org/10.1186/s12936-015-0616-4
Deveau, H., Barrangou, R., Garneau, J.E., Labonté, J., Fremaux, C., Boyaval, P., Romero, D.A., Horvath, P., Moineau, S., 2008. Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. Journal of bacteriology 190, 1390–1400.
Diabaté, A., Dabire, R.K., Kengne, P., Brengues, C., Baldet, T., Ouari, A., Simard, F., Lehmann, T., 2006. Mixed swarms of the molecular M and S forms of Anopheles gambiae (Diptera: Culicidae) in sympatric area from Burkina Faso. J Med Entomol 43, 480–483. https://doi.org/10.1603/0022-2585(2006)43[480:msotmm]2.0.co;2
DiCarlo, J.E., Chavez, A., Dietz, S.L., Esvelt, K.M., Church, G.M., 2015. Safeguarding CRISPR-Cas9 gene drives in yeast. Nature Biotechnology 33, 1250–1255. https://doi.org/10.1038/nbt.3412
Díez-Villaseñor, C., Guzmán, N.M., Almendros, C., García-Martínez, J., Mojica, F.J.M., 2013. CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of Escherichia coli. RNA Biol 10, 792–802. https://doi.org/10.4161/rna.24023
Doench, J.G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E.W., Donovan, K.F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., Virgin, H.W., Listgarten, J., Root, D.E., 2016. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat. Biotechnol. 34, 184–191. https://doi.org/10.1038/nbt.3437
Dong, Y., Simões, M.L., Marois, E., Dimopoulos, G., 2018. CRISPR/Cas9 -mediated gene knockout of Anopheles gambiae FREP1 suppresses malaria parasite infection. PLOS Pathogens 14, e1006898. https://doi.org/10.1371/journal.ppat.1006898
Downs, J.A., Jackson, S.P., 2004. A means to a DNA end: the many roles of Ku. Nat Rev Mol Cell Biol 5, 367–378. https://doi.org/10.1038/nrm1367
Duffy, P.E., Patrick Gorres, J., 2020. Malaria vaccines since 2000: progress, priorities, products. npj Vaccines 5, 1–9. https://doi.org/10.1038/s41541-020-0196-3
Dyck, V.A., Hendrichs, J., Robinson, A.S., 2021. Sterile Insect Technique : Principles And Practice In Area-Wide Integrated Pest Management. Taylor & Francis.
Eggenhofer, F., Tafer, H., Stadler, P.F., Hofacker, I.L., 2011. RNApredator: fast accessibility-based prediction of sRNA targets. Nucleic Acids Res 39, W149–W154. https://doi.org/10.1093/nar/gkr467
Egli, D., Hafen, E., Schaffner, W., 2004. An Efficient Method to Generate Chromosomal Rearrangements by Targeted DNA Double-Strand Breaks in Drosophila melanogaster. Genome Res. 14, 1382–1393. https://doi.org/10.1101/gr.2279804
Enkerlin, W., Gutiérrez-Ruelas, J.M., Cortes, A.V., Roldan, E.C., Midgarden, D., Lira, E., López, J.L.Z., Hendrichs, J., Liedo, P., Arriaga, F.J.T., 2015. Area Freedom in Mexico from Mediterranean Fruit Fly (Diptera: Tephritidae): A Review of Over 30 Years of a Successful Containment Program Using an Integrated Area-Wide SIT Approach. flen 98, 665–681. https://doi.org/10.1653/024.098.0242
Esvelt, K.M., Smidler, A.L., Catteruccia, F., Church, G.M., 2014. Concerning RNA-guided gene drives for the alteration of wild populations. eLife 3, e03401. https://doi.org/10.7554/eLife.03401
Fonfara, I., Le Rhun, A., Chylinski, K., Makarova, K.S., Lécrivain, A.-L., Bzdrenga, J., Koonin, E.V., Charpentier, E., 2014. Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas systems. Nucleic Acids Res 42, 2577–2590. https://doi.org/10.1093/nar/gkt1074
Fontaine, M.C., Pease, J.B., Steele, A., Waterhouse, R.M., Neafsey, D.E., Sharakhov, I.V., Jiang, X., Hall, A.B., Catteruccia, F., Kakani, E., Mitchell, S.N., Wu, Y.-C., Smith, H.A., Love, R.R., Lawniczak, M.K., Slotman, M.A., Emrich, S.J., Hahn, M.W., Besansky, N.J., 2015. Extensive

introgression in a malaria vector species complex revealed by phylogenomics. Science 347. https://doi.org/10.1126/science.1258524
Fuglede, B., Topsoe, F., 2004. Jensen-Shannon divergence and Hilbert space embedding, in: International Symposium OnInformation Theory, 2004. ISIT 2004. Proceedings. Presented at the International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings., pp. 31-. https://doi.org/10.1109/ISIT.2004.1365067
G. James, R., J. Ellison, C., P. Crutchfield, J., 2018. dit: a Python package for discrete information theory. JOSS 3, 738. https://doi.org/10.21105/joss.00738
Galizi, R., Doyle, L.A., Menichelli, M., Bernardini, F., Deredec, A., Burt, A., Stoddard, B.L., Windbichler, N., Crisanti, A., 2014. A synthetic sex ratio distortion system for the control of the human malaria mosquito. Nature Communications 5, 3977. https://doi.org/10.1038/ncomms4977
Galizi, R., Hammond, A., Kyrou, K., Taxiarchi, C., Bernardini, F., O'Loughlin, S.M., Papathanos, P.-A., Nolan, T., Windbichler, N., Crisanti, A., 2016. A CRISPR-Cas9 sex-ratio distortion system for genetic control. Scientific Reports 6, 31139. https://doi.org/10.1038/srep31139
Gantz, V.M., Bier, E., 2015. The mutagenic chain reaction: A method for converting heterozygous to homozygous mutations. Science 348, 442–444. https://doi.org/10.1126/science.aaa5945
Gantz, V.M., Jasinskiene, N., Tatarenkova, O., Fazekas, A., Macias, V.M., Bier, E., James, A.A., 2015. Highly efficient Cas9-mediated gene drive for population modification of the malaria vector mosquito Anopheles stephensi. PNAS 112, E6736–E6743. https://doi.org/10.1073/pnas.1521077112
Garrood, W.T., Kranjc, N., Petri, K., Kim, D.Y., Guo, J.A., Hammond, A.M., Morianou, I., Pattanayak, V., Joung, J.K., Crisanti, A., Simoni, A., 2021. Analysis of off-target effects in CRISPR-based gene drives in the human malaria mosquito. PNAS 118. https://doi.org/10.1073/pnas.2004838117
Gasiunas, G., Young, J.K., Karvelis, T., Kazlauskas, D., Urbaitis, T., Jasnauskaite, M., Grusyte, M.M., Paulraj, S., Wang, P.-H., Hou, Z., Dooley, S.K., Cigan, M., Alarcon, C., Chilcoat, N.D., Bigelyte, G., Curcuru, J.L., Mabuchi, M., Sun, Z., Fuchs, R.T., Schildkraut, E., Weigele, P.R., Jack, W.E., Robb, G.B., Venclovas, Č., Siksnys, V., 2020. A catalogue of biochemically diverse CRISPR-Cas9 orthologs. Nat Commun 11, 5512. https://doi.org/10.1038/s41467-020-19344-1
Giraldo-Calderon, G.I., Emrich, S.J., MacCallum, R.M., Maslen, G., Dialynas, E., Topalis, P., Ho, N., Gesing, S., Madey, G., Collins, F.H., Lawson, D., 2015. VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. Nucleic acids research 43, D707-13. https://doi.org/10.1093/nar/gku1117
Gloor, G.B., Moretti, J., Mouyal, J., Keeler, K.J., 2000. Distinct P-element excision products in somatic and germline cells of Drosophila melanogaster. Genetics 155, 1821–1830. https://doi.org/10.1093/genetics/155.4.1821
Gong, P., Epton, M.J., Fu, G., Scaife, S., Hiscox, A., Condon, K.C., Condon, G.C., Morrison, N.I., Kelly, D.W., Dafa'alla, T., Coleman, P.G., Alphey, L., 2005. A dominant lethal genetic system for autocidal control of the Mediterranean fruitfly. Nature Biotechnology 23, 453–456. https://doi.org/10.1038/nbt1071
Grajcarek, J., Monlong, J., Nishinaka-Arai, Y., Nakamura, M., Nagai, M., Matsuo, S., Lougheed, D., Sakurai, H., Saito, M.K., Bourque, G., Woltjen, K., 2019. Genome-wide microhomologies enable precise template-free editing of biologically relevant deletion mutations. Nat Commun 10, 4856. https://doi.org/10.1038/s41467-019-12829-8
Grunwald, H.A., Gantz, V.M., Poplawski, G., Xu, X.-R.S., Bier, E., Cooper, K.L., 2019. Super-Mendelian inheritance mediated by CRISPR–Cas9 in the female mouse germline. Nature 566, 105–109. https://doi.org/10.1038/s41586-019-0875-2
Hall, A.B., Papathanos, P.-A., Sharma, A., Cheng, C., Akbari, O.S., Assour, L., Bergman, N.H., Cagnetti, A., Crisanti, A., Dottorini, T., Fiorentini, E., Galizi, R., Hnath, J., Jiang, X., Koren, S., Nolan, T., Radune, D., Sharakhova, M.V., Steele, A., Timoshevskiy, V.A., Windbichler, N., Zhang, S., Hahn, M.W., Phillippy, A.M., Emrich, S.J., Sharakhov, I.V., Tu, Z.J., Besansky, N.J., 2016. Radical remodeling of the Y chromosome in a recent radiation of malaria mosquitoes. Proceedings of the National Academy of Sciences 113, E2114–E2123. https://doi.org/10.1073/pnas.1525164113

Hamilton, W.D., 1967. Extraordinary Sex Ratios. Science 156, 477–488. https://doi.org/10.1126/science.156.3774.477
Hammond, A., Galizi, R., 2017. Gene drives to fight malaria: current state and future directions. Pathogens and Global Health 111, 412–423. https://doi.org/10.1080/20477724.2018.1438880
Hammond, A., Galizi, R., Kyrou, K., Simoni, A., Siniscalchi, C., Katsanos, D., Gribble, M., Baker, D., Marois, E., Russell, S., Burt, A., Windbichler, N., Crisanti, A., Nolan, T., 2016. A CRISPR-Cas9 gene drive system targeting female reproduction in the malaria mosquito vector Anopheles gambiae. Nat Biotech 34, 78–83. https://doi.org/10.1038/nbt.3439
Hammond, A., Karlsson, X., Morianou, I., Kyrou, K., Beaghton, A., Gribble, M., Kranjc, N., Galizi, R., Burt, A., Crisanti, A., Nolan, T., 2021a. Regulating the expression of gene drives is key to increasing their invasive potential and the mitigation of resistance. PLOS Genetics 17, e1009321. https://doi.org/10.1371/journal.pgen.1009321
Hammond, A., Kyrou, K., Bruttini, M., North, A., Galizi, R., Karlsson, X., Kranjc, N., Carpi, F.M., D'Aurizio, R., Crisanti, A., Nolan, T., 2017. The creation and selection of mutations resistant to a gene drive over multiple generations in the malaria mosquito. PLOS Genetics 13, e1007039. https://doi.org/10.1371/journal.pgen.1007039
Hammond, A., Pollegioni, P., Persampieri, T., North, A., Minuz, R., Trusso, A., Bucci, A., Kyrou, K., Morianou, I., Simoni, A., Nolan, T., Müller, R., Crisanti, A., 2021b. Gene-drive suppression of mosquito populations in large cages as a bridge between lab and field. Nat Commun 12, 4589. https://doi.org/10.1038/s41467-021-24790-6
Harris, A.F., McKemey, A.R., Nimmo, D., Curtis, Z., Black, I., Morgan, S.A., Oviedo, M.N., Lacroix, R., Naish, N., Morrison, N.I., Collado, A., Stevenson, J., Scaife, S., Dafa'alla, T., Fu, G., Phillips, C., Miles, A., Raduan, N., Kelly, N., Beech, C., Donnelly, C.A., Petrie, W.D., Alphey, L., 2012. Successful suppression of a field mosquito population by sustained release of engineered male mosquitoes. Nat Biotechnol 30, 828–830. https://doi.org/10.1038/nbt.2350
Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E., 2020. Array programming with NumPy. Nature 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2
Heigwer, F., Kerr, G., Boutros, M., 2014. E-CRISP: fast CRISPR target site identification. Nat Methods 11, 122–123. https://doi.org/10.1038/nmeth.2812
Heler, R., Samai, P., Modell, J.W., Weiner, C., Goldberg, G.W., Bikard, D., Marraffini, L.A., 2015. Cas9 specifies functional viral targets during CRISPR–Cas adaptation. Nature 519, 199–202. https://doi.org/10.1038/nature14245
Hickey, W.A., Craig, G.B., 1966. Genetic distortion of sex ratio in a mosquito, Aedes aegypti. Genetics 53, 1177–1196.
Hirano, H., Gootenberg, J.S., Horii, T., Abudayyeh, O.O., Kimura, M., Hsu, P.D., Nakane, T., Ishitani, R., Hatada, I., Zhang, F., Nishimasu, H., Nureki, O., 2016. Structure and Engineering of Francisella novicida Cas9. Cell 164, 950–961. https://doi.org/10.1016/j.cell.2016.01.039
Hodgkinson, A., Eyre-Walker, A., 2011. Variation in the mutation rate across mammalian genomes. Nat Rev Genet 12, 756–766. https://doi.org/10.1038/nrg3098
Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R., Wincker, P., Clark, A.G., Ribeiro, J.C., Wides, R., Salzberg, S.L., Loftus, B., Yandell, M., Majoros, W.H., Rusch, D.B., Lai, Z., Kraft, C.L., Abril, J.F., Anthouard, V., Arensburger, P., Atkinson, P.W., Baden, H., Berardinis, V. de, Baldwin, D., Benes, V., Biedler, J., Blass, C., Bolanos, R., Boscus, D., Barnstead, M., Cai, S., Center, A., Chatuverdi, K., Christophides, G.K., Chrystal, M.A., Clamp, M., Cravchik, A., Curwen, V., Dana, A., Delcher, A., Dew, I., Evans, C.A., Flanigan, M., Grundschober-Freimoser, A., Friedli, L., Gu, Z., Guan, P., Guigo, R., Hillenmeyer, M.E., Hladun, S.L., Hogan, J.R., Hong, Y.S., Hoover, J., Jaillon, O., Ke, Z., Kodira, C., Kokoza, E., Koutsos, A., Letunic, I., Levitsky, A., Liang, Y., Lin, J.-J., Lobo, N.F., Lopez, J.R., Malek, J.A., McIntosh, T.C., Meister, S., Miller, J., Mobarry, C., Mongin, E., Murphy, S.D., O'Brochta, D.A., Pfannkoch, C., Qi, R., Regier, M.A., Remington, K., Shao, H., Sharakhova, M.V., Sitter, C.D., Shetty, J., Smith, T.J., Strong, R., Sun, J., Thomasova, D., Ton, L.Q., Topalis, P., Tu, Z., Unger, M.F., Walenz, B., Wang, A., Wang, J., Wang, M., Wang, X., Woodford, K.J.,

Wortman, J.R., Wu, M., Yao, A., Zdobnov, E.M., Zhang, H., Zhao, Q., Zhao, S., Zhu, S.C., Zhimulev, I., Coluzzi, M., Torre, A. della, Roth, C.W., Louis, C., Kalush, F., Mural, R.J., Myers, E.W., Adams, M.D., Smith, H.O., Broder, S., Gardner, M.J., Fraser, C.M., Birney, E., Bork, P., Brey, P.T., Venter, J.C., Weissenbach, J., Kafatos, F.C., Collins, F.H., Hoffman, S.L., 2002. The Genome Sequence of the Malaria Mosquito Anopheles gambiae. Science 298, 129–149. https://doi.org/10.1126/science.1076181

Hong, C.C., Hashimoto, C., 1996. The maternal nudel protein of Drosophila has two distinct roles important for embryogenesis. Genetics 143, 1653–1661.

Horvath, P., Romero, D.A., Coûté-Monvoisin, A.-C., Richards, M., Deveau, H., Moineau, S., Boyaval, P., Fremaux, C., Barrangou, R., 2008. Diversity, Activity, and Evolution of CRISPR Loci in Streptococcus thermophilus. J. Bacteriol. 190, 1401–1412. https://doi.org/10.1128/JB.01415-07

Hsu, P.D., Scott, D.A., Weinstein, J.A., Ran, F.A., Konermann, S., Agarwala, V., Li, Y., Fine, E.J., Wu, X., Shalem, O., Cradick, T.J., Marraffini, L.A., Bao, G., Zhang, F., 2013. DNA targeting specificity of RNA-guided Cas9 nucleases. Nature Biotechnology 31, 827. https://doi.org/10.1038/nbt.2647

Hu, J.H., Miller, S.M., Geurts, M.H., Tang, W., Chen, L., Sun, N., Zeina, C.M., Gao, X., Rees, H.A., Lin, Z., Liu, D.R., 2018. Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. Nature. https://doi.org/10.1038/nature26155

Hubisz, M.J., Pollard, K.S., Siepel, A., 2010. PHAST and RPHAST: phylogenetic analysis with space/time models. Briefings in Bioinformatics 12, 41–51. https://doi.org/10.1093/bib/bbq072

Huerta-Cepas, J., Serra, F., Bork, P., 2016. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Molecular Biology and Evolution 33, 1635–1638. https://doi.org/10.1093/molbev/msw046

Hwang, C.-L., Lai, Y.-J., Liu, T.-Y., 1993. A new approach for multiple objective decision making. Computers & Operations Research 20, 889–899. https://doi.org/10.1016/0305-0548(93)90109-V

Hwang, W.Y., Fu, Y., Reyon, D., Maeder, M.L., Tsai, S.Q., Sander, J.D., Peterson, R.T., Yeh, J.-R.J., Joung, J.K., 2013. Efficient genome editing in zebrafish using a CRISPR-Cas system. Nature Biotechnology 31, 227. https://doi.org/10.1038/nbt.2501

Iovino, N., Pane, A., Gaul, U., 2009. miR-184 has multiple roles in Drosophila female germline development. Dev Cell 17, 123–133. https://doi.org/10.1016/j.devcel.2009.06.008

Isaacs, A.T., Jasinskiene, N., Tretiakov, M., Thiery, I., Zettor, A., Bourgouin, C., James, A.A., 2012. Transgenic Anopheles stephensi coexpressing single-chain antibodies resist Plasmodium falciparum development. PNAS 109, E1922–E1930. https://doi.org/10.1073/pnas.1207738109

Jiang, F., Zhou, K., Ma, L., Gressel, S., Doudna, J.A., 2015. A Cas9–guide RNA complex preorganized for target DNA recognition. Science 348, 1477–1481. https://doi.org/10.1126/science.aab1452

Jiang, X., Peery, A., Hall, A.B., Sharma, A., Chen, X.-G., Waterhouse, R.M., Komissarov, A., Riehle, M.M., Shouche, Y., Sharakhova, M.V., Lawson, D., Pakpour, N., Arensburger, P., Davidson, V.L.M., Eiglmeier, K., Emrich, S., George, P., Kennedy, R.C., Mane, S.P., Maslen, G., Oringanje, C., Qi, Y., Settlage, R., Tojo, M., Tubio, J.M.C., Unger, M.F., Wang, B., Vernick, K.D., Ribeiro, J.M.C., James, A.A., Michel, K., Riehle, M.A., Luckhart, S., Sharakhov, I.V., Tu, Z., 2014. Genome analysis of a major urban malaria vector mosquito, Anopheles stephensi. Genome Biol 15, 459. https://doi.org/10.1186/s13059-014-0459-2

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., Charpentier, E., 2012. A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. Science 337, 816–821. https://doi.org/10.1126/science.1225829

Kamau, E., Campino, S., Amenga-Etego, L., Drury, E., Ishengoma, D., Johnson, K., Mumba, D., Kekre, M., Yavo, W., Mead, D., Bouyou-Akotet, M., Apinjoh, T., Golassa, L., Randrianarivelojosia, M., Andagalu, B., Maiga-Ascofare, O., Amambua-Ngwa, A., Tindana, P., Ghansah, A., MacInnis, B., Kwiatkowski, D., Djimde, A.A., 2015. K13-propeller polymorphisms in Plasmodium falciparum parasites from sub-Saharan Africa. J Infect Dis 211, 1352–1355. https://doi.org/10.1093/infdis/jiu608

Karvelis, T., Gasiunas, G., Siksnys, V., 2017. Methods for decoding Cas9 protospacer adjacent motif (PAM) sequences: A brief overview. Methods, CRISPR-Cas systems for Genome engineering and investigation 121, 3–8. https://doi.org/10.1016/j.ymeth.2017.03.006
Kent, T., Chandramouly, G., McDevitt, S.M., Ozdemir, A.Y., Pomerantz, R.T., 2015. Mechanism of microhomology-mediated end-joining promoted by human DNA polymerase θ. Nat Struct Mol Biol 22, 230–237. https://doi.org/10.1038/nsmb.2961
Kleinschmit, A., Koyama, T., Dejima, K., Hayashi, Y., Kamimura, K., Nakato, H., 2010. Drosophila heparan sulfate 6-O endosulfatase regulates Wingless morphogen gradient formation. Dev Biol 345, 204–214. https://doi.org/10.1016/j.ydbio.2010.07.006
Kleinstiver, B.P., Prew, M.S., Tsai, S.Q., Nguyen, N.T., Topkar, V.V., Zheng, Z., Joung, J.K., 2015a. Broadening the targeting range of *Staphylococcus aureus* CRISPR-Cas9 by modifying PAM recognition. Nature Biotechnology 33, 1293. https://doi.org/10.1038/nbt.3404
Kleinstiver, B.P., Prew, M.S., Tsai, S.Q., Topkar, V.V., Nguyen, N.T., Zheng, Z., Gonzales, A.P.W., Li, Z., Peterson, R.T., Yeh, J.-R.J., Aryee, M.J., Joung, J.K., 2015b. Engineered CRISPR-Cas9 nucleases with altered PAM specificities. Nature 523, 481–485. https://doi.org/10.1038/nature14592
Knipling, E.F., 1955. Possibilities of insect control or eradication through the use of sexually sterile males. Journal of Economic Entomology 48, 459–462.
Kranjc, N., Crisanti, A., Nolan, T., Bernardini, F., 2021. Anopheles gambiae Genome Conservation as a Resource for Rational Gene Drive Target Site Selection. Insects 12, 97. https://doi.org/10.3390/insects12020097
Kyrou, K., Hammond, A.M., Galizi, R., Kranjc, N., Burt, A., Beaghton, A.K., Nolan, T., Crisanti, A., 2018. A CRISPR–Cas9 gene drive targeting doublesex causes complete population suppression in caged Anopheles gambiae mosquitoes. Nature Biotechnology 36, 1062–1066. https://doi.org/10.1038/nbt.4245
Labun, K., Montague, T.G., Gagnon, J.A., Thyme, S.B., Valen, E., 2016. CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering. Nucleic Acids Res 44, W272–W276. https://doi.org/10.1093/nar/gkw398
Labun, K., Montague, T.G., Krause, M., Torres Cleuren, Y.N., Tjeldnes, H., Valen, E., 2019. CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. Nucleic Acids Research 47, W171–W174. https://doi.org/10.1093/nar/gkz365
Lacroix, R., McKemey, A.R., Raduan, N., Wee, L.K., Ming, W.H., Ney, T.G., A.a, S.R., Salman, S., Subramaniam, S., Nordin, O., A.t, N.H., Angamuthu, C., Mansor, S.M., Lees, R.S., Naish, N., Scaife, S., Gray, P., Labbé, G., Beech, C., Nimmo, D., Alphey, L., Vasan, S.S., Lim, L.H., A, N.W., Murad, S., 2012. Open Field Release of Genetically Engineered Sterile Male Aedes aegypti in Malaysia. PLOS ONE 7, e42771. https://doi.org/10.1371/journal.pone.0042771
Langmead, B., 2010. Aligning Short Sequencing Reads with Bowtie. Current Protocols in Bioinformatics 32, 11.7.1-11.7.14. https://doi.org/10.1002/0471250953.bi1107s32
Larkin, A., Marygold, S.J., Antonazzo, G., Attrill, H., dos Santos, G., Garapati, P.V., Goodman, J.L., Gramates, L.S., Millburn, G., Strelets, V.B., Tabone, C.J., Thurmond, J., FlyBase Consortium, 2021. FlyBase: updates to the Drosophila melanogaster knowledge base. Nucleic Acids Research 49, D899–D907. https://doi.org/10.1093/nar/gkaa1026
Laven, H., Cousserans, J., Guille, G., 1972. Eradicating Mosquitoes using Translocations: a First Field Experiment. Nature 236, 456–457. https://doi.org/10.1038/236456a0
Le Rhun, A., Escalera-Maurer, A., Bratovič, M., Charpentier, E., 2019. CRISPR-Cas in Streptococcus pyogenes. RNA Biol 16, 380–389. https://doi.org/10.1080/15476286.2019.1582974
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352
Li, P., Peng, J., Hu, J., Xu, Z., Xie, W., Yuan, L., 2011. Localized expression pattern of miR-184 in Drosophila. Mol Biol Rep 38, 355–358. https://doi.org/10.1007/s11033-010-0115-1
Liao, D., 2008. Concerted Evolution. https://doi.org/10.1002/9780470015902.a0005132.pub2
Lieber, M.R., 2010. The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End-Joining Pathway. Annual Review of Biochemistry 79, 181–211. https://doi.org/10.1146/annurev.biochem.052308.093131

Liu, G., Zhang, Y., Zhang, T., 2020. Computational approaches for effective CRISPR guide RNA design and evaluation. Computational and Structural Biotechnology Journal 18, 35–44. https://doi.org/10.1016/j.csbj.2019.11.006
Liu, N., 2015. Insecticide Resistance in Mosquitoes: Impact, Mechanisms, and Research Directions. Annual Review of Entomology 60, 537–559. https://doi.org/10.1146/annurev-ento-010814-020828
Liu, W., Hao, Z., Huang, L., Chen, L., Wei, Q., Cai, L., Liang, S., 2017. Comparative expression profile of microRNAs in Anopheles anthropophagus midgut after blood-feeding and Plasmodium infection. Parasites & Vectors 10, 86. https://doi.org/10.1186/s13071-017-2027-6
Lu, J.S., Bindewald, E., Kasprzak, W.K., Shapiro, B.A., 2018. RiboSketch: versatile visualization of multi-stranded RNA and DNA secondary structure. Bioinformatics 34, 4297–4299. https://doi.org/10.1093/bioinformatics/bty468
Lucas, E.R., Miles, A., Harding, N.J., Clarkson, C.S., Lawniczak, M.K.N., Kwiatkowski, D.P., Weetman, D., Donnelly, M.J., 2019. Whole-genome sequencing reveals high complexity of copy number variation at insecticide resistance loci in malaria mosquitoes. Genome Res 29, 1250–1261. https://doi.org/10.1101/gr.245795.118
Lynch, M., 2010. Evolution of the mutation rate. Trends in Genetics 26, 345–352. https://doi.org/10.1016/j.tig.2010.05.003
Magadán, A.H., Dupuis, M.-È., Villion, M., Moineau, S., 2012. Cleavage of Phage DNA by the Streptococcus thermophilus CRISPR3-Cas System. PLoS One 7. https://doi.org/10.1371/journal.pone.0040913
Maggert, K.A., Golic, K.G., 2002. The Y chromosome of Drosophila melanogaster exhibits chromosome-wide imprinting. Genetics 162, 1245–1258.
Mahmoudi, S., Keshavarz, H., 2017. Efficacy of phase 3 trial of RTS, S/AS01 malaria vaccine: The need for an alternative development plan. Human Vaccines & Immunotherapeutics 13, 2098–2101. https://doi.org/10.1080/21645515.2017.1295906
Maïga, H., Damiens, D., Niang, A., Sawadogo, S.P., Fatherhaman, O., Lees, R.S., Roux, O., Dabiré, R.K., Ouédraogo, G.A., Tripet, F., Diabaté, A., Gilles, J.R., 2014. Mating competitiveness of sterile male Anopheles coluzzii in large cages. Malaria Journal 13, 460. https://doi.org/10.1186/1475-2875-13-460
Makarova, K.S., Koonin, E.V., 2015. Annotation and Classification of CRISPR-Cas Systems, in: CRISPR, Methods in Molecular Biology. Humana Press, New York, NY, pp. 47–75. https://doi.org/10.1007/978-1-4939-2687-9_4
Makarova, K.S., Wolf, Y.I., Alkhnbashi, O.S., Costa, F., Shah, S.A., Saunders, S.J., Barrangou, R., Brouns, S.J.J., Charpentier, E., Haft, D.H., Horvath, P., Moineau, S., Mojica, F.J.M., Terns, R.M., Terns, M.P., White, M.F., Yakunin, A.F., Garrett, R.A., van der Oost, J., Backofen, R., Koonin, E.V., 2015. An updated evolutionary classification of CRISPR-Cas systems. Nat Rev Micro 13, 722–736. https://doi.org/10.1038/nrmicro3569
Makarova, K.S., Wolf, Y.I., Iranzo, J., Shmakov, S.A., Alkhnbashi, O.S., Brouns, S.J.J., Charpentier, E., Cheng, D., Haft, D.H., Horvath, P., Moineau, S., Mojica, F.J.M., Scott, D., Shah, S.A., Siksnys, V., Terns, M.P., Venclovas, Č., White, M.F., Yakunin, A.F., Yan, W., Zhang, F., Garrett, R.A., Backofen, R., van der Oost, J., Barrangou, R., Koonin, E.V., 2020. Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. Nat Rev Microbiol 18, 67–83. https://doi.org/10.1038/s41579-019-0299-x
Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E., Church, G.M., 2013. RNA-guided human genome engineering via Cas9. Science 339, 823–826. https://doi.org/10.1126/science.1232033
Marinotti, O., Cerqueira, G.C., de Almeida, L.G.P., Ferro, M.I.T., Loreto, E.L. da S., Zaha, A., Teixeira, S.M.R., Wespiser, A.R., Almeida E Silva, A., Schlindwein, A.D., Pacheco, A.C.L., Silva, A.L. da C. da, Graveley, B.R., Walenz, B.P., Lima, B. de A., Ribeiro, C.A.G., Nunes-Silva, C.G., de Carvalho, C.R., Soares, C.M. de A., de Menezes, C.B.A., Matiolli, C., Caffrey, D., Araújo, D.A.M., de Oliveira, D.M., Golenbock, D., Grisard, E.C., Fantinatti-Garboggini, F., de Carvalho, F.M., Barcellos, F.G., Prosdocimi, F., May, G., Azevedo Junior, G.M. de, Guimarães, G.M., Goldman, G.H., Padilha, I.Q.M., Batista, J. da S., Ferro, J.A., Ribeiro, J.M.C., Fietto, J.L.R., Dabbas, K.M., Cerdeira, L., Agnez-Lima, L.F., Brocchi, M., de Carvalho, M.O., Teixeira, M. de M., Diniz Maia, M. de M., Goldman, M.H.S., Cruz Schneider, M.P., Felipe, M.S.S., Hungria, M., Nicolás, M.F., Pereira, M., Montes, M.A., Cantão, M.E., Vincentz, M.,

Rafael, M.S., Silverman, N., Stoco, P.H., Souza, R.C., Vicentini, R., Gazzinelli, R.T., Neves, R. de O., Silva, R., Astolfi-Filho, S., Maciel, T.E.F., Urményi, T.P., Tadei, W.P., Camargo, E.P., de Vasconcelos, A.T.R., 2013. The genome of Anopheles darlingi, the main neotropical malaria vector. Nucleic Acids Res 41, 7387–7400. https://doi.org/10.1093/nar/gkt484
Marraffini, L.A., 2015. CRISPR-Cas immunity in prokaryotes. Nature 526, 55. https://doi.org/10.1038/nature15386
Marshall, J.M., Buchman, A., C, H.M.S., Akbari, O.S., 2017. Overcoming evolved resistance to population-suppressing homing-based gene drives. Sci Rep 7, 3776. https://doi.org/10.1038/s41598-017-02744-7
Marshall, R., Maxwell, C.S., Collins, S.P., Jacobsen, T., Luo, M.L., Begemann, M.B., Gray, B.N., January, E., Singer, A., He, Y., Beisel, C.L., Noireaux, V., 2018. Rapid and Scalable Characterization of CRISPR Technologies Using an E. coli Cell-Free Transcription-Translation System. Molecular Cell 69, 146-157.e3. https://doi.org/10.1016/j.molcel.2017.12.007
Mendoza, B.J., Trinh, C.T., 2018. In Silico Processing of the Complete CRISPR-Cas Spacer Space for Identification of PAM Sequences. Biotechnology Journal 13, 1700595. https://doi.org/10.1002/biot.201700595
Mi, H., Muruganujan, A., Ebert, D., Huang, X., Thomas, P.D., 2019. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Research 47, D419–D426. https://doi.org/10.1093/nar/gky1038
Miles, A., Harding, N.J., Bottà, G., Clarkson, C.S., Antão, T., Kozak, K., Schrider, D.R., Kern, A.D., Redmond, S., Sharakhov, I., Pearson, R.D., Bergey, C., Fontaine, M.C., Donnelly, M.J., Lawniczak, M.K.N., Kwiatkowski, D.P., Donnelly, M.J., Ayala, D., Besansky, N.J., Burt, A., Caputo, B., della Torre, A., Fontaine, M.C., Godfray, H.C.J., Hahn, M.W., Kern, A.D., Kwiatkowski, D.P., Lawniczak, M.K.N., Midega, J., Neafsey, D.E., O'Loughlin, S., Pinto, J., Riehle, M.M., Sharakhov, I., Vernick, K.D., Weetman, D., Wilding, C.S., White, B.J., Troco, A.D., Pinto, J., Diabaté, A., O'Loughlin, S., Burt, A., Costantini, C., Rohatgi, K.R., Besansky, N.J., Elissa, N., Pinto, J., Coulibaly, B., Riehle, M.M., Vernick, K.D., Pinto, J., Dinis, J., Midega, J., Mbogo, C., Bejon, P., Wilding, C.S., Weetman, D., Mawejje, H.D., Donnelly, M.J., Weetman, D., Wilding, C.S., Donnelly, M.J., Stalker, J., Rockett, K., Drury, E., Mead, D., Jeffreys, A., Hubbart, C., Rowlands, K., Isaacs, A.T., Jyothi, D., Malangone, C., Vauterin, P., Jeffery, B., Wright, I., Hart, L., Kluczyński, K., Cornelius, V., MacInnis, B., Henrichs, C., Giacomantonio, R., Kwiatkowski, D.P., 2017. Genetic diversity of the African malaria vector Anopheles gambiae. Nature. https://doi.org/10.1038/nature24995
Miles, A., Kirkham, J., Durant, M., Bourbeau, J., Onalan, T., Hamman, J., Patel, Z., shikharsg, Rocklin, M., dussin, raphael, Schut, V., Andrade, E.S. de, Abernathey, R., Noyes, C., sbalmer, bot, pyup io, Tran, T., Saalfeld, S., Swaney, J., Moore, J., Jevnik, J., Kelleher, J., Funke, J., Sakkis, G., Barnes, C., Banihirwe, A., 2020. zarr-developers/zarr-python: v2.4.0. Zenodo. https://doi.org/10.5281/zenodo.3773450
Miles, A., Ralph, P., Rae, S., Pisupati, R., 2019. cggh/scikit-allel: v1.2.1.
Miller, S.M., Wang, T., Randolph, P.B., Arbab, M., Shen, M.W., Huang, T.P., Matuszek, Z., Newby, G.A., Rees, H.A., Liu, D.R., 2020. Continuous evolution of SpCas9 variants compatible with non-G PAMs. Nat Biotechnol 38, 471–481. https://doi.org/10.1038/s41587-020-0412-8
Mita, T., Tanabe, K., Takahashi, N., Tsukahara, T., Eto, H., Dysoley, L., Ohmae, H., Kita, K., Krudsood, S., Looareesuwan, S., Kaneko, A., Björkman, A., Kobayakawa, T., 2007. Independent Evolution of Pyrimethamine Resistance in Plasmodium falciparum Isolates in Melanesia. Antimicrobial Agents and Chemotherapy 51, 1071–1077. https://doi.org/10.1128/AAC.01186-06
Mojica, F.J.M., 2009. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology 155, 733–740. https://doi.org/10.1099/mic.0.023960-0
Moreno-Mateos, M.A., Vejnar, C.E., Beaudoin, J.-D., Fernandez, J.P., Mis, E.K., Khokha, M.K., Giraldez, A.J., 2015. CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. Nat Methods 12, 982–988. https://doi.org/10.1038/nmeth.3543
Morrison, N.I., Simmons, G.S., Fu, G., O'Connell, S., Walker, A.S., Dafa'alla, T., Walters, M., Claus, J., Tang, G., Jin, L., Marubbi, T., Epton, M.J., Harris, C.L., Staten, R.T., Miller, E., Miller, T.A., Alphey, L., 2012. Engineered repressible lethality for controlling the pink bollworm, a lepidopteran pest of cotton. PLoS One 7, e50922. https://doi.org/10.1371/journal.pone.0050922

Mota, M.B.S., Carvalho, M.A., Monteiro, A.N.A., Mesquita, R.D., 2019. DNA damage response and repair in perspective: Aedes aegypti, Drosophila melanogaster and Homo sapiens. Parasites & Vectors 12, 533. https://doi.org/10.1186/s13071-019-3792-1

Mount, S.M., Gotea, V., Lin, C.-F., Hernandez, K., Makałowski, W., 2007. Spliceosomal small nuclear RNA genes in 11 insect genomes. RNA 13, 5–14. https://doi.org/10.1261/rna.259207

Moynahan, M.E., Jasin, M., 2010. Mitotic homologous recombination maintains genomic stability and suppresses tumorigenesis. Nature Reviews Molecular Cell Biology 11, 196–207. https://doi.org/10.1038/nrm2851

Neafsey, D.E., Waterhouse, R.M., Abai, M.R., Aganezov, S.S., Alekseyev, M.A., Allen, J.E., Amon, J., Arcà, B., Arensburger, P., Artemov, G., Assour, L.A., Basseri, H., Berlin, A., Birren, B.W., Blandin, S.A., Brockman, A.I., Burkot, T.R., Burt, A., Chan, C.S., Chauve, C., Chiu, J.C., Christensen, M., Costantini, C., Davidson, V.L.M., Deligianni, E., Dottorini, T., Dritsou, V., Gabriel, S.B., Guelbeogo, W.M., Hall, A.B., Han, M.V., Hlaing, T., Hughes, D.S.T., Jenkins, A.M., Jiang, X., Jungreis, I., Kakani, E.G., Kamali, M., Kemppainen, P., Kennedy, R.C., Kirmitzoglou, I.K., Koekemoer, L.L., Laban, N., Langridge, N., Lawniczak, M.K.N., Lirakis, M., Lobo, N.F., Lowy, E., MacCallum, R.M., Mao, C., Maslen, G., Mbogo, C., McCarthy, J., Michel, K., Mitchell, S.N., Moore, W., Murphy, K.A., Naumenko, A.N., Nolan, T., Novoa, E.M., O'Loughlin, S., Oringanje, C., Oshaghi, M.A., Pakpour, N., Papathanos, P.A., Peery, A.N., Povelones, M., Prakash, A., Price, D.P., Rajaraman, A., Reimer, L.J., Rinker, D.C., Rokas, A., Russell, T.L., Sagnon, N., Sharakhova, M.V., Shea, T., Simão, F.A., Simard, F., Slotman, M.A., Somboon, P., Stegniy, V., Struchiner, C.J., Thomas, G.W.C., Tojo, M., Topalis, P., Tubio, J.M.C., Unger, M.F., Vontas, J., Walton, C., Wilding, C.S., Willis, J.H., Wu, Y.-C., Yan, G., Zdobnov, E.M., Zhou, X., Catteruccia, F., Christophides, G.K., Collins, F.H., Cornman, R.S., Crisanti, A., Donnelly, M.J., Emrich, S.J., Fontaine, M.C., Gelbart, W., Hahn, M.W., Hansen, I.A., Howell, P.I., Kafatos, F.C., Kellis, M., Lawson, D., Louis, C., Luckhart, S., Muskavitch, M.A.T., Ribeiro, J.M., Riehle, M.A., Sharakhov, I.V., Tu, Z., Zwiebel, L.J., Besansky, N.J., 2015. Highly evolvable malaria vectors: The genomes of 16 Anopheles mosquitoes. Science 347. https://doi.org/10.1126/science.1258522

Nishimasu, H., Ran, F.A., Hsu, P.D., Konermann, S., Shehata, S.I., Dohmae, N., Ishitani, R., Zhang, F., Nureki, O., 2014. Crystal structure of Cas9 in complex with guide RNA and target DNA. Cell 156, 935–949. https://doi.org/10.1016/j.cell.2014.02.001

Noedl, H., Se, Y., Schaecher, K., Smith, B.L., Socheat, D., Fukuda, M.M., 2008. Evidence of artemisinin-resistant malaria in western Cambodia. New England Journal of Medicine 359, 2619–2620.

Nolan, T., 2021. Control of malaria-transmitting mosquitoes using gene drives. Philosophical Transactions of the Royal Society B: Biological Sciences 376, 20190803. https://doi.org/10.1098/rstb.2019.0803

Nolan, T., Bower, T.M., Brown, A.E., Crisanti, A., Catteruccia, F., 2002. piggyBac-mediated germline transformation of the malaria mosquito Anopheles stephensi using the red fluorescent protein dsRED as a selectable marker. J Biol Chem 277, 8759–8762. https://doi.org/10.1074/jbc.C100766200

Nussenzweig, R.S., Vanderberg, J., Most, H., Orton, C., 1967. Protective Immunity produced by the Injection of X-irradiated Sporozoites of Plasmodium berghei. Nature 216, 160–162. https://doi.org/10.1038/216160a0

Oberhofer, G., Ivy, T., Hay, B.A., 2018. Behavior of homing endonuclease gene drives targeting genes required for viability or female fertility with multiplexed guide RNAs. PNAS 115, E9343–E9352. https://doi.org/10.1073/pnas.1805278115

O'Brochta, D.A., Gomez, S.P., Handler, A.M., 1991. P element excision in Drosophila melanogaster and related drosophilids. Mol Gen Genet 225, 387–394. https://doi.org/10.1007/BF00261678

Oh, S., Harvey, A., Zimbric, J., Wang, Y., Nguyen, T., Jackson, P.J., Hendrickson, E.A., 2014. DNA ligase III and DNA ligase IV carry out genetically distinct forms of end joining in human somatic cells. DNA Repair 21, 97–110. https://doi.org/10.1016/j.dnarep.2014.04.015

O'Loughlin, S.M., Forster, A.J., Fuchs, S., Dottorini, T., Nolan, T., Crisanti, A., Burt, A., 2021. Ultra-conserved sequences in the genomes of highly diverse Anopheles mosquitoes, with implications for malaria vector control. G3 Genes|Genomes|Genetics 11. https://doi.org/10.1093/g3journal/jkab086

Overcash, J.M., Aryan, A., Myles, K.M., Adelman, Z.N., 2015. Understanding the DNA damage response in order to achieve desired gene editing outcomes in mosquitoes. Chromosome Res 23, 31–42. https://doi.org/10.1007/s10577-014-9450-8
Papathanos, P.A., Bourtzis, K., Tripet, F., Bossin, H., Virginio, J.F., Capurro, M.L., Pedrosa, M.C., Guindo, A., Sylla, L., Coulibaly, M.B., Yao, F.A., Epopa, P.S., Diabate, A., 2018. A perspective on the need and current status of efficient sex separation methods for mosquito genetic control. Parasites & Vectors 11, 654. https://doi.org/10.1186/s13071-018-3222-9
Park, J., Bae, S., Kim, J.-S., 2015. Cas-Designer: a web-based tool for choice of CRISPR-Cas9 target sites. Bioinformatics 31, 4014–4016. https://doi.org/10.1093/bioinformatics/btv537
Phuc, H.K., Andreasen, M.H., Burton, R.S., Vass, C., Epton, M.J., Pape, G., Fu, G., Condon, K.C., Scaife, S., Donnelly, C.A., Coleman, P.G., White-Cooper, H., Alphey, L., 2007. Late-acting dominant lethal genetic systems and mosquito control. BMC Biology 5, 11. https://doi.org/10.1186/1741-7007-5-11
Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., Siepel, A., 2010. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 20, 110–121. https://doi.org/10.1101/gr.097857.109
Prykhozhij, S.V., Rajan, V., Gaston, D., Berman, J.N., 2015. CRISPR MultiTargeter: A Web Tool to Find Common and Unique CRISPR Single Guide RNA Targets in a Set of Similar Sequences. PLOS ONE 10, e0119372. https://doi.org/10.1371/journal.pone.0119372
Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. https://doi.org/10.1093/bioinformatics/btq033
Radhakrishnan, S.K., Jette, N., Lees-Miller, S.P., 2014. Non-homologous end joining: Emerging themes and unanswered questions. DNA Repair, Recent Developments in Non-Homologous End Joining 17, 2–8. https://doi.org/10.1016/j.dnarep.2014.01.009
Raghavendra, K., Barik, T.K., Reddy, B.P.N., Sharma, P., Dash, A.P., 2011. Malaria vector control: from past to future. Parasitol Res 108, 757–779. https://doi.org/10.1007/s00436-010-2232-0
Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A., Zhang, F., 2013. Genome engineering using the CRISPR-Cas9 system. Nat Protoc 8, 2281–2308. https://doi.org/10.1038/nprot.2013.143
Ranson, H., N'Guessan, R., Lines, J., Moiroux, N., Nkuni, Z., Corbel, V., 2011. Pyrethroid resistance in African anopheline mosquitoes: what are the implications for malaria control? Trends in Parasitology 27, 91–98. https://doi.org/10.1016/j.pt.2010.08.004
Reback, J., McKinney, W., jbrockmendel, Bossche, J.V. den, Augspurger, T., Cloud, P., Hawkins, S., gfyoung, Sinhrks, Roeschke, M., Klein, A., Petersen, T., Tratner, J., She, C., Ayd, W., Naveh, S., patrick, Garcia, M., Schendel, J., Hayden, A., Saxton, D., Jancauskas, V., Gorelli, M., Shadrach, R., McMaster, A., Battiston, P., Seabold, S., Dong, K., chris-b1, h-vetinari, 2021. pandas-dev/pandas: Pandas 1.2.4. Zenodo. https://doi.org/10.5281/zenodo.4681666
Redman, M., King, A., Watson, C., King, D., 2016. What is CRISPR/Cas9? Archives of Disease in Childhood - Education and Practice 101, 213–215. https://doi.org/10.1136/archdischild-2016-310459
Rodríguez-Sanchez, I.P., Saldaña-Torres, D.R., Villanueva-Segura, O.K., Garza-Rodriguez, M.L., Gómez-Govea, M.A., Liang, G., de Lourdes Ramírez-Ahuja, M., De La Luz Martinez-Fierro, M., Delgado-Enciso, I., Martinez-de-Villarreal, L.E., Zhou, Y., Flores-Suarez, A.E., Chen, X., Resendez-Pérez, D., Zhang, C.-Y., Ponce-Garcia, G., 2021. miRNAs of Aedes aegypti (Linnaeus 1762) conserved in six orders of the class Insecta. Sci Rep 11, 10706. https://doi.org/10.1038/s41598-021-90095-9
Ronshaugen, M., Biemar, F., Piel, J., Levine, M., Lai, E.C., 2005. The Drosophila microRNA iab-4 causes a dominant homeotic transformation of halteres to wings. Genes Dev 19, 2947–2952. https://doi.org/10.1101/gad.1372505
RTS,S Clinical Trials Partnership, 2015. Efficacy and safety of RTS,S/AS01 malaria vaccine with or without a booster dose in infants and children in Africa: final results of a phase 3, individually randomised, controlled trial. The Lancet 386, 31–45. https://doi.org/10.1016/S0140-6736(15)60721-8
Scali, C., Catteruccia, F., Li, Q., Crisanti, A., 2005. Identification of sex-specific transcripts of the Anopheles gambiae doublesex gene. J Exp Biol 208, 3701–3709. https://doi.org/10.1242/jeb.01819

Schmidt, H., Collier, T.C., Hanemaaijer, M.J., Houston, P.D., Lee, Y., Lanzaro, G.C., 2020. Abundance of conserved CRISPR-Cas9 target sites within the highly polymorphic genomes of Anopheles and Aedes mosquitoes. Nature Communications 11, 1425. https://doi.org/10.1038/s41467-020-15204-0
Scott, M.J., Concha, C., Welch, J.B., Phillips, P.L., Skoda, S.R., 2017. Review of research advances in the screwworm eradication program over the past 25 years. Entomologia Experimentalis et Applicata 164, 226–236.
Segers, M.E., Lebeer, S., 2014. Towards a better understanding of Lactobacillus rhamnosus GG - host interactions. Microb Cell Fact 13, S7. https://doi.org/10.1186/1475-2859-13-S1-S7
Sekelsky, J., 2017. DNA Repair in Drosophila: Mutagens, Models, and Missing Genes. Genetics 205, 471–490. https://doi.org/10.1534/genetics.116.186759
Serebrovskii, A.S., 1940. On the possibility of a new method for the control of insect pests. Zool Zhurnal 19, 618–630.
Sfeir, A., Symington, L.S., 2015. Microhomology-Mediated End Joining: A Back-up Survival Mechanism or Dedicated Pathway? Trends in Biochemical Sciences 40, 701–714. https://doi.org/10.1016/j.tibs.2015.08.006
Shah, S.A., Erdmann, S., Mojica, F.J.M., Garrett, R.A., 2013. Protospacer recognition motifs. RNA Biol 10, 891–899. https://doi.org/10.4161/rna.23764
Sharakhova, M.V., Hammond, M.P., Lobo, N.F., Krzywinski, J., Unger, M.F., Hillenmeyer, M.E., Bruggner, R.V., Birney, E., Collins, F.H., 2007. Update of the Anopheles gambiae PEST genome assembly. Genome Biol 8, R5. https://doi.org/10.1186/gb-2007-8-1-r5
Sharma, S., Javadekar, S.M., Pandey, M., Srivastava, M., Kumari, R., Raghavan, S.C., 2015. Homology and enzymatic requirements of microhomology-dependent alternative end joining. Cell Death Dis 6, e1697. https://doi.org/10.1038/cddis.2015.58
Shen, M.W., Arbab, M., Hsu, J.Y., Worstell, D., Culbertson, S.J., Krabbe, O., Cassa, C.A., Liu, D.R., Gifford, D.K., Sherwood, R.I., 2018. Predictable and precise template-free CRISPR editing of pathogenic variants. Nature 563, 646–651. https://doi.org/10.1038/s41586-018-0686-x
Shirley, M., 2014. pyfaidx: efficient pythonic random access to fasta subsequences. https://doi.org/10.6084/m9.figshare.972933.v3
Shmakov, S., Abudayyeh, O.O., Makarova, K.S., Wolf, Y.I., Gootenberg, J.S., Semenova, E., Minakhin, L., Joung, J., Konermann, S., Severinov, K., Zhang, F., Koonin, E.V., 2015. Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. Molecular Cell 60, 385–397. https://doi.org/10.1016/j.molcel.2015.10.008
Shmakov, S., Smargon, A., Scott, D., Cox, D., Pyzocha, N., Yan, W., Abudayyeh, O.O., Gootenberg, J.S., Makarova, K.S., Wolf, Y.I., Severinov, K., Zhang, F., Koonin, E.V., 2017. Diversity and evolution of class 2 CRISPR–Cas systems. Nature Reviews Microbiology 15, 169–182. https://doi.org/10.1038/nrmicro.2016.184
Shmakov, S.A., Makarova, K.S., Wolf, Y.I., Severinov, K.V., Koonin, E.V., 2018. Systematic prediction of genes functionally linked to CRISPR-Cas systems by gene neighborhood analysis. PNAS 201803440. https://doi.org/10.1073/pnas.1803440115
Shmakov, S.A., Sitnik, V., Makarova, K.S., Wolf, Y.I., Severinov, K.V., Koonin, E.V., 2017. The CRISPR Spacer Space Is Dominated by Sequences from Species-Specific Mobilomes. mBio 8, e01397-17. https://doi.org/10.1128/mBio.01397-17
Shmakov, S.A., Wolf, Y.I., Savitskaya, E., Severinov, K.V., Koonin, E.V., 2020. Mapping CRISPR spaceromes reveals vast host-specific viromes of prokaryotes. Commun Biol 3. https://doi.org/10.1038/s42003-020-1014-1
Siksnys, V., Gasiunas, G., 2016. Rewiring Cas9 to Target New PAM Sequences. Molecular Cell 61, 793–794. https://doi.org/10.1016/j.molcel.2016.03.002
Simoni, A., Hammond, A., Beaghton, A.K., Galizi, R., Taxiarchi, C., Kyrou, K., Meacci, D., Gribble, M., Morselli, G., Burt, A., Nolan, T., Crisanti, A., 2020. A male-biased sex-distorter gene drive for the human malaria vector Anopheles gambiae. Nat Biotechnol 38, 1054–1060. https://doi.org/10.1038/s41587-020-0508-1
Simoni, A., Siniscalchi, C., Chan, Y.-S., Huen, D.S., Russell, S., Windbichler, N., Crisanti, A., 2014. Development of synthetic selfish elements based on modular nucleases in Drosophila melanogaster. Nucleic Acids Research 42, 7461–7472. https://doi.org/10.1093/nar/gku387

Sinka, M.E., Bangs, M.J., Manguin, S., Rubio-Palis, Y., Chareonviriyaphap, T., Coetzee, M., Mbogo, C.M., Hemingway, J., Patil, A.P., Temperley, W.H., Gething, P.W., Kabaria, C.W., Burkot, T.R., Harbach, R.E., Hay, S.I., 2012. A global map of dominant malaria vectors. Parasites & Vectors 5, 69. https://doi.org/10.1186/1756-3305-5-69
Slaymaker, I.M., Gao, L., Zetsche, B., Scott, D.A., Yan, W.X., Zhang, F., 2016. Rationally engineered Cas9 nucleases with improved specificity. Science 351, 84–88. https://doi.org/10.1126/science.aad5227
Slotman, M., Torre, A.D., Powell, J.R., 2005. Female Sterility in Hybrids Between Anopheles Gambiae and a. Arabiensis, and the Causes of Haldane's Rule. Evolution 59, 1016–1026. https://doi.org/10.1111/j.0014-3820.2005.tb01040.x
Stemmer, M., Thumberger, T., Keyer, M. del S., Wittbrodt, J., Mateo, J.L., 2015. CCTop: An Intuitive, Flexible and Reliable CRISPR/Cas9 Target Prediction Tool. PLOS ONE 10, e0124633. https://doi.org/10.1371/journal.pone.0124633
Sternberg, S.H., LaFrance, B., Kaplan, M., Doudna, J.A., 2015. Conformational control of DNA target cleavage by CRISPR–Cas9. Nature 527, 110–113. https://doi.org/10.1038/nature15544
Sternberg, S.H., Redding, S., Jinek, M., Greene, E.C., Doudna, J.A., 2014. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature 507, 62–67. https://doi.org/10.1038/nature13011
Sturm, A., Amino, R., Sand, C. van de, Regen, T., Retzlaff, S., Rennenberg, A., Krueger, A., Pollok, J.-M., Menard, R., Heussler, V.T., 2006. Manipulation of Host Hepatocytes by the Malaria Parasite for Delivery into Liver Sinusoids. Science 313, 1287–1290. https://doi.org/10.1126/science.1129720
Swarts, D.C., Mosterd, C., van Passel, M.W.J., Brouns, S.J.J., 2012. CRISPR interference directs strand specific spacer acquisition. PLoS One 7, e35888. https://doi.org/10.1371/journal.pone.0035888
Sweeny, T.L., Barr, A.R., 1978. Sex Ratio Distortion Caused by Meiotic Drive in a Mosquito, Culex pipiens L. Genetics 88, 427–446.
Tajkarimi, M., Wexler, H.M., 2017. CRISPR-Cas Systems in Bacteroides fragilis, an Important Pathobiont in the Human Gut Microbiome. Front Microbiol 8. https://doi.org/10.3389/fmicb.2017.02234
Takeda, S., Nakamura, K., Taniguchi, Y., Paull, T.T., 2007. Ctp1/CtIP and the MRN complex collaborate in the initial steps of homologous recombination. Mol Cell 28, 351–352. https://doi.org/10.1016/j.molcel.2007.10.016
Tan, G., Polychronopoulos, D., Lenhard, B., 2019. CNEr: A toolkit for exploring extreme noncoding conservation. PLOS Computational Biology 15, e1006940.
Taxiarchi, C., Kranjc, N., Kriezis, A., Kyrou, K., Bernardini, F., Russell, S., Nolan, T., Crisanti, A., Galizi, R., 2019. High-resolution transcriptional profiling of Anopheles gambiae spermatogenesis reveals mechanisms of sex chromosome regulation. Scientific Reports 9, 14841. https://doi.org/10.1038/s41598-019-51181-1
Thawornwattana, Y., Dalquen, D., Yang, Z., 2018. Coalescent Analysis of Phylogenomic Data Confidently Resolves the Species Relationships in the Anopheles gambiae Species Complex. Molecular Biology and Evolution 35, 2512–2527. https://doi.org/10.1093/molbev/msy158
The Anopheles gambiae 1000 Genomes Consortium, 2021. Ag1000G phase 3 SNP data release [WWW Document]. MalariaGEN.
The Anopheles gambiae 1000 Genomes Consortium, 2020. Genome variation and population structure among 1142 mosquitoes of the African malaria vector species Anopheles gambiae and Anopheles coluzzii. Genome Research 30, 1533–1546. https://doi.org/10.1101/gr.262790.120
The Anopheles gambiae 1000 Genomes Consortium, 2017. Ag1000G phase 2 AR1 data release. [WWW Document]. MalariaGEN.
Thistlewood, H.M.A., Judd, G.J.R., 2019. Twenty-five Years of Research Experience with the Sterile Insect Technique and Area-Wide Management of Codling Moth, Cydia pomonella (L.), in Canada. Insects 10, 292. https://doi.org/10.3390/insects10090292
Thomas, D.D., Donnelly, C.A., Wood, R.J., Alphey, L.S., 2000. Insect Population Control Using a Dominant, Repressible, Lethal Genetic System. Science 287, 2474–2476. https://doi.org/10.1126/science.287.5462.2474

Thyme, S.B., Boissel, S.J.S., Arshiya Quadri, S., Nolan, T., Baker, D.A., Park, R.U., Kusak, L., Ashworth, J., Baker, D., 2014. Reprogramming homing endonuclease specificity through computational design and directed evolution. Nucleic Acids Research 42, 2564–2576. https://doi.org/10.1093/nar/gkt1212
Truong, L.N., Li, Y., Shi, L.Z., Hwang, P.Y.-H., He, J., Wang, H., Razavian, N., Berns, M.W., Wu, X., 2013. Microhomology-mediated End Joining and Homologous Recombination share the initial end resection step to repair DNA double-strand breaks in mammalian cells. Proceedings of the National Academy of Sciences 110, 7720–7725. https://doi.org/10.1073/pnas.1213431110
Turner, J.M.A., 2007. Meiotic sex chromosome inactivation. Development 134, 1823–1831. https://doi.org/10.1242/dev.000018
Uwimana, A., Legrand, E., Stokes, B.H., Ndikumana, J.-L.M., Warsame, M., Umulisa, N., Ngamije, D., Munyaneza, T., Mazarati, J.-B., Munguti, K., Campagne, P., Criscuolo, A., Ariey, F., Murindahabi, M., Ringwald, P., Fidock, D.A., Mbituyumuremyi, A., Menard, D., 2020. Emergence and clonal expansion of in vitro artemisinin-resistant Plasmodium falciparum kelch13 R561H mutant parasites in Rwanda. Nature Medicine 26, 1602–1608. https://doi.org/10.1038/s41591-020-1005-2
Vaidya, A.B., Mather, M.W., 2000. Atovaquone resistance in malaria parasites. Drug Resist Updat 3, 283–287. https://doi.org/10.1054/drup.2000.0157
van der Oost, J., Westra, E.R., Jackson, R.N., Wiedenheft, B., 2014. Unravelling the structural and mechanistic basis of CRISPR–Cas systems. Nature Reviews Microbiology 12, 479–492. https://doi.org/10.1038/nrmicro3279
van der Ploeg, J.R., 2009. Analysis of CRISPR in Streptococcus mutans suggests frequent occurrence of acquired immunity against infection by M102-like bacteriophages. Microbiology 155, 1966–1976. https://doi.org/10.1099/mic.0.027508-0
van Overbeek, M., Capurso, D., Carter, M.M., Thompson, M.S., Frias, E., Russ, C., Reece-Hoyes, J.S., Nye, C., Gradia, S., Vidal, B., Zheng, J., Hoffman, G.R., Fuller, C.K., May, A.P., 2016. DNA Repair Profiling Reveals Nonrandom Outcomes at Cas9-Mediated Breaks. Mol Cell 63, 633–646. https://doi.org/10.1016/j.molcel.2016.06.037
Venugopal, K., Hentzschel, F., Valkiūnas, G., Marti, M., 2020. Plasmodium asexual growth and sexual development in the haematopoietic niche of the host. Nat Rev Microbiol 18, 177–189. https://doi.org/10.1038/s41579-019-0306-2
Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17, 261–272. https://doi.org/10.1038/s41592-019-0686-2
Vreysen, M.J.B., Saleh, K.M., Ali, M.Y., Abdulla, A.M., Zhu, Z.-R., Juma, K.G., Dyck, V.A., Msangi, A.R., Mkonyi, P.A., Feldmann, H.U., 2000. Glossina austeni (Diptera: Glossinidae) Eradicated on the Island of Unguja, Zanzibar, Using the Sterile Insect Technique. Journal of Economic Entomology 93, 123–135. https://doi.org/10.1603/0022-0493-93.1.123
Walton, R.T., Christie, K.A., Whittaker, M.N., Kleinstiver, B.P., 2020. Unconstrained genome targeting with near-PAMless engineered CRISPR-Cas9 variants. Science 368, 290–296. https://doi.org/10.1126/science.aba8853
Wang, H., Xu, X., 2017. Microhomology-mediated end joining: new players join the team. Cell & Bioscience 7, 6. https://doi.org/10.1186/s13578-017-0136-8
Weetman, D., Djogbenou, L.S., Lucas, E., 2018. Copy number variation (CNV) and insecticide resistance in mosquitoes: evolving knowledge or an evolving problem? Curr Opin Insect Sci 27, 82–88. https://doi.org/10.1016/j.cois.2018.04.005
Wei, D.S., Rong, Y.S., 2007. A Genetic Screen For DNA Double-Strand Break Repair Mutations in Drosophila. Genetics 177, 63–77. https://doi.org/10.1534/genetics.107.077693
Weinfeld, M., Mani, R.S., Abdou, I., Aceytuno, R.D., Glover, J.N.M., 2011. Tidying up loose ends: the role of polynucleotide kinase/phosphatase in DNA strand break repair. Trends Biochem Sci 36, 262–271. https://doi.org/10.1016/j.tibs.2011.01.006
White, G.B., 1985. Anopheles bwambae sp.n., a malaria vector in the Semliki Valley, Uganda, and its relationships with other sibling species of the An. gambiae complex (Diptera:

Culicidae). Systematic Entomology 10, 501–522. https://doi.org/10.1111/j.1365-3113.1985.tb00155.x
WHO, 2021. Fact sheet about Malaria [WWW Document]. URL https://www.who.int/news-room/fact-sheets/detail/malaria (accessed 6.21.21).
WHO, 2020. WHO World Malaria Report 2020, Malaria report 2020.
WHO, 2012. Global plan for insecticide resistance management in malaria vectors. World Health Organization.
WHO, Global Malaria Programme, 2015. Global technical strategy for malaria, 2016-2030.
Williams, G.J., Hammel, M., Radhakrishnan, S.K., Ramsden, D., Lees-Miller, S.P., Tainer, J.A., 2014. Structural insights into NHEJ: building up an integrated picture of the dynamic DSB repair super complex, one component and interaction at a time. DNA Repair (Amst) 17, 110–120. https://doi.org/10.1016/j.dnarep.2014.02.009
Willis, K., Burt, A., 2021. Double drives and private alleles for localised population genetic control. PLOS Genetics 17, e1009333. https://doi.org/10.1371/journal.pgen.1009333
Windbichler, N., Menichelli, M., Papathanos, P.A., Thyme, S.B., Li, H., Ulge, U.Y., Hovde, B.T., Baker, D., Monnat, R.J., Burt, A., Crisanti, A., 2011. A synthetic homing endonuclease-based gene drive system in the human malaria mosquito. Nature 473, 212–215. https://doi.org/10.1038/nature09937
Windbichler, N., Papathanos, P.A., Catteruccia, F., Ranson, H., Burt, A., Crisanti, A., 2007. Homing endonuclease mediated gene targeting in Anopheles gambiae cells and embryos. Nucleic Acids Res 35, 5922–5933. https://doi.org/10.1093/nar/gkm632
Windbichler, N., Papathanos, P.A., Crisanti, A., 2008. Targeting the X Chromosome during Spermatogenesis Induces Y Chromosome Transmission Ratio Distortion and Early Dominant Embryo Lethality in Anopheles gambiae. PLoS Genetics 4, e1000291. https://doi.org/10.1371/journal.pgen.1000291
Wu, X., Scott, D.A., Kriz, A.J., Chiu, A.C., Hsu, P.D., Dadon, D.B., Cheng, A.W., Trevino, A.E., Konermann, S., Chen, S., Jaenisch, R., Zhang, F., Sharp, P.A., 2014. Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. Nat Biotechnol 32, 670–676. https://doi.org/10.1038/nbt.2889
WWARN Artemisinin based Combination Therapy (ACT) Africa Baseline Study Group, 2015. Clinical determinants of early parasitological response to ACTs in African patients with uncomplicated falciparum malaria: a literature review and meta-analysis of individual patient data. BMC Medicine 13, 212. https://doi.org/10.1186/s12916-015-0445-x
Wyss, J.H., 2000. Screwworm eradication in the Americas. Annals of the New York Academy of Sciences 916, 186–193.
Xiao, Y., Ng, S., Nam, K.H., Ke, A., 2017. How type II CRISPR–Cas establish immunity through Cas1–Cas2-mediated spacer integration. Nature 550, 137–141. https://doi.org/10.1038/nature24020
Xu, X., Duan, D., Chen, S.-J., 2017. CRISPR-Cas9 cleavage efficiency correlates strongly with target-sgRNA folding stability: from physical mechanism to off-target assessment. Sci Rep 7, 143. https://doi.org/10.1038/s41598-017-00180-1
Yosef, I., Goren, M.G., Qimron, U., 2012. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. Nucleic Acids Research 40, 5569–5576. https://doi.org/10.1093/nar/gks216
Zavadskas, E.K., Turskis, Z., Kildienė, S., 2014. State of art surveys of overviews on MCDM/MADM methods. Technological and Economic Development of Economy 20, 165–179. https://doi.org/10.3846/20294913.2014.892037
Zetsche, B., Gootenberg, J.S., Abudayyeh, O.O., Slaymaker, I.M., Makarova, K.S., Essletzbichler, P., Volz, S.E., Joung, J., van der Oost, J., Regev, A., Koonin, E.V., Zhang, F., 2015. Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. Cell 163, 759–771. https://doi.org/10.1016/j.cell.2015.09.038
Zhang, Y., Heidrich, N., Ampattu, B.J., Gunderson, C.W., Seifert, H.S., Schoen, C., Vogel, J., Sontheimer, E.J., 2013. Processing-independent CRISPR RNAs limit natural transformation in Neisseria meningitidis. Mol. Cell 50, 488–503. https://doi.org/10.1016/j.molcel.2013.05.001
Zulkower, V., Rosser, S., 2020. DNA Features Viewer: a sequence annotation formatting and plotting library for Python. Bioinformatics 36, 4350–4352. https://doi.org/10.1093/bioinformatics/btaa213

Appendix A

Licences and Permissions

Page	Name of work	Source of work	Copyright holder	Requested	Granted	Permission note
21	Figure 1 - Geography, vector status, and molecular phylogeny of 16 anopheline mosquitoes and selected other dipterans.	Highly evolvable malaria vectors: The genomes of 16 Anopheles mosquitoes	© 2015, AAAS Neafsey et al. 2015	22. 07. 2021	Yes	Allowed free use for dissertations
48	Figure 4 - MMEJ and NHEJ repair mechanisms.	Risky business: Microhomology-mediated end joining	© 2016, Elsevier Supriya et al.	11. 02. 2022	Yes	Allowed free use for dissertations
59	Figure 6 - The principal building blocks of CRISPR-Cas system types.	Origins and evolution of CRISPR-Cas systems	© 2019, Published by the Royal Society under the terms of the Creative Commons Attribution License Koonin and Makarova, 2019	22. 07. 2021		Creative Commons Attribution (CC BY 4.0)
60	Figure 7 - The repurposed CRISPR-Cas9 system for genome editing.	What is CRISPR/Cas9? Archives of Disease in Childhood - Education and Practice vol.101, pp.213-215 (2016)	© 2016, BMJ Publishing Group Ltd. Redman et al. 2016 Open Access article distributed under the terms of CC BY.	22. 07. 2021		Creative Commons Attribution (CC BY 4.0)

Appendix B

Publications

The following publication is attached to the next page.

Kranjc, Nace, Andrea Crisanti, Tony Nolan, and Federica Bernardini. 2021. "Anopheles Gambiae Genome Conservation as a Resource for Rational Gene Drive Target Site Selection." *Insects* 12 (2): 97. https://doi.org/10.3390/insects12020097.

insects

MDPI

Article

Anopheles gambiae Genome Conservation as a Resource for Rational Gene Drive Target Site Selection

Nace Kranjc [1], Andrea Crisanti [1], Tony Nolan [2,*] and Federica Bernardini [1,*]

[1] Department of Life Sciences, Imperial College, London SW7 2AZ, UK; n.kranjc@imperial.ac.uk (N.K.); a.drcrisanti@imperial.ac.uk (A.C.)
[2] Department of Vector Biology, Liverpool School of Tropical Medicine, Liverpool L3 5QA, UK
* Correspondence: tony.nolan@lstmed.ac.uk (T.N.); f.bernardini11@imperial.ac.uk (F.B.)

Simple Summary: Malaria is a huge public health burden that affects predominantly sub-Saharan Africa and is transmitted by *Anopheles* mosquitoes. As a measure for population control, a method called gene drive has been recently developed, which relies on genetic engineering to introduce specific genetic traits into mosquito populations. Gene drives are designed to insert at specific target sites in the mosquito genome. The efficacy of gene drives greatly depends on the selection of appropriate target sites that are functionally or structurally constrained and less likely to tolerate mutations that can hinder the spread of the desired trait in the population. The aim of this study was to perform a genome-wide analysis of highly conserved genomic regions in *Anopheles gambiae* and introduce a measure of conservation that could indicate sites of functional or structural constraint. The results of this analysis are gathered in a publicly available dataset that can support gene drive target selection and can offer further insights in the nature of conserved genomic regions.

Abstract: The increase in molecular tools for the genetic engineering of insect pests and disease vectors, such as *Anopheles* mosquitoes that transmit malaria, has led to an unprecedented investigation of the genomic landscape of these organisms. The understanding of genome variability in wild mosquito populations is of primary importance for vector control strategies. This is particularly the case for gene drive systems, which look to introduce genetic traits into a population by targeting specific genomic regions. Gene drive targets with functional or structural constraints are highly desirable as they are less likely to tolerate mutations that prevent targeting by the gene drive and consequent failure of the technology. In this study we describe a bioinformatic pipeline that allows the analysis of whole genome data for the identification of highly conserved regions that can point at potential functional or structural constraints. The analysis was conducted across the genomes of 22 insect species separated by more than hundred million years of evolution and includes the observed genomic variation within field caught samples of *Anopheles gambiae* and *Anopheles coluzzii*, the two most dominant malaria vectors. This study offers insight into the level of conservation at a genome-wide scale as well as at per base-pair resolution. The results of this analysis are gathered in a data storage system that allows for flexible extraction and bioinformatic manipulation. Furthermore, it represents a valuable resource that could provide insight into population structure and dynamics of the species in the complex and benefit the development and implementation of genetic strategies to tackle malaria.

Keywords: *Anopheles gambiae* complex; vector control strategies; gene drive; genetic variability; genome conservation; genetic resistance

check for updates

Citation: Kranjc, N.; Crisanti, A.; Nolan, T.; Bernardini, F. *Anopheles gambiae* Genome Conservation as a Resource for Rational Gene Drive Target Site Selection. *Insects* **2021**, *12*, 97. https://doi.org/10.3390/insects12020097

Received: 18 December 2020
Accepted: 19 January 2021
Published: 23 January 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Malaria is a threat to life, health and wellbeing. It is a parasitic disease transmitted exclusively by mosquito species from the *Anopheles* genus. In 2018 there were 228 million cases recorded worldwide and 405,000 of these resulted in death [1]. The investment

in malaria control programs relying on insecticide-treated mosquito nets (ITNs), indoor residual spraying and artemisinin-based combination drug therapies has been crucial for the decline of malaria incidents between 2000 and 2015, with the vast majority of this decline being attributed to the control of mosquito numbers [2]. However, from 2014 to 2018, the rate of change slowed dramatically due to the rise of resistance to drugs and insecticides [1].

In sub-Saharan Africa, where the malaria burden is highest, the majority of malaria transmission is caused by members of the *Anopheles gambiae* species complex. Eight species are currently included in this complex of which three, *Anopheles gambiae*, *Anopheles coluzzii* and *Anopheles arabiensis*, are responsible for 95% of malaria transmission in Africa [3]. Recently, a new species, *Anopheles fontenillei*, was discovered in Central Africa and attributed to the complex [4]. Transgenesis tools for *Anopheles* mosquitoes have been in use since the turn of the century [5,6] and in recent years, a number of genetic control methods have been generated that have proven, in a laboratory setting, to have great potential to tackle malaria [7–11]. Among the genetic control methods available for suppression or alteration of *Anopheles* mosquito populations, technologies based on gene drive systems seem to be the most promising. Naturally occurring gene drives alter Mendelian rules of segregation by increasing the probability that a specific allele will be included in the gametes that contribute to the offspring of any one individual. This higher than normal rate of inheritance allows the rapid spread of that specific allele through the population [12–14].

In recent years, genome editing technologies have facilitated the development of synthetic gene drives designed to enable the spread of altered traits mimicking the dynamics of the gene drives found in nature [10,11,15–17]. Due to their simple molecular components and efficiency, CRISPR/Cas9-based homing gene drives have quickly become a system of choice in this context. This technology is based on the presence of a bacterial endonuclease (Cas9) and a guide RNA (gRNA) that is responsible for directing the sequence specificity of Cas9 [18]. A genetic construct encoding the units of the Cas9-gRNA complex can be inserted into the genome of an organism to cause disruption of a target gene. The driving nature of this system relies on the ability of the gRNA to direct Cas9 to cut the wild-type version of the target gene on the homologous chromosome. This cut triggers the activation of the cell DNA break repair machinery, which will use the homologous regions flanking the genetic construct as a template to repair the double strand break. The repair mechanism can thus lead to the copy of the gene drive construct (and any encoded trait) and the conversion of the cell to a homozygous genotype. When this process happens in the germline it results in gametes with a higher than 50% probability of containing the gene drive, which in turn causes the allele to increase in frequency each generation and spread through the population. We have previously shown that a CRISPR-Cas9 gene drive construct can be used to target genes with essential roles in the female fertility of *Anopheles* mosquitoes [9]. However, the high fitness cost imposed by this type of gene drive can create a strong selection pressure for mutations that restore function to the target gene yet are resistant to CRISPR-Cas9 cleavage. This highlights the requirement, when designing gene drive systems, for a meticulous selection of genic target sequences that are under high levels of functional constraint and therefore less likely to tolerate resistant mutations. Indeed, in caged populations, suppression was achieved only when *doublesex*, a gene involved in sex determination and highly conserved within the *Anopheles* complex was targeted, showing that this aspect is crucial for the success of mosquito population control through this approach [11,19–21].

High levels of sequence conservation can be a good proxy for functional constraint. Recently, high-throughput DNA sequencing of wild-caught *Anopheles* mosquitoes revealed complex population structure, patterns of gene flow and a genetic diversity higher than expected [22–24]. Here we describe a bioinformatics pipeline that integrates a systematic analysis of the data on genetic variation in more than 1000 wild-caught *Anopheles gambiae* individuals and conserved syntenic regions of 19 *Anopheles* species, and three phylogenetically more distant species within the same Diptera order. This user-friendly pipeline

performs a sliding window analysis across the whole genome for every possible gene drive target sequence, attributing to it a conservation score (Cs), which is an indicator of their tendency to tolerate mutations. Cs incorporates interspecies variation, intraspecies variation and a measure of the selection forces applying there. This should prove a useful resource when selecting for functionally constrained target site within genes of interest in the context of vector control via gene drive. In addition, the compiled dataset offers the flexibility to identify conserved genomic regions within species subsets, which will be of significant interest for the implementation of any vector control strategies in geographical areas populated by sympatric species and where one wishes to target only a particular species.

2. Materials and Methods

2.1. Genomic Data

Genomic data of 19 *Anopheles* species, *Aedes aegypti*, *Culex quinquefasciatus* and *Drosophila melanogaster* were used in the analysis. The following reference genomes were obtained from VectorBase [25]: AgamP4 (*An. gambiae*), AcolM1 (*An. coluzzii*), AaraD1 (*An. arabiensis*), AquaS1 (*An. quadriannulatus*), AmelC2 (*An. melas*), AmerM2 (*An. merus*), AepiE1 (*An. epiroticus*), AchrA1 (*An. christyi*), AsinC2 (*An. sinensis*), AminM1 (*An. minimus*), AmacM1 (*An. maculatus*), AculA1 (*An. culicifacies*), AsteI2 (*An. stephsnsi*), AfunF1 (*An. funestus*), AatrE3 (*An. atroparvus*), AdirW1 (*An. dirus*), AfarF2 (*An. farauti*), AdarC3 (*An. darlingi*), AalbS2 (*An. albimanus*), AaegL5 (*Aedes aegypti*), CpipJ2 (*Culex quinquefasciatus*). The reference genome of *Drosophila melanogaster* DmelP6 was obtained from Flybase [26].

2.2. Genome Alignment

To identify conserved regions, we used the CNEr package [26], which can be used to detect highly conserved noncoding elements (CNE). CNEr streamlines the process of large-scale identification of CNEs. CNEr uses LASTZ genome aligner together with *axtChain*, *chainMergeSort*, *chainPreNet*, *chainNet*, *netSyntenic*, *netToAxt* and *axtSort* tools for chaining and netting. Genome alignment, chaining and netting were performed by using default parameters (C = 0 E = 30 H = 0 K = 3000 L = 3000 M = 50 O = 400 T = 1 Y = 9400). All of the selected genomes were aligned and scanned for highly identical regions relative to *An. gambiae* AgamP4 genome with the pipeline provided by CNEr. Two sets of identity thresholds were used: 90%, 97%, 100% and 70%, 90%, 96%, 98%, 100% identity over a scanning window of 30 and 50 bp, respectively. Although CNEr was designed for detection of conserved noncoding elements, by omitting the filtering step it retained detected conserved elements in the coding regions as well. This procedure yielded a list of conserved genomic regions between each selected genome and *An. gambiae* where the sequence identity was higher than a threshold set in each scan. The minimal detected identity was set to 70%. For example, an identity value of 98% corresponds to detecting 49 identical positions in the 50 bp alignment window.

2.3. Conservation Score Calculation

Genomic intervals with their corresponding CNEr identity values were mapped to a matrix of the size of each chromosome, allowing for per-base-pair resolution of identity scores and easier comparison across different species. Additionally, positions of SNPs found in 1142 wild-caught individuals within the Ag1000g phase 2 project [23,27] were mapped in the same data table and average SNP density in the sliding window of 20 bp was calculated across all chromosomes.

To calculate an overall conservation score (Cs), identity values were normalised with the phylogenetic distance between each species and *An. gambiae.* The phylogenetic tree was generated from the genome alignments using PHAST phyloFit [28]. Phylogenetic distances within the tree were calculated using ETE Toolkit [29]. Average normalised identity values were calculated for each position and SNP density data were used to skew values higher when SNP density was low. Furthermore, phyloP was used to calculate

scores with likelihood ratio test (method LRT) representing conservation or acceleration (mode CONACC) for each genomic position. Incorporating phyloP scores into the model indicated whether positions within highly identical genomic regions are under negative or positive selection, based on the phylogenetic model of neutral evolution (REV substitution model).

A matrix of normalised identities was created (Equation (1)), where index m is the number of genomes in comparison, index n is the length of a chromosome and d_m is the phylogenetic distance between genome m and AgamP4.

$$\tilde{\mathrm{I}}_{m,n} = \begin{bmatrix} I_{1,1} & \cdots & I_{1,n} \\ \vdots & \ddots & \vdots \\ I_{m,1} & \cdots & I_{m,n} \end{bmatrix} \begin{bmatrix} d_1 \\ \vdots \\ d_m \end{bmatrix} \quad (1)$$

Normalised identities were averaged for each nucleotide position n in each chromosome. SNP density σ calculated for a sliding window of 20 bp on each position n. PhyloP score for position n is denoted by p_n (Equation (2)).

$$Cs_n = \frac{\sum_{i=0}^{m} \tilde{\mathrm{I}}_{i,m}}{m} \frac{1-\sigma_n}{1+\sigma_n} \frac{1}{2^{-p_n}} \quad (2)$$

Finally, *Cs* values were scaled using MinMax scaler to a range between 0 and 1, where 1 means high conservation per chromosome arm. Normalisation was also performed on the whole genome and showed negligible difference in values. All calculations were done using Python modules Pandas, Numpy, Scipy, Zarr, h5py, scikit-learn and scikit-allel [30]. The data were stored in HDF5 format and plotted using Matplotlib and Seaborn. Genome accessibility data were obtained from Ag1000g phase 2 [23,27].

The dataset produced in the analysis can be found at https://zenodo.org/record/4304586. A Python tool for easier accessibility of the results was created and can be found together with the scripts used to generate the dataset at https://github.com/nkran/AgamP4_conservation_score.

To account for the missing data, a list of regions and the multisequence alignment depth, indicating the number of analyzed assemblies aligned across the genome, was gathered and deposited at https://github.com/nkran/AgamP4_conservation_score/tree/master/data.

3. Results

3.1. Conservation Score Calculation

We selected available reference genomes of 18 *Anopheles* species, namely *An. coluzzii*, *An. arabiensis*, *An. quadriannulatus*, *An. melas*, *An. merus*, *An. epiroticus*, *An. christyi*, *An. sinensis*, *An. minimus*, *An. maculatus*, *An. culicifacies*, *An. stephensi*, *An. funestus*, *An. atroparvus*, *An. dirus*, *An. farauti*, *An. darlingi*, *An. albimanus*, as well as reference genomes of the more distantly related species from the order Diptera: *Aedes aegypti*, *Culex quinquefasciatus* and *Drosophila melanogaster*. Genome sequences of each species were aligned to *An. gambiae* reference genome to find syntenic regions for each pair. Syntenic regions obtained from the alignments were analyzed to identify genomic intervals of highly conserved regions (>70%) relative to *An. gambiae*. We then calculated the average identity across all species included in the analysis and normalised the values based on the phylogenetic distance between *An. gambiae* and the other species. Normalisation was performed to balance the sequence identity scores, giving more weight to the same identity scores in species that are phylogenetically more distant than the species that are phylogenetically closer to *An. gambiae*.

To account for the evolutionary selection, we incorporated a measure of positive or negative selection, PhyloP score, in the calculation of Cs for each position in the genome. The PhyloP score is calculated by taking into account a substitution model of neutral

selection and it is used to perform a statistical test at each position in the multiple alignment for a significant increase or decrease in the rate of substitution.

Additionally, intraspecies variation was included in the calculation of the final conservation score (Cs), where a value of Cs = 1.0 represents loci with the highest conservation. Cs was generated as a measure of conservation with a base-pair resolution that takes into account genomic variation of compared species and intraspecific variation observed in 1142 wild-caught mosquito individuals.

To investigate differences in genome conservation of different species, we directed our analysis toward two groups, one including five sibling species within the *Anopheles gambiae* complex (*An. coluzzii*, *An. arabiensis*, *An. quadriannulatus*, *An. melas*, *An. merus*), and the other including five more distantly related species (*An. darlingi*, *An. albimanus*, *Aedes aegypti*, *Culex quinquefasciatus*, *Drosophila melanogaster).*

The genome identity at the chromosomal level was investigated and compared between the two clusters. We calculated mean identity of 81.0% across the genomes of species within the *Anopheles gambiae* complex (84.4% in autosomes and 67.2% in X chromosome) and 4.8% for the genomes of distantly related species (5.12% in autosomes and 3.69% in X chromosomes). Both groups showed decreased conservation toward the centromeric end of each chromosome and on X chromosome compared to autosomes (Figure 1).

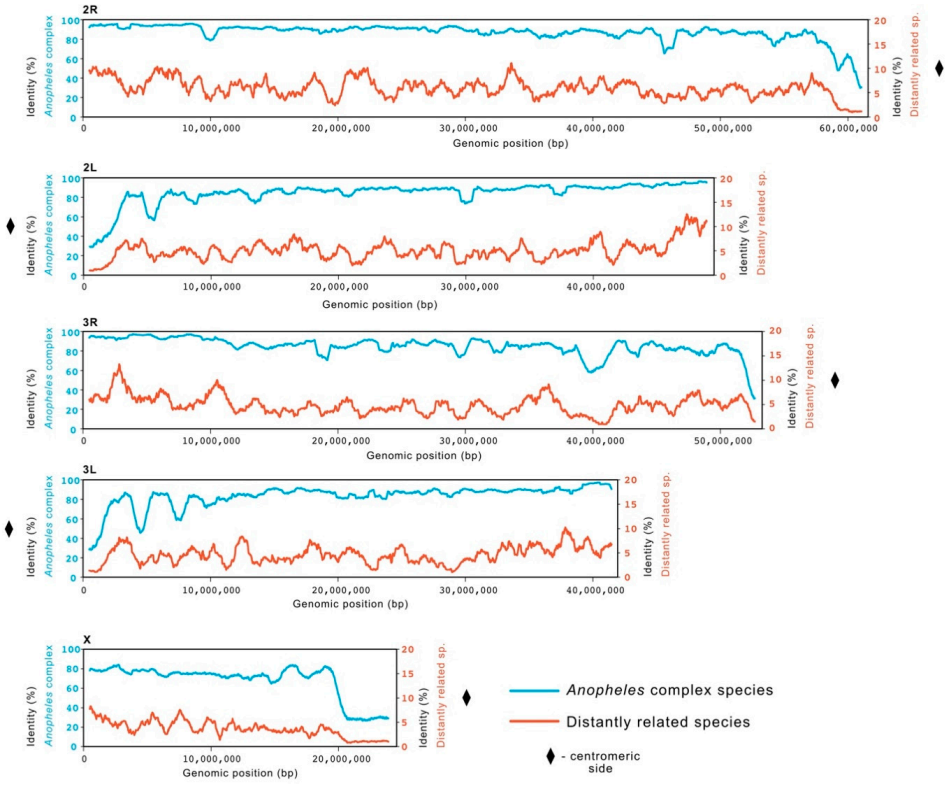


Figure 1. Conservation across *An. gambiae* chromosomes. A sliding window of length 1 Mb was used to calculate average identity (in percentage) for two groups of species in comparison to *An. gambiae*. A comparison with a group of five species of the *Anopheles gambiae* complex (in blue; left scale 0–100%), namely *An. coluzzii*, *An. arabiensis*, *An. quadriannulatus*, *An. melas* and *An. merus*, showed high average identity (81%) across the genome. The group of species more distantly related to *An. gambiae* (in red; right scale 0–20%), namely *An. darlingi*, *An. albimanus*, *Aedes aegypti*, *Culex quinquefasciatus*, *Drosophila melanogaster*, showed lower average identity (<5%). A decrease in conservation in both groups can be observed closer to the centromeric regions of each chromosome arm (represented by ⧫).

Additionally, to test the bias introduced by the inclusion of the five more distantly related species on the overall Cs performance, Cs was calculated by omitting these species

from the analysis. A strong correlation (Pearson's $r > 0.83$, $p < 0.001$) was observed between both Cs calculations. Inclusion of the five species skews the Cs values, however the highest-ranking positions are not considerably affected (Figure A2).

3.2. Conservation of Gene Drive Target Sites

Homing-based gene drives targeting female fertility have shown varying degrees of efficacy, in part related to the propensity of the target genes to tolerate resistant mutations. With the rationale that those target genes may have additional gene drive target sites less prone to the development of resistance, and as a proof of principle for the pipeline developed herein, we scanned for alternative target sites and compared conservation scores among them in order to identify possible regions of high functional constraint. We focused our analysis on two genes previously selected for the development of gene drive technologies against *An. gambiae* mosquitoes; the female-fertility gene *AGAP007280*, whose original target site was very prone to resistance, and *doublesex*, a gene whose female-specific isoform can be targeted by a gene drive without, to date, selection for resistance.

Consistent with the observation that it could tolerate a range of viable mutations, the original 18 bp-long target site that was chosen to disrupt gene function of the female fertility gene *AGAP007280* (2L:44994011–44994029) had a low conservation score with a mean Cs = 0.06 (Cs_{min} = 0.01, Cs_{max} = 0.24) (Figure 2A). We calculated the conservation score for all 987 SpCas9 accessible target sites in this gene. We found an alternative target site in exon 6 (2L:44993662–44993685) with the highest mean conservation score of 0.44 (Cs_{min} = 0.05, Cs_{max} = 0.73). Mean Cs of 0.06 and 0.44 correspond to the 57th and 94th percentile of all Cs values (Cs > 0), respectively, for chromosome arm 2L (Figure A1). This target site is located in a locus that is part of a conserved region annotated as peptidase S1 domain, which spans exons 4, 5 and 6.

In comparison, the previously validated gene drive target site spanning intron 4 and the female-specific exon (exon 5) of the *dsx* gene (*AGAP004050*), shown to be essential for female sex determination and intolerant of mutations, showed high mean Cs of 0.43 (Cs_{min} = 0.17, Cs_{max} = 0.68). The same site was 100% conserved across the *Anopheles gambiae* complex and showed high conservation for three out of the five distantly related species (92.8%, 86.5%, 86.2% in *An. albimanus*, *Aedes Aegypti* and *Culex quinquefasciatus*, respectively). Interestingly, the whole of exon 5 has a relatively high mean conservation score (Cs = 0.50) across its relatively short (87 bp) length with only one SNP present (Figure 2B). This allowed us to identify eight alternative SpCas9 accessible sites in exon 5 (Table 1) of which the highest-ranking alternative target site (2R:48714594–48714616) had an even higher conservation score Cs = 0.60 (Cs_{min} = 0.26, Cs_{max} = 0.70) than the original gene drive target site. Both mean Cs values for the previously validated gene drive target site and the alternative target site in exon 5 of *dsx* are in the top 10th percentile of all Cs values for chromosome arm 2R.

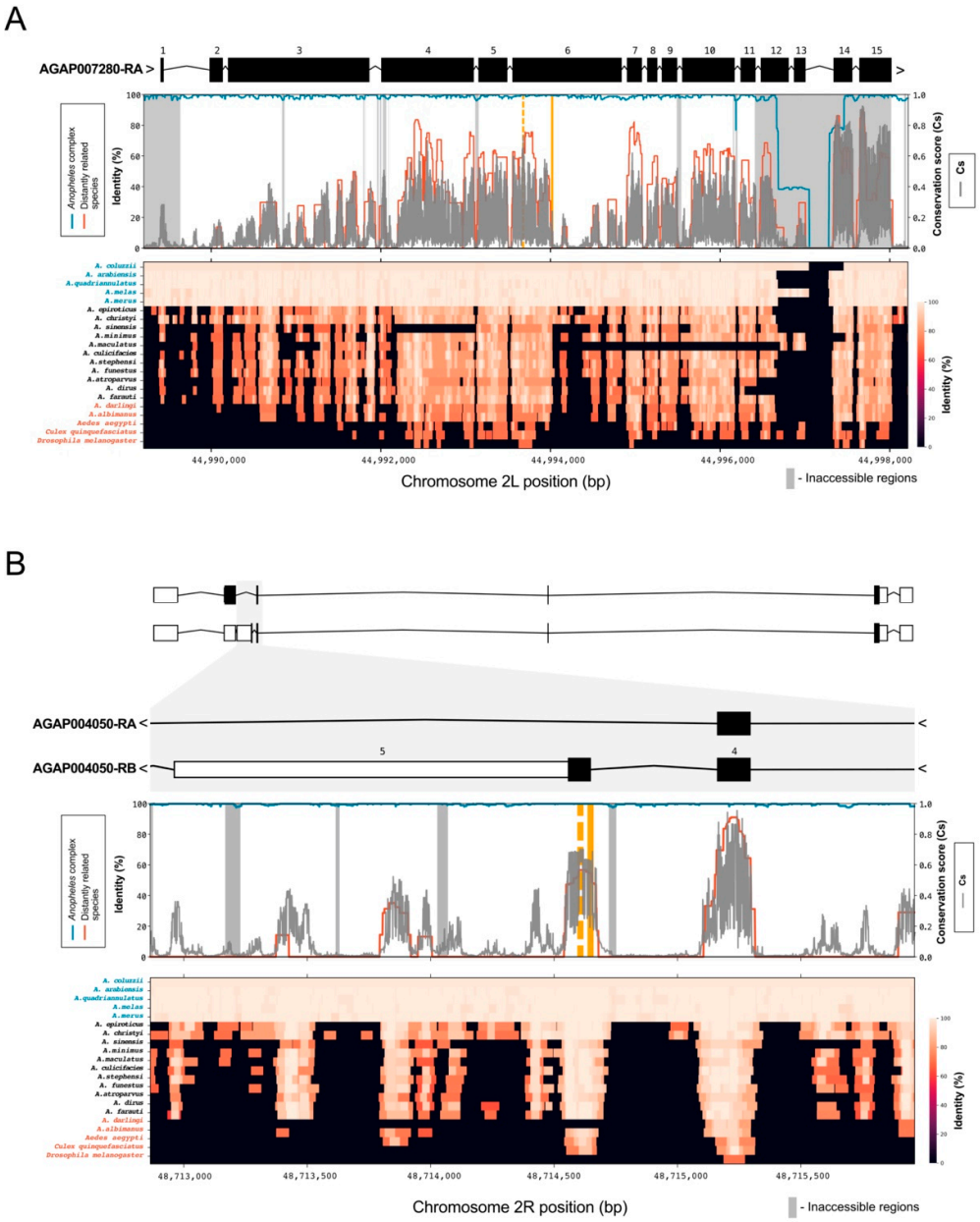


Figure 2. Conservation at *AGAP007280* and *AGAP004050* (*dsx*) gene drive target sites. (**A**) Top panel: the gene drive target site [9] in exon 6 of *AGAP007280* gene is highlighted (solid orange line). Middle panel: average identity (in percentage) was calculated for the five genome species within the *Anopheles gambiae* complex (blue line) and the genomes of the five most distantly related species included in the analysis (red line). Conservation score (Cs) in gray was also calculated, considering all 21 analyzed species and *An. gambiae* variation. The gray blocks are masked regions considered as inaccessible [22,23] due to the low confidence of genotyping. Bottom panel: heatmap represents the identity (in percentage) for each analyzed species. Species from the *An. gambiae* complex are highlighted in blue, mosquito species most distantly related from *An. gambiae* are highlighted in red. (**B**) Target site for *AGAP004050* (*dsx*) falls between intron 4 and exon 5 [11] and it is marked in orange. With the exception of two species (*An. darlingi* and *D. melanogaster*), the target site shows high conservation (>70%) for the *Anopheles gambiae* complex and more distantly related species. The alternative target sites, identified in this study, are marked with a dashed orange line and show a higher conservation score.

Table 1. Alternative gene drive target sites in the coding sequence of exon 5 in *dsx* gene. The existing gene drive target site is highlighted.

Chromosome	Start	End	Strand	Cs	Cs_{max}	Cs_{min}
2R	48714594	48714616	–	0.60	0.70	0.26
2R	48714589	48714611	–	0.56	0.70	0.20
2R	48714550	48714572	+	0.51	0.68	0.19
2R	48714551	48714573	+	0.50	0.68	0.19
2R	48714554	48714576	+	0.48	0.68	0.19
2R	48714561	48714583	–	0.45	0.68	0.15
2R	48714637	48714659	-	0.44	0.68	0.17
2R	48714640	48714662	-	0.41	0.68	0.17
2R	48714648	48714670	-	0.36	0.68	0.17

The high level of sequence conservation at the gene drive target site within dsx gene that was observed across species of the *Anopheles* complex as well as more distantly related species, gave the rationale for their inclusion in the analysis.

4. Discussion

The availability of whole genome data for species of the *Anopheles gambiae* complex has enabled us to conduct a bioinformatic analysis where a conservation score, which accounts for interspecies and intraspecies variation, was attributed to genomic regions across these species as well as to more distantly related ones.

A strong pattern of conservation was observed for species of the *Anopheles* complex with an average genome identity of 84% across the autosomes and lower values only falling in those regions of the chromosome annotated as centromeric. This discrepancy is in agreement with a number of works that describe centromeric sequences as highly variable in size and composition [31,32]. Overall, a lower average sequence identity (67%) was observed for the X chromosome, reflecting the expectation that genetic divergence tends to be higher on sex chromosomes [33,34] and confirming the previously shown diversity of this chromosome within the species of the complex [24,35–38]. This conservation pattern across the more distantly related species, albeit at a lower level (4.8%), is consistent with longer evolutionary distances that separate them [24]. These observations confirmed the suitability of using sequence identity deriving from multiple alignment as the base of the proposed conservation model, which was further improved by including the intraspecies variation.

It has recently been shown that, despite high genetic variation between genomes of *Aedes* and *Anopheles* mosquitoes, conserved target sites for Cas9-based gene drives are abundant within protein-coding sequences [39]. We made use of existing bioinformatic methods to account for positive or purifying selection on the genomic sequences used in this study [28] and strengthened our analysis by developing a bioinformatic pipeline that relies on intraspecies genetic variation as well as highly conserved sequences of syntenic regions between several species of dipterans for the identification of ultraconserved regions. This analysis offers a flexible workflow where conserved regions can be identified on the genome-wide scale as well as with per base-pair resolution.

For some previous CRISPR-based gene drives designed to interrupt female fertility genes (e.g., *AGAP007280*), genotypic analysis of mosquito samples collected during cage trial experiments tracking the gene drive over time showed that end-joining repair of CRISPR-induced DNA breaks at the selected target site generated resistance, thus preventing the further spread of the gene drive [9,40]. Highlighting the utility of our pipeline, we were able to identify additional gene drive target sites in *AGAP007280* with a conservation score much higher than the one originally chosen (Cs 0.42 versus 0.06) (Figure 2A). Notably, this score is similar to the conservation score attributed to the previously validated target

site in *dsx*, which has so far failed to yield resistant alleles, consistent with this region of the gene showing functional or structural constraint. Thus, our pipeline offers additional target sites, likely less prone to resistance, in genes previously validated as suitable for targeting by population suppression such as *AGAP007280*. The ability to multiplex gene drives by incorporating several gRNAs recognizing separate target sites in the same target gene is anticipated to further delay the onset of resistance since the probability of individuals being resistant at both target sites is much lower [41–44]. Thus, it is prudent to select multiple sites, each with the highest conservation score where possible. In this respect it is encouraging that alternative target sites with high conservation scores were found when exon 5 of *dsx* gene was analyzed (the highest-ranking target site had Cs = 0.60). The identification of these sites through our pipeline should thus facilitate the design of multiplexed gene drive constructs targeting the *dsx* gene and attempts to future-proof these constructs for release in the field—although to-date the original single target site fails to yield resistant alleles in the laboratory, the situation in the wild, at increased geographical and temporal scale, and with larger population sizes, may be more challenging.

The conservation scores described herein are a good indicator for conserved genomic regions under functional or structural constraint. That said, at any two functionally constrained sequences that show ultraconservation the actual level of constraint applying at each might be considerably different—the force of purifying selection could be sufficient to ensure conservation of sequences where changes caused minimal or intermediate reductions in fitness as it would at sequence changes that were nonviable. The choice between target sites at each extreme of functional constraint very much depends on the type of gene drive approach, and the intended effect of the drive.

It must also be noted that confounding sources might lead to ultraconservation that do not necessarily reflect functional constraint. For instance, repetitive regions of the genome and their representation in the reference genome assemblies can lead to ambiguous genome alignments and false identical regions. We used genome accessibility data generated in the Ag1000g project that includes already annotated, masked, low-complexity genomic regions and DNA sequences with low coverage and poor mapping quality [22,23] to label these regions in the analyzed genes. High conservation score in regions marked as inaccessible (Figure 2) should be considered and taken cautiously as indicator of functional constraint. In addition, many families of repetitive sequences, such as satellite DNA and ribosomal DNA genes, are maintained homogenous within populations by concerted evolution rather than unbreakable constraint [45]. Concerted evolution for such elements should be taken into account when conservation within and among different species is evaluated.

The understanding of the real constraint within the ultraconserved regions identified in this study represents our next step to satisfy the increasing request for suitable target sites for genetic vector control strategies based on gene drive as well as for other approaches to vector control. Furthermore, the dataset and tools generated in this study have the ability to offer a solid base for future insights into the underlying biology of conserved elements in the *An. gambiae* genome.

Author Contributions: Conceptualization, N.K., A.C., T.N. and F.B.; methodology, N.K.; software, N.K.; validation, T.N. and F.B.; formal analysis, N.K.; investigation, N.K. and F.B.; resources, A.C.; data curation, N.K.; writing—original draft preparation, N.K. and F.B.; writing—review and editing, N.K., T.N. and F.B.; visualization, N.K.; supervision, A.C., T.N. and F.B.; project administration, A.C.; funding acquisition, A.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Bill and Melinda Gates Foundation (LCII PSC656).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in Zenodo at 10.5281/zenodo.4304586, reference number 4304586.

Acknowledgments: We would like to thank Alistair Miles for sharing the code for plotting transcripts with us and to Silke Fuchs and John Mumford for their comments.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

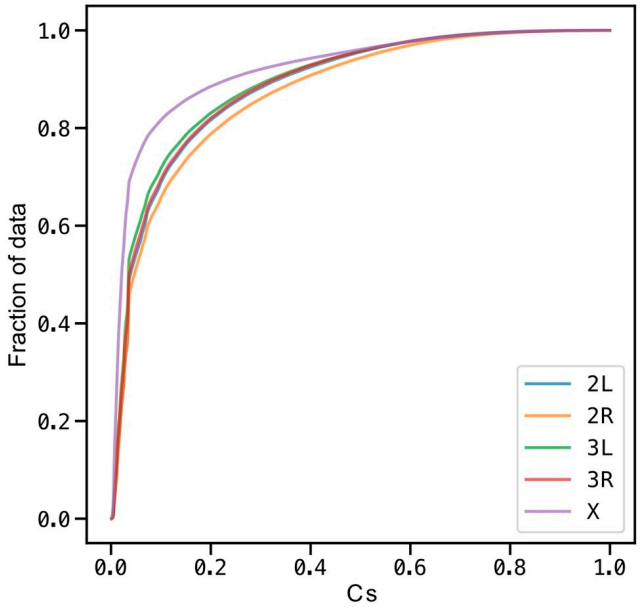


Figure A1. Conservation at CDF was calculated for Cs values (Cs > 0) for each chromosome arm. On average, 90% of the genomic positions that show some conservation have Cs less than 0.32. Cs = 1 occurs on 150 positions across the analyzed genome (chromosome arms 2L, 2R, 3L, 3R and X).

Appendix B

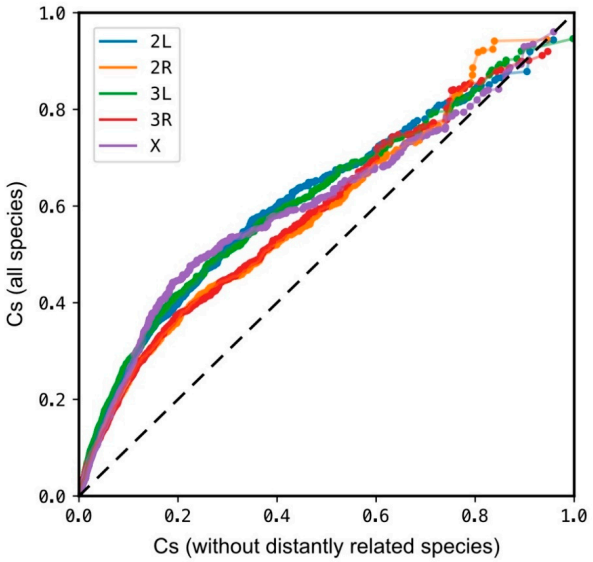


Figure A2. Q–Q plot of comparison between Cs calculated from the genome assembly data from all the species included in the analysis and Cs from a subset of genome assemblies without distantly related species. All chromosome arms show strong correlation between both groups (Pearson's $r > 0.83$, $p < 0.001$). The inclusion of the group of more distantly related genomes skews Cs towards higher values, however, both compared groups showed similar distribution of higher scores. Plot was generated by random sampling of 10,000 genomic positions from each chromosome arm.

References

1. WHO. *World Malaria Report 2018*; World Health Organization: Geneva, Switzerland, 2018; ISBN 9789241565653.
2. Bhatt, S.; Weiss, D.J.; Cameron, E.; Bisanzio, D.; Mappin, B.; Dalrymple, U.; Battle, K.E.; Moyes, C.L.; Henry, A.; Eckhoff, P.A.; et al. The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015. *Nature* **2015**, *526*, 207–211. [CrossRef]
3. Sinka, M.E.; Bangs, M.J.; Manguin, S.; Palis, R.Y.; Chareonviriyaphap, T.; Coetzee, M.; Mbogo, C.M.; Hemingway, J.; Patil, A.P.; Temperley, W.H.; et al. A global map of dominant malaria vectors. *Parasit. Vectors* **2012**, *5*. [CrossRef]
4. Barrón, M.G.; Paupy, C.; Rahola, N.; Akone, E.O.; Ngangue, M.F.; Bahun, W.T.A.; Pombi, M.; Kengne, P.; Costantini, C.; Simard, F.; et al. A new species in the major malaria vector complex sheds light on reticulated species evolution. *Sci. Rep.* **2019**, *9*. [CrossRef]
5. Catteruccia, F.; Nolan, T.; Loukeris, T.G.; Blass, C.; Savakis, C.; Kafatos, F.C.; Crisanti, A. Stable germline transformation of the malaria mosquito *Anopheles stephensi*. *Nature* **2000**, *405*, 959–962. [CrossRef]
6. Grossman, G.L.; Rafferty, C.S.; Clayton, J.R.; Stevens, T.K.; Mukabayire, O.; Benedict, M.Q. Germline transformation of the malaria vector, *Anopheles gambiae*, with the piggyBac transposable element. *Insect Mol. Biol.* **2001**, *10*, 597–604. [CrossRef]
7. Galizi, R.; Doyle, L.A.; Menichelli, M.; Bernardini, F.; Deredec, A.; Burt, A.; Stoddard, B.L.; Windbichler, N.; Crisanti, A. A synthetic sex ratio distortion system for the control of the human malaria mosquito. *Nat. Commun.* **2014**, *5*. [CrossRef]
8. Bernardini, F.; Kriezis, A.; Galizi, R.; Nolan, T.; Crisanti, A. Introgression of a synthetic sex ratio distortion system from *Anopheles gambiae* into *Anopheles arabiensis*. *Sci. Rep.* **2019**, *9*. [CrossRef]
9. Hammond, A.; Galizi, R.; Kyrou, K.; Simoni, A.; Siniscalchi, C.; Katsanos, D.; Gribble, M.; Baker, D.; Marois, E.; Russell, S.; et al. A CRISPR-Cas9 gene drive system targeting female reproduction in the malaria mosquito vector *Anopheles gambiae*. *Nat. Biotechnol.* **2016**, *34*, 78–83. [CrossRef]
10. Gantz, V.M.; Jasinskiene, N.; Tatarenkova, O.; Fazekas, A.; Macias, V.M.; Bier, E.; James, A.A. Highly efficient Cas9-mediated gene drive for population modification of the malaria vector mosquito *Anopheles stephensi*. *Proc. Natl. Acad. Sci. USA* **2015**, *112*. [CrossRef]
11. Kyrou, K.; Hammond, A.M.; Galizi, R.; Kranjc, N.; Burt, A.; Beaghton, A.K.; Nolan, T.; Crisanti, A. A CRISPR–Cas9 gene drive targeting doublesex causes complete population suppression in caged *Anopheles gambiae* mosquitoes. *Nat. Biotechnol.* **2018**, *36*, 1062–1066. [CrossRef]
12. Burt, A. Site-specific selfish genes as tools for the control and genetic engineering of natural populations. *Proc. Biol. Sci.* **2003**, *270*, 921–928. [CrossRef]
13. Burt, A.; Trivers, R. *Genes in Conflict*; Harvard University Press: Cambridge, MA, USA, 2009; ISBN 978-0-674-02911-8.
14. Werren, J.H.; Nur, U.; Wu, C.I. Selfish genetic elements. *Trends Ecol. Evol.* **1988**, *3*, 297–302. [CrossRef]
15. Grunwald, H.A.; Gantz, V.M.; Poplawski, G.; Xu, X.R.S.; Bier, E.; Cooper, K.L. Super-Mendelian inheritance mediated by CRISPR–Cas9 in the female mouse germline. *Nature* **2019**, *566*, 105–109. [CrossRef]
16. Shapiro, R.S.; Chavez, A.; Porter, C.B.M.; Hamblin, M.; Kaas, C.S.; Carlo, J.E.; Zeng, G.; Xu, X.; Revtovich, A.; Kirienko, N.; et al. A CRISPR–Cas9-based gene drive platform for genetic interaction analysis in Candida albicans. *Nat. Microbiol.* **2018**, *3*, 73–82. [CrossRef]
17. Carlo, J.E.; Chavez, A.; Dietz, S.L.; Esvelt, K.M.; Church, G.M. Safeguarding CRISPR-Cas9 gene drives in yeast. *Nat. Biotechnol.* **2015**, *33*, 1250–1255. [CrossRef]
18. Jiang, F.; Doudna, J.A. CRISPR-Cas9 Structures and Mechanisms. *Annu. Rev. Biophys.* **2017**, *46*, 505–529. [CrossRef]
19. Hammond, A.M.; Kyrou, K.; Bruttini, M.; North, A.; Galizi, R.; Karlsson, X.; Kranjc, N.; Carpi, F.M.; Aurizio, R.; Crisanti, A.; et al. The creation and selection of mutations resistant to a gene drive over multiple generations in the malaria mosquito. *PLoS Genet.* **2017**, *13*. [CrossRef]
20. Burtis, K.C.; Baker, B.S. *Drosophila* doublesex gene controls somatic sexual differentiation by producing alternatively spliced mRNAs encoding related sex-specific polypeptides. *Cell* **1989**, *56*, 997–1010. [CrossRef]
21. Graham, P.; Penn, J.K.M.; Schedl, P. Masters change, slaves remain. *Bioessays* **2003**, *25*, 1–4. [CrossRef]
22. Miles, A.; Harding, N.J.; Bottà, G.; Clarkson, C.S.; Antão, T.; Kozak, K.; Schrider, D.R.; Kern, A.D.; Redmond, S.; Sharakhov, I.; et al. Genetic diversity of the African malaria vector *Anopheles gambiae*. *Nature* **2017**, *552*, 96–100. [CrossRef]
23. The Anopheles gambiae 1000 Genomes Consortium. Genome variation and population structure among 1142 mosquitoes of the African malaria vector species *Anopheles gambiae* and *Anopheles coluzzii*. *Genome Res.* **2020**. [CrossRef]
24. Neafsey, D.E.; Waterhouse, R.M.; Abai, M.R.; Aganezov, S.S.; Alekseyev, M.A.; Allen, J.E.; Amon, J.; Arca, B.; Arensburger, P.; Artemov, G.; et al. Highly evolvable malaria vectors: The genomes of 16 *Anopheles* mosquitoes. *Science* **2015**, *347*. [CrossRef]
25. Calderon, G.G.I.; Emrich, S.J.; Callum, R.M.; Maslen, G.; Dialynas, E.; Topalis, P.; Ho, N.; Gesing, S.; Madey, G.; Collins, F.H.; et al. VectorBase: An updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. *Nucleic Acids Res.* **2015**, *43*, 707–713. [CrossRef]
26. Tan, G.; Polychronopoulos, D.; Lenhard, B. CNEr: A toolkit for exploring extreme noncoding conservation. *PLoS Comput. Biol.* **2019**, *15*. [CrossRef]
27. The Anopheles Gambiae 1000 Genomes Consortium. Ag1000G Phase 2 AR1 Data Release. *MalariaGEN*. 2017. Available online: http://www.malariagen.net/data/ag1000g-phase2-ar1 (accessed on 22 January 2021).
28. Hubisz, M.J.; Pollard, K.S.; Siepel, A. PHAST and RPHAST: Phylogenetic analysis with space/time models. *Brief. Bioinform.* **2010**, *12*, 41–51. [CrossRef]

29. Cepas, H.J.; Serra, F.; Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **2016**, *33*, 1635–1638. [CrossRef]
30. Miles, A.; pyup.io bot; Murillo, R.; Ralph, P.; Harding, N.; Pisupati, R.; Rae, S.; Millar, T. cggh/scikit-allel: v1.3.2. 2020. Available online: https://zenodo.org/record/3976233#.YAvTZRYRVPY (accessed on 22 January 2021). [CrossRef]
31. Henikoff, S.; Ahmad, K.; Malik, H.S. The Centromere Paradox: Stable Inheritance with Rapidly Evolving DNA. *Science* **2001**, *293*, 1098–1102. [CrossRef]
32. Thomae, A.W.; Schade, G.O.M.; Padeken, J.; Borath, M.; Vetter, I.; Kremmer, E.; Heun, P.; Imhof, A. A pair of centromeric proteins mediates reproductive isolation in *Drosophila* species. *Dev. Cell* **2013**, *27*, 412–424. [CrossRef]
33. Hall, A.B.; Papathanos, P.A.; Sharma, A.; Cheng, C.; Akbari, O.S.; Assour, L.; Bergman, N.H.; Cagnetti, A.; Crisanti, A.; Dottorini, T.; et al. Radical remodeling of the Y chromosome in a recent radiation of malaria mosquitoes. *Proc. Natl. Acad. Sci. USA* **2016**, *113*. [CrossRef]
34. Sayres, W.M.A. Genetic Diversity on the Sex Chromosomes. *Genome Biol. Evol.* **2018**, *10*, 1064–1078. [CrossRef]
35. Favia, G.; Lanfrancotti, A.; Spanos, L.; Kiamos, S.I.; Louis, C. Molecular characterization of ribosomal DNA polymorphisms discriminating among chromosomal forms of *Anopheles gambiae s.s.*. *Insect Mol. Biol.* **2001**, *10*, 19–23. [CrossRef]
36. Stump, A.D.; Fitzpatrick, M.C.; Lobo, N.F.; Traoré, S.; Sagnon, N.; Costantini, C.; Collins, F.H.; Besansky, N.J. Centromere-proximal differentiation and speciation in *Anopheles gambiae*. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 15930–15935. [CrossRef]
37. Crawford, J.E.; Riehle, M.M.; Guelbeogo, W.M.; Gneme, A.; Sagnon, N.; Vernick, K.D.; Nielsen, R.; Lazzaro, B.P. Reticulate Speciation and Barriers to Introgression in the *Anopheles gambiae* Species Complex. *Genome Biol. Evol.* **2015**, *7*, 3116–3131. [CrossRef]
38. Sharma, A.; Kinney, N.A.; Timoshevskiy, V.A.; Sharakhova, M.; Sharakhov, I. Structural Variation of the X Chromosome Heterochromatin in the *Anopheles gambiae* Complex. *Genes* **2020**, *11*, 327. [CrossRef]
39. Schmidt, H.; Collier, T.C.; Hanemaaijer, M.J.; Houston, P.D.; Lee, Y.; Lanzaro, G.C. Abundance of conserved CRISPR-Cas9 target sites within the highly polymorphic genomes of *Anopheles* and *Aedes mosquitoes*. *Nat. Commun.* **2020**, *11*, 1–6. [CrossRef]
40. Hammond, A.; Karlsson, X.; Morianou, I.; Kyrou, K.; Beaghton, A.; Gribble, M.; Kranjc, N.; Galizi, R.; Burt, A.; Crisanti, A.; et al. Regulation of gene drive expression increases invasive potential and mitigates resistance. *bioRxiv* **2020**. [CrossRef]
41. Champer, J.; Liu, J.; Oh, S.Y.; Reeves, R.; Luthra, A.; Oakes, N.; Clark, A.G.; Messer, P.W. Reducing resistance allele formation in CRISPR gene drive. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 5522–5527. [CrossRef]
42. Kandul, N.P.; Liu, J.; Buchman, A.; Gantz, V.M.; Bier, E.; Akbari, O.S. Assessment of a Split Homing Based Gene Drive for Efficient Knockout of Multiple Genes. *G3 Genes Genomes Genet.* **2020**, *10*, 827–837. [CrossRef]
43. Marshall, J.M.; Buchman, A.; Sánchez, C.H.M.; Akbari, O.S. Overcoming evolved resistance to population-suppressing homing-based gene drives. *bioRxiv* **2017**, *7*, 3776. [CrossRef]
44. Champer, S.E.; Oh, S.Y.; Liu, C.; Wen, Z.; Clark, A.G.; Messer, P.W.; Champer, J. Computational and experimental performance of CRISPR homing gene drive strategies with multiplexed gRNAs. *Sci. Adv.* **2020**, *6*. [CrossRef]
45. Liao, D. Concerted Evolution. In *Encyclopedia of Life Sciences*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2008. [CrossRef]